BIENNIAL REPORT

2008 – 2009

Adaptive Informatics Research Centre

Department of Information and Computer Science

Aalto University School of Science and Technology

P.O. Box 15400

FI–00076 Aalto, Finland

K. Raivio, J. Peltonen, and L. Koivisto, editors

Otaniemi, April 2010

Yliopistopaino
Helsinki 2010

Contents

Preface **7**

Personnel **9**

Awards and activities **13**

Doctoral dissertations **29**

Theses **41**

1 Introduction **47**

Algorithms and Methods

2 Bayesian learning of latent variable models **51**
Juha Karhunen, Antti Honkela, Tapani Raiko, Alexander Ilin, Koen Van Leemput, Jaakko Luttinen, Matti Tornio, Markus Harva . . . . . .
2.1 Bayesian modeling and variational learning 52
2.2 Algorithmic improvements for variational inference 54
2.3 Nonlinear state-space models for model-predictive control 57
2.4 Extensions of probabilistic PCA 59
2.5 Time-series modelling in bioinformatics 62
2.6 Estimation of time delays in gravitational lensing in astronomy 63
2.7 Automated segmentation of brain MR images 65

3 Blind and semi-blind source separation **71**
Erkki Oja, Alexander Ilin, Zhirong Yang, Zhijian Yuan, Jaakko Luttinen . .
3.1 Introduction 72
3.2 Non-negative projections 74
3.3 Reconstruction of historical climate data by Gaussian-process factor analysis 76

4 Multi-source machine learning **79**
Samuel Kaski, Arto Klami, Gayle Leen, Jaakko Peltonen, Kai Puolamäki, Janne Sinkkonen, Zhirong Yang, Helena Aidos, Ilkka Huopaniemi, Kristian Nybo, Juuso Parkkinen, Eerika Savia, Tommi Suvitaival, Abhishek Tripathi
4.1 Introduction 80
4.2 Multi-view learning 81
4.3 Multi-task learning 83
4.4 Multi-way learning 87
4.5 Information visualization 89
4.6 Networks 93

Bioinformatics and Neuroinformatics

5 Bioinformatics **97**
Samuel Kaski, Jarkko Salojärvi, Gayle Leen, Arto Klami, Jaakko Peltonen, José Caldas, Andrey Ermolov, Ali Faisal, Ilkka Huopaniemi, Leo Lahti, Juuso Parkkinen, Abhishek Tripathi . . . . . .
5.1 Introduction 98
5.2 Translational medicine on metabolic level 99
5.3 Retrieval and visualization of relevant experiments 101
5.4 Fusion of heterogeneous biomedical data 103

6 Neuroinformatics **107**
Ricardo Vigário, Miguel Almeida, Nicolau Gonçalves, Nima Reyhani, Jarkko Ylipaavalniemi, Elina Karp, Jayaprakash Rajasekharan, Jyri Soppela, Janne Nikkilä, Eerika Savia, Samuel Kaski, Erkki Oja . . . . . .
6.1 Introduction 108
6.2 Complex neural responses to complex stimuli 110
6.3 Phase synchrony 111
6.4 Single trial event related studies 112
6.5 Tissue segmentation in MRI 113
6.6 Document mining 114

Multimodal interfaces

7 Content-based information retrieval and analysis **117**
Erkki Oja, Jorma Laaksonen, Markus Koskela, Zhirong Yang, Ville Viitaniemi, Mats Sjöberg, He Zhang . . . . . .
7.1 Introduction 118
7.2 Semantic concept detection from images and videos 118
7.3 Video search and retrieval 119
7.4 Video analysis applications 120

8 Automatic speech recognition **123**
Mikko Kurimo, Kalle Palomäki, Teemu Hirsimäki, Janne Pylkkönen, Ville Turunen, Sami Virpioja, Matti Varjokallio, Ulpu Remes, Heikki Kallasjoki, Reima Karhila, Teemu Ruokolainen, Tanel Alumäe, Sami Keronen, Andre Mansikkaniemi . . . . . .
8.1 Introduction 124
8.2 Acoustic modeling 126
8.3 Language modeling 129
8.4 Applications and tasks 131
8.5 Noise robust speech recognition 134

9 Proactive Interfaces **137**
Samuel Kaski, Jorma Laaksonen, Mikko Kurimo, Arto Klami, Markus Koskela, Kai Puolamäki, Jarkko Salojärvi, Antti Ajanki, Mats Sjöberg, Ville Viitaniemi, He Zhang, Melih Kandemir, Laszlo Kozma, Lu Wei, Teemu Ruokolainen, Xi Chen, Erkki Oja . . . . . .
9.1 Introduction 138
9.2 Inferring interest from gaze patterns 138
9.3 Eye-movement enhanced image retrieval 138

9.4 Contextual information interfaces 140

10 Natural language processing **143**

Krista Lagus, Mikko Kurimo, Timo Honkela, Sami Virpioja, Oskar Kohonen, Mari-Sanna Paukkeri, Mikaela Klami, Ville Turunen, Matti Varjokallio, Matti Pöllä, Ilari Nieminen, Tommi Vatanen . . . . .

10.1 Introduction 144
10.2 Unsupervised learning of morphology 145
10.3 Unsupervised discovery of constructions 147
10.4 Keyphrase extraction 148
10.5 Morpho Challenge 149

Computational Cognitive Systems

11 Cognitive Systems Research **153**

Timo Honkela, Krista Lagus, Oskar Kohonen, Tiina Lindh-Knuutila, Mari-Sanna Paukkeri, Matti Pöllä, Juha Raitio, Sami Virpioja, Jaakko J. Väyrynen . . . . .

11.1 Introduction 154
11.2 Summary of collaboration 154
11.3 Summary of cognitive systems research areas 156

12 Conceptual modeling and learning **157**

Krista Lagus, Timo Honkela, Tiina Lindh-Knuutila, Mari-Sanna Paukkeri, Juha Raitio, Oskar Kohonen, Paul Wagner . . . . .

12.1 Introduction 158
12.2 Intersubjective communication model 160
12.3 Multiagent simulation model of conceptual development 163
12.4 Analysis of philosophy students' conceptions 165

13 Learning to translate **167**

Jaakko J. Väyrynen, Sami Virpioja, Timo Honkela, Mikko Kurimo, Marcus Dobrinkat, Tero Tapiovaara, Tommi Vatanen . . . . .

13.1 Introduction 168
13.2 Analysis of complexity of European languages 168
13.3 Learning interlingual mappings 168
13.4 Applying morphology learning to statistical machine translation 169
13.5 Experiments in speech-to-speech machine translation 169
13.6 Automatic machine translation evaluation 169
13.7 Within-language translation 170

14 Socio-cognitive modeling **173**

Timo Honkela, Krista Lagus, Tiina Lindh-Knuutila, Mari-Sanna Paukkeri, Juha Raitio, Eric Malmi . . . . .

14.1 Introduction 174
14.2 Modeling expertise at individual and social level 176
14.3 Knowledge practices and pragmatic web 178
14.4 Social simulation and ensemble models 179
14.5 Text mining in qualitative research 179
14.6 Analysis of consumer data 180
14.7 Supporting democratic innovation in organizations 180

14.8 Analysis of political popularity patterns . . . 181

15 Immune system inspired computing **185**
Matti Pöllä and Timo Honkela . . .
15.1 Introduction . . . 186
15.2 Anomaly detection . . . 186

Adaptive Informatics Applications

16 Intelligent data engineering **189**
Miki Sirola, Kimmo Raivio, Pasi Lehtimäki, Jukka Parviainen, Jaakko Talonen, Golan Lampi, Teemu Poikela, Eimontas Augilius, Olli Simula . . . .
16.1 Data analysis in industrial operator support . . . 190
16.2 Cellular network optimization . . . 192

17 Time series prediction **195**
Amaury Lendasse, Francesco Corona, Federico Montesino-Pouzols, Patrick Bas, Antti Sorjamaa, Mark van Heeswijk, Laura Kainulainen, Eric Severin, Elia Liitiäinen, Tuomas Kärnä, Yu Qi, Yoan Miche, Emil Eirola, Dušan Sovilj, Olli Simula . . .
17.1 Introduction . . . 196
17.2 European Symposium on Time Series Prediction . . . 197
17.3 Tools for long-term prediction of time series . . . 202
17.4 Nonparametric noise estimation . . . 203
17.5 Imputation of missing data in climatology and finance . . . 205
17.6 OP-ELM and ensembles of ELM . . . 207
17.7 Chemoinformatics . . . 209
17.8 Steganography and steganalysis . . . 210
17.9 Bankruptcy prediction . . . 211

Individual projects
A. On the quantization error in SOM vs. VQ: A critical and systematic study . . 215

Publications of the Adaptive Informatics Research Centre **221**

Preface

The **Adaptive Informatics Research Centre (AIRC, adaptiivisen informatiikan tutkimusyksikkö)** was nominated as one of the national Centers of Excellence (CoE) by the Academy of Finland for the period 2006 - 2011. It is financed by the Academy, Tekes, HUT, and Nokia Co.

The present report covers the activities of AIRC during the years 2008 and 2009. It concentrates on the research projects, but also lists the degrees and awards given to the staff. The achievements and developments of the previous two years have been reported in the Biennial Report 2006 - 2007. The web pages of AIRC, `http://www.cis.hut.fi/` also contain up-to-date texts.

During 2008 - 2009, the AIRC was operating within the new Department of Information and Computer Science (ICS), belonging to the Faculty of Information and Natural Sciences of Helsinki University of Technology. Professor Erkki Oja was the director of AIRC, and Professor Samuel Kaski was the vice-director, with Professors Olli Simula and Juha Karhunen participating in its research projects. In addition, 16 post-doctoral researchers, ca. 45 graduate students, and a number of undergraduate students were working in the AIRC projects in 2009. In terms of the personnel and budget, AIRC comprises about two thirds of the ICS Department.

To briefly list the main numerical outputs of AIRC during the period 2008 - 2009, the laboratory produced 10 D.Sc. (Eng.) degrees and 31 M.Sc. (Eng.) degrees. The number of scientific publications appearing during the period was 176, of which 37 were journal papers. It can be also seen that the impact of our research is clearly increasing, measured by the citation numbers to our previously published papers and books, as well as the number of users of our public domain software packages.

A large number of talks, some of them plenary and invited, were given by our staff in the major conferences in our research field. We had several foreign visitors participating in our research, and our own researchers made visits to universities and research institutes abroad. The total number of visit months both to and from AIRC was 187. The research staff were active in international organizations, editorial boards of journals, and conference committees, including NIPS 2008, AKRR / ESTSP 2008 and ICML / MLG / UAI / COLT 2008. Also, some prices and honours, both national and international, were granted to members of our staff. All these are detailed in this report.

The second meeting of the Scientific Advisory Board of AIRC was held on May 12 - 13, 2008. The evaluation report written by the members of the Board, Professors Risto Miikkulainen and José C. Príncipe, was quite positive. It begins by stating that "The group ... continues to be a reference in the international research community in the area of adaptive informatics. Therefore, the prestigious status of Center of Excellence bestowed by the Academy of Finland is a well deserved recognition. The vision and scientific impact of the CoE is on par with the best centers in the world in this area, not only in terms of productivity, but primarily because of the quality of its scientific production and in the novelty of its contributions."

Another evaluation for the whole ICS laboratory was carried out in summer 2009 in the context of the international evaluation of all the departments of the new Aalto University, which started its operations in January 2010. The ICS laboratory won the first place among the 46 departments, shared with the Department of Applied Physics, with an almost perfect numerical score of 24 / 25. In that evaluation, the foreign experts state e.g. that "The research quality of the department is outstanding ...(it) has a long tradition of excellence in research and doctoral education... The citation impact is very impressive ... the department is one of the top five centers in the world in their research area."

Erkki Oja

Professor
Director,
Adaptive Informatics
Research Centre

Samuel Kaski

Professor
Vice-Director,
Adaptive Informatics
Research Centre

Personnel

Professors

Oja, Erkki; D.Sc. (Tech.), Director

Karhunen, Juha; D.Sc. (Tech.), part-time

Kaski, Samuel; D.Sc. (Tech.), Vice Head of the department

Kohonen, Teuvo; D.Sc. (Tech.), Emeritus Professor, Academician

Simula, Olli; D.Sc. (Tech.), Dean, Faculty of Information and Natural Sciences

Post-doc researchers

Alumäe, Tanel; Ph.D., from April 2009

Corona, Francesco; Ph.D.

Girdziusas, Ramunas; D.Sc. (Tech.), until July 2008

Harva, Markus; D.Sc. (Tech.), until August 2008

Hirsimäki, Teemu; D.Sc. (Tech.), on leave from October 2009

Honkela, Antti; D.Sc. (Tech.)

Honkela, Timo; Ph.D., Chief research scientist

Ilin, Alexander; D.Sc. (Tech.)

Klami, Arto; D.Sc. (Tech.)

Koskela, Markus; D.Sc. (Tech.)

Kujala, Jussi; D.Sc. (Tech.), from November 2009

Kurimo, Mikko; D.Sc. (Tech.), Teaching research scientist, Chief research scientist from August 2008

Laaksonen, Jorma; D.Sc. (Tech.), Teaching research scientist

Lagus, Krista; D.Sc. (Tech.)

Leen, Gayle; Ph.D., from April 2008

Lendasse, Amaury; Ph.D.

Montesino-Pouzols, Federico; Ph.D., from June 2009

Nikkilä, Janne; D.Sc. (Tech.), Teaching research scientist until March 2008, Researcher from April 2008 until August 2009

Ogul, Hasan; Ph.D., from August 2008 until July 2009

Palomäki, Kalle; D.Sc. (Tech.)

Peltonen, Jaakko; D.Sc. (Tech.)

Raiko, Tapani; D.Sc. (Tech.)

Raivio, Kimmo; D.Sc. (Tech.), Teaching research scientist

Salojärvi, Jarkko; D.Sc. (Tech.), Teaching research scientist from April 2008 until August 2009

Vigário, Ricardo; D.Sc. (Tech.)

Yang, Zhirong; D.Sc. (Tech.)

Yuan, Zhijian; D.Sc. (Tech.), from May 2008

Post-graduate researchers

Aidos, Helena; M.Sc., Grant researcher until December 2009

Ajanki, Antti; M.Sc. (Tech.)

Almeida, Miguel; M.Sc., Grant researcher until December 2009

Caldas, Jose; M.Sc., Grant researcher

Chen, Xi; M.Sc. (Tech.), from September 2009

Dobrinkat, Marcus; M.Sc., from April 2009

Ermolov, Andrey; M.Sc. (Tech.), until October 2009

Faisal, Ali; M.Sc.

Gonçalves, Nicolau; M.Sc., Grant researcher

van Heeswijk, Mark; M.Sc., from June 2008

Huopaniemi, Ilkka; M.Sc. (Tech.)

Kallasjoki, Heikki; M.Sc. (Tech.)

Kandemir, Melih; M.Sc., from August 2008

Keronen, Sami; M.Sc. (Tech.), from September 2008

Klami, Mikaela; M.Sc. (Tech.), part-time

Kohonen, Oskar; M.Sc. (Tech.)

Lahti, Leo; M.Sc. (Tech.)

Lehtimäki, Pasi; M.Sc. (Tech.), until January 2008

Liitiäinen, Elia; M.Sc. (Tech.)

Lindh-Knuutila, Tiina; M.Sc. (Tech.)

Miche, Yoan; M.Sc., Grant researcher, Researcher from April 2009

Nybo, Kristian; M.Sc. (Tech.)

Pajarinen, Joni; M.Sc. (Tech.), from February 2008

Paukkeri, Mari-Sanna; M.Sc. (Tech.)

Parviainen, Jukka; M.Sc. (Tech.), University teacher

Pylkkönen, Janne; M.Sc. (Tech.)

Pöllä, Matti; M.Sc. (Tech.)

Raitio, Juha; M.Sc. (Tech.)

Remes, Ulpu; M.Sc. (Tech.)

Reyhani, Nima; M.Sc. (Tech.)

Ruokolainen, Teemu; M.Sc. (Tech.), from May 2008

Savia, Eerika; M.Sc. (Tech.), until October 2009

Sjöberg, Mats; M.Sc. (Tech.)

Sorjamaa, Antti; M.Sc. (Tech.)

Sovilj, Dušan; M.Sc. (Tech.), from February 2008

Suvitaival, Tommi; M.Sc. (Tech.), from May 2008

Talonen, Jaakko; M.Sc. (Tech.)

Turunen, Ville; M.Sc. (Tech.)

Varjokallio, Matti; M.Sc. (Tech.), until October 2009

Viitaniemi, Ville; M.Sc. (Tech.), Assistant

Virpioja, Sami; M.Sc. (Tech.)

Wei, Lu; M.Sc., from May until August 2008

Väyrynen, Jaakko; M.Sc. (Tech.)

Ylipaavalniemi, Jarkko; M.Sc. (Tech.)

Yu, Qi; M.Sc. (Tech.), from March until July 2008, from August 2008 Grant researher, from January 2009 Researcher

Zhang, He; M.Sc. (Tech.)

Under-graduate researchers

Cho, Kyunghyun; from October 2009

Eirola, Otto

Ellonen, Sakari

Gillberg, Leo

Kainulainen, Laura; from June 2009

Karhila, Reima; from February 2008

Karp, Elina; from June 2008

Kivimäki, Ilkka; from June 2009

Klapuri, Antti; from May until August 2008

Koistinen, Olli-Pekka; from June until December 2008

Korsakova, Natalia; from July 2008 until March 2009

Kozma, Laszlo; from February 2008 until June 2009

Kuusela, Mikael; from May 2008

Lampi, Golan; on leave from May 2009

Luttinen, Jaakko

Malmi, Eric; from May 2008

Mansikkaniemi, Andre; from September 2008

Mohammadi, Pejman; from August 2008

Nevala, Maija; from May 2008

Nieminen, Ilari

Osmala, Maria; from June 2009

Parkkinen, Juuso; from June 2008 until June 2009

Peltola, Veli

Poikela, Teemu; from June until August 2008

Raatikainen, Niklas

Salminen, Eero; from January until June 2008

Smit, Pieter; from June until August 2009
Soppela, Jyri; from June 2008
Suikkanen, Saara; from June until August 2008
Tapiovaara, Tero; from June 2008
Tiinanen, Teemu; from June 2008
Wagner, Paul
Vatanen, Tommi; from June 2008
Viinikanoja, Jaakko; from June 2009
Virtanen, Seppo
Wu, Jing; from March 2009
Yao, Li; from June 2009
Zhu, Zhanxing; from October 2009

Support staff

Koivisto, Leila; Secretary
Pihamaa, Tarja; Secretary
Ranta, Markku; B.Eng., Works engineer
Sirola, Miki; D.Sc. (Tech.), Laboratory engineer

Awards and activities

Prizes and academic awards received by personnel of the unit

Academician Teuvo Kohonen:

- IEEE Frank Rosenblatt Award, IEEE, USA, 2008

Professor Erkki Oja:

- Honorary doctor of philosophy, Uppsala University, Sweden, 2008
- Honorary doctor of technology, Lappeenranta University of Technology, Finland, 2008

Dr. Markus Harva:

- Best Nordic PhD-thesis Award in the field of Image Analysis and Pattern Recognition, 2007-2008, Nordic Image Processing and Pattern Recognition Research Societies, Norway, 2009

Dr. Jaakko Peltonen:

- Outstanding reviewer for the Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS 2008), Canada, 2008

Dr. Zhirong Yang:

- Chinese Government Award for Outstanding Self-Financed Students Abroad, April 2008.

M.Sc. Jukka Parviainen:

- Teacher of Year 2008 in the Computer Science Engineering Programme, Helsinki University of Technology

Important international positions of trust held by personnel of the unit

Professor Juha Karhunen:

- Program Committee Member:
  The 2008 European Signal Processing Conference (EUSIPCO-2008), Lausanne, Switzerland, August 2008.
  Int. Workshop on Computational Intelligence in Security for Information Systems (CISIS'08), Genoa, Italy, October 2008.
  The 16th European Symposium on Artificial Neural Networks (ESANN2008), Bruges, Belgium, April 2008.
  The 2008 IEEE Int. Joint Conf. on Neural Networks (IJCNN2008), Hong Kong, China, June 2008.
  The 9th Int. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL2008), Daejeon, South Korea, November 2008.
  The 2008 IEEE Workshop on Machine Learning in Signal Processing (MLSP2008), Cancun, Mexico, October 2008.
  The 3rd Int. Workshop on Hybrid Artificial Intelligence Systems (HAIS'08), Burgos, Spain, September 2008.
  IEEE Symposium on Computational Intelligence in Data Mining (CIMD2009), Nashville, Tennessee, USA, April 2009.
  The 4th Int. Workshop on Hybrid Artificial Intelligence Systems (HAIS'09), Salamanca, Spain, June 2009.
  The 2009 Int. Conf. on Independent Component Analysis and Signal Separation (ICA 2009), Paraty, Brazil, March 2009.
  The European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning (ESANN'2009), Bruges, Belgium, April 2009.
  The 10th Int. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL2009), Burgos, Spain, September 2009.
  The 2009 IEEE International Workshop on Machine Learning in Signal Processing (MLSP2009), Grenoble, France, September 2009.
- Editorial Board Member, Neural Processing Letters, The Netherlands.

Professor Samuel Kaski:

- Program Committee Member:
  NIPS 2008 Workshop on Learning from Multiple Sources, Whistler, Canada, 13 Dec., 2008.
  International Symposium on Intelligent Data Engineering and Automated Learning, Daejeon, Korea, November 2-5, 2008.
  European Conference on Machine Learning, Antwerp, Belgium, September 15-19, 2008.
  Machine Learning for Signal Processing, Cancun, Mexico, October 16-19, 2008.
  Web Intelligence, Sydney, Australia, December 9-12, 2008.
  European Symposium on Artificial Neural Networks, Bruges, Belgium, April 23-25, 2008.
  Computational Intelligence Methods for Bioinformatics and Biostatistics, Vietri sul Mare, Salerno, Italy, October 3-4, 2008.
  Machine Learning for Systems Biology, Brussels, Belgium, September 13-14, 2008.
  Sixth Asia-Pacific Bioinformatics Conference (APBC 2008), Kyoto, Japan, January

14-17, 2008.
ICPR-08, The 19th International Conference on Pattern Recognition, Tampa, Florida, USA, December 8-11, 2008.
Fifth International Workshop on Computational Systems Biology, WCSB 2008, Leipzig, Germany, June 11-13, 2008.
ECCB08, European Conference on Computational Biology, Cagliari, Italy, September 22-26, 2008.
ECAI 2008, European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008.
WABI 2008, 8th Workshop on Algorithms in Bioinformatics, Karlsruhe, Germany, September 15-17, 2008.
International Symposium on Intelligent Data Engineering and Automated Learning IDEAL'09, Burgos, Spain, September 23-26, 2009.
Workshop on Self-Organizing Maps, WSOM09, St Augustine, Florida, USA, June 8-10, 2009.
IEEE/WIC International Conference on Web Intelligence, WI 2009, Milano, Italy, September 15-18, 2009.
European Symposium on Artificial Neural Networks, Bruges, Belgium, April 22-24, 2009.
Machine Learning for Structural and Systems Biology, MLSB'09, Ljubljana, Slovenia, September 5-6, 2009.
International Conference on Machine Learning CML 2009, Montreal, Canada, June 14-18, 2009.
Asia-Pacific Bioinformatics Conference APBC2009, Beijing, China, January 13-16, 2009.
International Workshop on Computational Systems Biology, WCSB 2009, Aarhus, Denmark, June 10-12, 2009.
The 9th IEEE International Symposium on Bioinformatics and Bio Engineering BIBE 2009, Taichung, Taiwan, June 22-24, 2009.
Mining and Learning with Graphs (MLG-2009), Leuven, Belgium, July 2-4, 2009.
PRIB 2009, Pattern Recognition in Bioinformatics, Sheffield, U.K., September 7-9, 2009.

- Session Chairman: NIPS 2008 Workshop on Learning from Multiple Sources, Whistler, Canada, December 13, 2008.

- Member of Steering Committee, EU NoE PASCAL2, UK.

- Evaluator of CoE applications, Slovenian Research Agency ARRS, Slovenia, 2009.

- Associate Editor, International Journal of Knowledge Discovery in Bioinformatics, Singapore, 2009.

- Editorial Board Member:
  Intelligent Data Analysis, The Netherlands.
  International Journal of Neural Systems, Singapore.
  Cognitive Neurodynamics, Germany.

- Pre-examiner of a doctoral thesis:
  Zhu Yan, Nanyang Technological University, Singapore, 2008.
  Steffen Bickel, Universitat Potsdam, Germany, 2008.

- Opponent at the doctoral dissertation of Jörg Ontrup, University of Bielefeld, Germany, 2008.
- Issuing a statement for appointment to office of a professor: NUI Galway, Bioinformatics, Ireland, 2008.
- Invited talks:
  EMMDS 2009, European Workshop on Challenges in Modern Massive Data Sets, Copenhagen, Denmark July 1–4, 2009
  4th IAPR International Conference in Pattern Recognition for Bioinformatics (PRIB 2009), special session on Machine learning in integrative genomics
- Keynote talk:
  16th Scandinavian Conference on Image Analysis (SCIA), Oslo, Norway, 15.–18.6. 2009

Academician Teuvo Kohonen:

- Member of the committee, College of Fellows, International Neural Network Society, INNS, USA, 2009.
- Plenary talk, "On the quantization error in SOM vs. VQ: a critical and systematic study" coauthored by I. Nieminen and T. Honkela, 7th International Workshop on Self-Organizing Maps (WSOM 2009), St. Augustine, Florida, USA, June 8–10, 2009.

Professor Erkki Oja:

- Program Committee Member, NIPS, Vancouver, Canada, December 9-13, 2008.
- Governing Board Member, International Neural Network Society, INNS, USA, 2008.
- Member of the committee, College of Fellows, International Neural Network Society, INNS, USA, 2009.
- Member of the Award Committee, IEEE Computational Intelligence Society, USA, 2008.
- Member of the Fellowship Committee, IEEE Computational Intelligence Society, USA, 2009.
- Member of the IAPR, K.S. Fu Award Committee, USA, 2008.
- Evaluator of research program, Science Foundation Ireland, 2008.
- Editorial Board Member:
  Natural Computing - An International Journal, The Netherlands.
  Neural Computation, USA.
  International Journal of Pattern Recognition and Artificial Intelligence, Singapore.
- Opponent at the doctoral dissertation of Andreas Brinch Nielsen, Technical University of Denmark (DTU), 2009.
- Issuing a statement for appointment to office of a professor:
  Oxford University, Professorship in Information Engineering, U.K., 2008
  University College Dublin, SFI Stokes Professorship in Computer Science and Informatics, Ireland, 2008.
  Katholieke Universiteit Leuven, Research Professor, Belgium, 2009.

- Plenary talk in International Conference on Artificial Neural Networks (ICANN), Limassol, Cyprus, September 15 – 18, 2009

Professor Olli Simula:

- Program Committee Member: International Conference on Artificial Neural Networks, ICANN 2008, Prague, Czech, September 3-6, 2008.
- Scientific Council Member, Institute Eurecom, France.

Dr. Antti Honkela:

- Program Committee Member, 4th IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB 2009, Sheffield, U.K., September 7-9, 2009.

Dr. Timo Honkela:

- Program Committee Member:
  ICANN'08, 18th International Conference on Artificial Neural Networks, Prague, Czech Republic, September 3-6, 2008.
  IIP 2008, International Conference on Intelligent Information Processing, Lyon, France, September 10-12, 2008.
  ICANN 2009, International Conference on Artificial Neural Networks, Limassol, Cyprus, September 14-17, 2009.
- Chair, International Federation on Information Processing (IFIP), WG12.1 (Knowledge Representation and Reasoning), Austria.
- Representative of Finland, International Federation on Information Processing (IFIP), TC12 (Artificial Intelligence), Austria.
- Member of Executive Board, ENNS, European Neural Networks Society, The Netherlands.
- Editorial Board Member, Constructivist Foundations, Austria.
- Evaluator, European Commission Information Society & Media DG, Belgium, 2009.
- Issuing a statement for appointment to office of a professor:
  Lund University, Professor of Cognitive Science, Sweden, 2009.
- Opponent at the doctoral dissertation:
  Basilio Calderone, Scuola Normale Superiore di Pisa, Italy, 2008.
  Ling Feng, Technical University of Denmark, Denmark, 2008.
  Innar Liiv, Tallinn University of Technology, Estonia, 2008.
- Keynote talk "From quantification of information to quantification of meaning using socio-cognitive computing" at the 2008 IAPR Workshop on Cognitive Information Processing in June 2008.
- AERFAISS08 Summer School Tutorial, Bilbao, Spain, June 23-28, 2008.

Dr. Arto Klami:

- Program Committee Member:
  Learning from Multiple sources with applications to Robotics Workshop, Whistler, Canada, December 13, 2009 (Workshop in connection with NIPS 2009).
  Sixth International Symposium on Neural Networks (ISSN 2009), Wuhan, China, May 25-29, 2009.

- Organizing Committee Member:
  Mining and Learning with Graphs 2008 (MLG), Helsinki, Finland, July 4-5, 2008.
  International Conference on Machine Learning (ICML), Helsinki, Finland, July 5-9, 2008.
  Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, July 10-12, 2008.
  Conference on Learning Theory (COLT), Helsinki, Finland, July 10-12, 2008.

Dr. Markus Koskela:

- Program Committee Member:
  10th International Conference on Visual Information Systems, Salerno, Italy, September 11-12, 2008.
  19th International Conference on Pattern Recognition, Tampa, Florida, USA, September 8-11, 2008.
  TRECVID BBC Rushes Summarization Workshop at ACM Multimedia '08 Vancouver, British Columbia, Canada, Oct. 27- Nov. 1, 2008.
  The 2009 IEEE Pacific-Rim Conference on Multimedia (IEEE-PCM2009), Bangkok, Thailand, December 15-18, 2009.
  16th Scandinavian Conference on Image Analysis 2009 (SCIA 2009), Oslo, Norway, June 15-18, 2009.
  15th International MultiMedia Modeling Conference (MMM 2009), Sophia Antipolis, France, January 7-9, 2009.

Dr. Mikko Kurimo:

- Program Committee Member:
  Morpho Challenge Workshop, Aarhus, Denmark, September 17, 2008.
  Morpho Challenge Workshop 2009, Corfu, Greece, September 30, 2009.

- Editor, Advances in Multilingual and Multimodal Information Retrieval 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, Morpho Challenge, Hungary, 2008.

- Editor, Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, Lecture Notes in Computer Science Vol. 5706, 2009.

- Opponent at the doctoral dissertation of Ebru Arisoy, Bogazici University, Istanbul, Turkey, 2009.

- Invited talks:
  "Speech recognition and retrieval using unsupervised and adaptive sub-word LMs" International Computer Science Institute, Berkeley, CA, USA, February 4, 2008
  "Speech recognition and retrieval using unsupervised sub-word LMs" Google Inc.,

Mountain View, CA, USA, February 8, 2008
"Speech recognition and retrieval using unsupervised sub-word language models" Nokia Research Center, Palo Alto, CA, USA, February 13, 2008
"Large-vocabulary continuous speech recognition (LVCSR) of spontaneous speech" Speech Reduction Workshop, Max Planc Institute, Nijmegen, Netherlands, June 16, 2008
"Status of large-vocabulary continuous speech recognition (LVCSR) in Finnish" Finnish-Japanese speech technology seminar, Helsinki, Finland, August 7, 2008
"Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard and IR Experiments" Morpho Challenge Workshop 2008, Aarhus, Denmark, August 17, 2008
"CLEF at Morpho Challenge 2008 - Unsupervised Morpheme Analysis" CLEF Workshop 2008, Aarhus, Denmark, August 18, 2008
"Decomposition of words for speech recognition, retrieval and translation", Bogazici University, Istanbul, Turkey, December 2009.
"Overview of Morpho Challenge task at CLEF 2009", CLEF 2009 Workshop, Corfu, Greece, September 2009.
"Unsupervised decomposition of words for speech recognition and retrieval", Keynote of International Conference Speech and Computer SPECOM 2009, St. Petersburg, Russia, June2009.

Dr. Jorma Laaksonen:

- Program Committee Member:
  10th International Conference on Visual Information Systems, Salerno, Italy, September 11-12, 2008.
  The International Conference on Image Analysis and Recognition (ICIAR 2009), Halifax, Canada, July 6-8, 2009.

- Secretary, IEEE Finland Section, IEEE Institute of Electrical and Electronics Engineers, USA.

- Referee, Lithuanian State Science and Studies Foundation, Lithuania, 2009.

Dr. Amaury Lendasse:

- Guest Editor:
  Neurocomputing, The Netherlands.
  International Journal of Forecasting, UK.

- Opponent at the doctoral dissertation of Federico Montesino Pouzols, University of Sevilla, Spain, 2009.

Dr. Jaakko Peltonen:

- Program Committee Member:
  10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2008), Turin, Italy, Sept. 1-5, 2008.
  Learning from Multiple Sources Workshop, Whistler, Canada, Dec. 13, 2008. (Workshop in connection with NIPS 2008).
  Sixth International Symposium on Neural Networks (ISNN 2009), Wuhan, China, May 25-29, 2009.

- Program Committee Member and Session Chairman, Learning from Multiple Sources with Applications to Robotics (Workshop in connection with NIPS 2009), Whistler, Canada, December 12, 2009.
- Editorial Board Member, Neural Processing Letters, The Netherlands.

Dr. Tapani Raiko:

- Program and Organizing Committee Member, Mining and Learning with Graphs (MLG 2008), Helsinki, Finland, July 4-5, 2008.
- Program Committee Member:
  Scandinavian Conference on Artificial Intelligence (SCAI 2008), Stockholm, Sweden, May 26-28, 2008.
  7th International Workshop on Mining and Learning with Graphs, Leuven, Belgium, 2009.
  Scandinavian Conference on Image Analysis, Oslo, Norway, June 15-18, 2009.

Dr. Kimmo Raivio:

- ICT Domain expert, COST (European Cooperation in Science and Technology), Belgium.

Dr. Miki Sirola:

- Program Committee Member:
  International Conference on Modelling, Identification and Control, Innsbruck, Austria, February 11-13, 2008.
  International Conference on Modelling and Simulation, Quebec, Canada, May 26-28, 2008.
  International Conference on Applied Simulation and Modelling, Corfu, Greece, June 23-25, 2008.
  IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2007), Rende, Italy, September 21-23, 2009.
  International Conference on Modelling and Simulation, Banff, Canada, July 6-8, 2009.
  International Conference on Computational Intelligence, Honolulu, Hawaii, USA, August 17-19, 2009.

Dr. Zhirong Yang:

- Program Committee Member, 16th International Conference on Neural Information Processing, ICONIP 2009, Bangkok, Thailand, December 1-5, 2009.

M.Sc. Leo Lahti:

- Member of the COST/EUGESMA working group, U.K., 2009.

Important domestic positions of trust held by personnel of the unit

Professor Samuel Kaski:

- Pre-examiner of a doctoral thesis:
  Petri Kontkanen, University of Helsinki, 2009.
  Ying Tang, University of Helsinki, 2009.
- Opponent at the doctoral dissertation of Reija Autio, Tampere University of Technology, 2008.
- Issuing a statement for appointment to office of a docent, University of Helsinki, 2009.
- Invited talks:
  "Modeling, mining, and fusing gene expression," FIMM & Biomedicum Medical Bioinformatics Day, Helsinki, March 13, 2008.

  "Fuusio, relaatio ja visualisaatio: uusilla työkaluilla relevanssia biodatasta," Bioinformatiikan päivä, Bioinformatiikan seuran vuotoinen päätapahtuma, May 16, 2008.

  "Tolkkua biodataan," Suomalainen tiedeakatemia 100 vuotta: rajaton tietojenkäsittelytiede -seminaari, September 8, 2008.

Professor Erkki Oja:

- Chairman, Research council for natural sciences and engineering, Academy of Finland
- Member, Governing Board of the Academy of Finland
- Vice chairman, Finnish Academy of Sciences and Letters, group of Mathematics and Computer Science
- Opponent at the doctoral dissertation of Timo Ahonen, University of Oulu, 2009.

Professor Olli Simula:

- Chairman, IEEE Computer Chapter, Finland.
- Program Committee Member: European Symposium on Time Series Prediction, ESTSP 2008, Porvoo, Finland, Sept. 17-19, 2008.
- Opponent at the doctoral dissertation of Mikko Laurikkala, Tampere University of Technology, 2009.

Dr. Timo Honkela:

- Program Committee Member:
  AKRR'08, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, Porvoo, Finland, Sept. 17-19, 2008.
  ESTSP'08, European Symposium on Time Series Prediction, Porvoo, Finland, September 17-19, 2008.
  Tieteen päivät (Science Days), Helsinki, Finland, January 7-11, 2009.
- Editorial Board Member, Puhe ja kieli, Finland.

- Editorial Board Member, Tieteessä tapahtuu, Finland, 2009.

- Editor, Timo Honkela, Mari-Sanna Paukkeri, Matti Pöllä, and Olli Simula, Eds. Proceedings of AKRR'08, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning. Helsinki University of Technology, Espoo, Finland, 2008.

- Member of Steering Committee, Langnet - Finnish Graduate School in Language Studies, Finland, 2009.

- Board Member, Hecse - Helsinki Graduate School in Computer Science and Engineering, Finland, 2009.

- Opponent at the doctoral dissertation:
  Antti Järvelin, University of Tampere.
  Tapio Pahikkala, University of Turku.

- Panel discussion "How evolution theory influenced and influences in different scientific disciplines", organized by the Federation of Finnish Learned Societies and the Committee for Public Information, Turku, October 6, 2008.

- Invited talks:
  "With our languages, with our minds: the deep structures of communication", in Communications seminar, organized by Finnish Association of Marketing Communication Agencies MTL, October 28, 2008.

  "Linguistic string manipulation", the annual symposium on string methods, Helsinki University of Technology, January 8, 2009.

- Talk "The co-evolution of man and machine", in the Science Forum (Tieteen päivät), University of Helsinki, January 10, 2009.

- Chairing the session "Future of intelligence", at the Science Forum (Tieteen päivät), University of Helsinki, January 10, 2009.

- Chairing the seminar "Interdisciplinarity in language research", organized by the Langnet graduate school, November 26-27, 2009.

Dr. Markus Koskela:

- Member of the board, Suomen hahmontunnistuksen seura ry, Pattern Recognition Society of Finland, 2008.

- Vice President, Suomen hahmontunnistuksen seura ry, Pattern Recognition Society of Finland, 2009.

Dr. Mikko Kurimo:

- Program Committee Member: EMIME FP7 August meeting, Espoo, Finland, August 4-6, 2008.

- Opponent at the doctoral dissertation of Juha Yli-Sipilä, Tampere University of Technology, 2008.

- Invited talks:
  "Computational models of speech recognition (and synthesis)" NeuroCafe, NeuroHUT & Finnish Graduate School of Neuroscience, Espoo, Finland, October 21, 2008

  "Large-Vocabulary Speech Recognition and Information Retrieval" Hatutus Fall Seminar 2008 - Pattern Recognition in Human-Technology Interaction, Tampere, Finland, November 21, 2008

Dr. Jorma Laaksonen:

- Pre-examiner of a doctoral thesis, Jussi Lindgren, University of Helsinki, 2008.
- Opponent at the doctoral dissertation of Ville Ojansivu, University of Oulu, 2009.
- Invited talk: "Research and teaching of digital image processing at TKK" in a joint research meeting of Nokia Research Center, Nokia Devices and Finnish universities in Tampere on February 5, 2008.

Dr. Jaakko Peltonen:

- Pre-examiner of a doctoral thesis, Jussi Kollin, University of Helsinki, 2009.

Dr. Tapani Raiko:

- Program and Organizing Committee Chairman: Suomen tekoälytutkimuksen päivät (STeP 2008), Espoo, Finland, August 20, 2008.
- Chairman, Finnish Artificial Intelligence Society, Suomen Tekoälyseura, Finland, 2008.
- Vice Chairman, Finnish Artificial Intelligence Society, Suomen Tekoälyseura, Finland, 2009.

Dr. Kimmo Raivio:

- Pre-examiner of a doctoral thesis:
  Toni Huovinen, Tampere University of Technology, 2008.
  Miika Rajala, Tampere University of Technology, 2009.

Dr. Ricardo Vigário:

- Chairman and Program Committee Member, AKRR'08, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, Porvoo, Finland, September 17-19, 2008.
- Issuing a statement for appointment to office of a docent, Tampere University of Technology, Department of Signal Processing, 2008.

M.Sc. Leo Lahti:

- Board Member and PR-officer, Finnish Society for Bioinformatics.

M.Sc. Tiina Lindh-Knuutila:

- Program Committee Member, Second International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, AKRR'08, Porvoo, Finland, September 17-19, 2008.

M.Sc. Matti Pöllä:

- Program Committee Member, Second International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, AKRR'08, Porvoo, Finland, September 17-19, 2008.

Research visits abroad by personnel of the unit

- Dr. Antti Honkela, University of Manchester, 2 months, 2009.
- Dr. Mikko Kurimo, SRI, Stanford Research Institute, USA, 2 months, 2008.
- Mikael Kuusela, CERN, Switzerland, 3 months, 2009.
- M.Sc. Tiina Lindh-Knuutila, International Computer Science Institute, UC Berkeley, USA, 6 months, 2009.
- Eric Malmi, CERN, Switzerland, 3 months, 2009.
- M.Sc. Yu Qi, University of Lille, France, 2 months, 2009.
- M.Sc. Mari-Sanna Paukkeri, University of Edinburgh, 6 months, 2009.
- M.Sc. Antti Sorjamaa, Universidad de Granada, Spain, 2 months, 2009.
- Dr. Ricardo Vigário, Grenoble Institute of Technology, France, 1 month, 2008.
- Dr. Zhirong Yang, Microsoft, Cambridge, U.K., 3 months, 2008.
- Dr. Zhirong Yang, Chinese University of Hong Kong, 4 months, 2009.
- M.Sc. Yoan Miche, Gipsa-Lab, INPG, Grenoble, France, 1 month, 2008.
- M.Sc. Nima Reyhani, University of California at Berkeley, USA, 11 months, 2008-2009.

Research visits by foreign researchers to the unit

- M.Sc. Bhattacharya Shourangshu, Indian Institute of Science, Bangalore, India, 4 months, 2008.
- M.Sc. Nils Gehlenborg, European Bioinformatics Institute, Cambridge, U.K., 2 months, 2008; 3 weeks, 2009.
- Ph.D. Gayle Leen, University of Paisley, U.K., from April 2008-.
- Ph.D. Sounak Chakraborty, Ass. Prof. University of Missouri-Columbia, USA, 2 months, 2008.
- M.Sc. Jose Vellez Caldas, INESC, Portugal, from January 2008-.
- M.Sc. Ali Faisal, University of Edinburgh, U.K., from August 2008-.
- M.Sc. Alba Martinetz-Ruiz, Universitat Politècnica de Catalunya, Spain, 3 months, 2008.
- M.Sc. Melih Kandemir, Bilkent University, Turkey, from August 2008-.
- Ph.D. Hasan Ogul, Baskent University, Turkey, 1 year, 2008-2009.
- M.Sc. Yusuf Yaslan, Istanbul Technical University, Turkey, 2 months, 2008.
- Ph.D. Alberto Guillen, University of Granada, Spain, 2 months, 2008.
- M.Sc. Federico Montesino Pouzols, University of Sevilla, Spain, 2 months, 2008.
- M.Sc. Fernando Mateo Jimenez, Polytechnic University of Valencia, Spain, 6 months, 2008.
- Ph.D. Hidekazu Yanagimoto, Osaka Prefecture University, Japan, 1 year, 2008-2009.
- M.Sc. Marcin Blachnik, Silesian University of Technology, Katowice, Poland, 1 month, 2008.
- M.Sc. Miguel Almeida, Technical University of Lisbon, Portugal, 2 years, 2008-2009.
- M.Sc. Helena Aidos, Technical University of Lisbon, Portugal, 2 years, 2008-2009.
- M.Sc. Ane Amaia Orue-Etxebarria Apellaniz, University of the Basque Country, Spain, 11 months, 2008.
- M.Sc. Philip Prentis, Czech Technical University, Prague, Czechoslovakia, 2 weeks, 2008.
- M.Sc. Alberto Perez Garcia-Plaza, The Universidad Nacional de Educacion a Distancia (UNED), Spain, 3 months, 2009.
- Eugene Seo, Information and Communications University (ICU), South Korea, 6 months, 2009.
- M.Sc. Indre Zliobaite, Vilnius University, Lithuania, 3 months, 2009.
- Ph.D. Tanel Alumäe, Institute of Cybernetics at Tallinn University of Technology, Estonia, from April 2009-.

- M.Sc. Jort Gemmeke, Radboud University Nijmegen, The Netherlands, 2 months, 2009.
- Ph.D. Alex Leung, University of Leoben, Austria, 2 weeks, 2009.
- Ph.D. Zakria Hussain, University College London, U.K., 2 weeks, 2009.
- Ph.D. Federico Montesino Pouzols, University of Sevilla, Spain, 7 months, 2009.
- Prof. Eric Severin, University of Lille, France, 2 months, 2009.
- Prof. Concolacio Gil Montoya, Universidad de Almeria, Spain, 1 month, 2009.
- M.Sc. Mirtha Albanez-Lucero, CICIMAR, Instituto Politécnico Nacional, Mexico, 4 months, 2009.
- Ph.D. Koen van Leemput, Massachusetts General Hospital; Harvard Medical School, USA, 11 months, 2009.
- Prof. James Ramsay, McGill University, Canada, 5 days, 2009.

Other activities

Professor Erkki Oja:

- Interview on speech recognition, TV1 science program Prisma, September 8, 2009.

Dr. Mikko Kurimo:

- Talk titled "Speech recognition and retrieval using sub-word language models" `http://www.youtube.com/watch?v=KEqJpIDwO2Y` in Google Tech Talks on February 8, 2008.

Dr. Timo Honkela:

- Talk and chairing the seminar "Golf research, development, design and business", Helsinki University of Technology, December 11, 2009.

Doctoral dissertations

Learning from environmental data: methods for analysis of forest nutrition time series

Mika Sulkava

Dissertation for the degree of Doctor of Science in Technology on 18 January 2008.

External examiners:
Sašo Džeroski (Jožef Stefan Institute, Ljubljana, Slovenia)
Alfredo Vellido (Universitat Politècnica de Catalunya, Barcelona, Spain)
Opponent:
Thomas Martinez (University of Lübeck, Germany)

Abstract:
Data analysis methods play an important role in increasing our knowledge of the environment as the amount of data measured from the environment increases. This thesis fits under the scope of environmental informatics and environmental statistics. They are fields, in which data analysis methods are developed and applied for the analysis of environmental data.

The environmental data studied in this thesis are time series of nutrient concentration measurements of pine and spruce needles. In addition, there are data of laboratory quality and related environmental factors, such as the weather and atmospheric depositions.

The most important methods used for the analysis of the data are based on the self-organizing map and linear regression models. First, a new clustering algorithm of the self-organizing map is proposed. It is found to provide better results than two other methods for clustering of the self-organizing map. The algorithm is used to divide the nutrient concentration data into clusters, and the result is evaluated by environmental scientists. Based on the clustering, the temporal development of the forest nutrition is modeled and the effect of nitrogen and sulfur deposition on the foliar mineral composition is assessed.

Second, regression models are used for studying how much environmental factors and properties of the needles affect the changes in the nutrient concentrations of the needles between their first and second year of existence. The aim is to build understandable models with good prediction capabilities. Sparse regression models are found to outperform more traditional regression models in this task.

Third, fusion of laboratory quality data from different sources is performed to estimate the precisions of the analytical methods. Weighted regression models are used to quantify how much the precision of observations can affect the time needed to detect a trend in environmental time series. The results of power analysis show that improving the quality may decrease the time needed for detection of the trend by many years.

The data analysis methods developed and applied in this thesis are found to produce results which are understandable for the environmental scientists. They are, therefore, useful for studying the condition of the environment and evaluating the possible causes for changes in it.

Stability and inference in discrete diffusion scale-spaces

Ramunas Girdziusas

Dissertation for the degree of Doctor of Science in Technology on 29 February 2008.

External examiners:
Samuli Siltanen (Tampere University of Technology)
Keijo Ruotsalainen (University of Oulu)
Opponents:
Samuli Siltanen (Tampere University of Technology)
Atanas Gotchev (Tampere University of Technology)

Abstract:
Taking averages of observations is the most basic method to make inferences in the presence of uncertainty. In late 1980's, this simple idea has been extended to the principle of successively average less where the change is faster, and applied to the problem of revealing a signal with jump discontinuities in additive noise.

Successive averaging results in a family of signals with progressively decreasing amount of details, which is called the scale-space and further conveniently formalized by viewing it as a solution to a certain diffusion-inspired evolutionary partial differential equation (PDE). Such a model is known as the diffusion scale-space and it possesses two long-standing problems: (i) model analysis which aims at establishing stability and guarantees that averaging does not distort important information, and (ii) model selection, such as identification of the optimal scale (diffusion stopping time) given an initial noisy signal and an incomplete model.

This thesis studies both problems in the discrete space and time. Such a setting has been strongly advocated by Lindeberg [1991] and Weickert [1996] among others. The focus of the model analysis part is on necessary and sufficient conditions which guarantee that a discrete diffusion possesses the scale-space property in the sense of sign variation diminishing. Connections with the total variation diminishing and the open problem in a multivariate case are discussed too.

Considering the model selection, the thesis unifies two optimal diffusion stopping principles: (i) the time when the Shannon entropy-based Liapunov function of Sporring and Weickert [1999] reaches its steady state, and (ii) the time when the diffusion outcome has the least correlation with the noise estimate, contributed by Mrázek and Navara [2003]. Both ideas are shown to be particular cases of the marginal likelihood inference. Moreover, the suggested formalism provides first principles behind such criteria, and removes a variety of inconsistencies. It is suggested that the outcome of the diffusion should be interpreted as a certain expectation conditioned on the initial signal of observations instead of being treated as a random sample or probabilities. This removes the need to normalize signals in the approach of Sporring and Weickert [1999], and it also better justifies application of the correlation criterion of Mrázek and Navara [2003].

Throughout this work, the emphasis is given on methods that enable to reduce the problem to that of establishing the positivity of a quadratic form. The necessary and sufficient conditions can then be approached via positivity of matrix minors. A supplementary appendix is provided which summarizes a novel method of evaluating matrix minors. Intuitive examples of difficulties with statistical inference conclude the thesis.

Data analysis methods for cellular network performance optimization

Pasi Lehtimäki

Dissertation for the degree of Doctor of Science in Technology on 3 April 2008.

External examiners:
Jyrki Joutsensalo (University of Jyväskylä)
Ari Hämäläinen (Nokia Research Center)
Opponent:
Tapani Ristaniemi (University of Jyväskylä)

Abstract:
Modern cellular networks including GSM/GPRS and UMTS networks offer faster and more versatile communication services for the network subscribers. As a result, it becomes more and more challenging for the cellular network operators to enhance the usage of available radio resources in order to meet the expectations of the customers.

Cellular networks collect vast amounts of measurement information that can be used to monitor and analyze the network performance as well as the quality of service. In this thesis, the application of various data-analysis methods for the processing of the available measurement information is studied in order to provide more efficient methods for performance optimization.

In this thesis, expert-based methods have been presented for the monitoring and analysis of multivariate cellular network performance data. These methods allow the analysis of performance bottlenecks having an effect in multiple performance indicators.

In addition, methods for more advanced failure diagnosis have been presented aiming in identification of the causes of the performance bottlenecks. This is important in the analysis of failures having effect on multiple performance indicators in several network elements.

Finally, the use of measurement information in selection of most useful optimization action have been studied. In order to obtain good network performance efficiently, the expected performance of the alternative optimization actions must be possible to evaluate. In this thesis, methods to combine measurement information and application domain models are presented in order to build predictive regression models that can be used to select the optimization actions providing the best network performance.

Algorithms for approximate Bayesian inference with applications to astronomical data analysis

Markus Harva

Dissertation for the degree of Doctor of Science in Technology on 9 May 2008.

External examiners:
Petri Myllymäki (University of Helsinki)
Mark Plumbley (University College of London, United Kingdom)
Opponent:
Manfred Opper (Technical University of Berlin)

Abstract:
Bayesian inference is a theoretically well-founded and conceptually simple approach to data analysis. The computations in practical problems are anything but simple though, and thus approximations are almost always a necessity. The topic of this thesis is approximate Bayesian inference and its applications in three intertwined problem domains.

Variational Bayesian learning is one type of approximate inference. Its main advantage is its computational efficiency compared to the much applied sampling based methods. Its main disadvantage, on the other hand, is the large amount of analytical work required to derive the necessary components for the algorithm. One part of this thesis reports on an effort to automate variational Bayesian learning of a certain class of models.

The second part of the thesis is concerned with heteroscedastic modelling which is synonymous to variance modelling. Heteroscedastic models are particularly suitable for the Bayesian treatment as many of the traditional estimation methods do not produce satisfactory results for them. In the thesis, variance models and algorithms for estimating them are studied in two different contexts: in source separation and in regression.

Astronomical applications constitute the third part of the thesis. Two problems are posed. One is concerned with the separation of stellar subpopulation spectra from observed galaxy spectra; the other is concerned with estimating the time-delays in gravitational lensing. Solutions to both of these problems are presented, which heavily rely on the machinery of approximate inference.

Modeling of mutual dependencies

Arto Klami

Dissertation for the degree of Doctor of Science in Technology on 5 September 2008.

External examiners:
Jukka Corander (Åbo Akademi)
Volker Roth (University of Basel, Switzerland)
Opponent:
Tobias Scheffer (Max Planck Institut for Computer Science, Germany)

Abstract:
Data analysis means applying computational models to analyzing large collections of data, such as video signals, text collections, or measurements of gene activities in human cells. Unsupervised or exploratory data analysis refers to a subtask of data analysis, in which the goal is to find novel knowledge based on only the data. A central challenge in unsupervised data analysis is separating relevant and irrelevant information from each other. In this thesis, novel solutions to focusing on more relevant findings are presented.

Measurement noise is one source of irrelevant information. If we have several measurements of the same objects, the noise can be suppressed by averaging over the measurements. Simple averaging is, however, only possible when the measurements share a common representation. In this thesis, we show how irrelevant information can be suppressed or ignored also in cases where the measurements come from different kinds of sensors or sources, such as video and audio recordings of the same scene.

For combining the measurements, we use mutual dependencies between them. Measures of dependency, such as mutual information, characterize commonalities between two sets of measurements. Two measurements can hence be combined to reduce irrelevant variation by finding new representations for the objects so that the representations are maximally dependent. The combination is optimal, given the assumption that what is in common between the measurements is more relevant than information specific to any one of the sources.

Several practical models for the task are introduced. In particular, novel Bayesian generative models, including a Bayesian version of the classical method of canonical correlation analysis, are given. Bayesian modeling is especially justified approach to learning from small data sets. Hence, generative models can be used to extract dependencies in a more reliable manner in, for example, medical applications, where obtaining a large number of samples is difficult. Also, novel non-Bayesian models are presented: Dependent component analysis finds linear projections which capture more general dependencies than earlier methods.

Mutual dependencies can also be used for supervising traditional unsupervised learning methods. The learning metrics principle describes how a new distance metric focusing on relevant information can be derived based on the dependency between the measurements and a supervising signal. In this thesis, the approximations and optimization methods required for using the learning metrics principle are improved.

Discriminative learning with application to interactive facial image retrieval

Zhirong Yang

Dissertation for the degree of Doctor of Science in Technology on 14 November, 2008.

External examiners:
Sami Brandt (University of Malmö)
Joni-Kristian Kämäräinen (Lappeenranta University of Technology)
Opponent:
Irwin King (The Chinese University of Hong Kong)

Abstract:
The amount of digital images is growing drastically and advanced tools for searching in large image collections are therefore becoming urgently needed. Content-based image retrieval is advantageous for such a task in terms of automatic feature extraction and indexing without human labor and subjectivity in image annotations. The semantic gap between high-level semantics and low-level visual features can be reduced by the relevance feedback technique. However, most existing interactive content-based image retrieval (ICBIR) systems require a substantial amount of human evaluation labor, which leads to the evaluation fatigue problem that heavily restricts the application of ICBIR.

In this thesis a solution based on discriminative learning is presented. It extends an existing ICBIR system, PicSOM, towards practical applications. The enhanced ICBIR system allows users to input partial relevance which includes not only relevance extent but also relevance reason. A multi-phase retrieval with partial relevance can adapt to the user's searching intention in a from-coarse-to-fine manner.

The retrieval performance can be improved by employing supervised learning as a preprocessing step before unsupervised content-based indexing. In this work, Parzen Discriminant Analysis (PDA) is proposed to extract discriminative components from images. PDA regularizes the Informative Discriminant Analysis (IDA) objective with a greatly accelerated optimization algorithm. Moreover, discriminative Self-Organizing Maps trained with resulting features can easily handle fuzzy categorizations.

The proposed techniques have been applied to interactive facial image retrieval. Both a query example and a benchmark simulation study are presented, which indicate that the first image depicting the target subject can be retrieved in a small number of rounds.

Inferring relevance from eye movements with wrong models

Jarkko Salojärvi

Dissertation for the degree of Doctor of Science in Technology on 21 November 2008.

External examiners:
Petri Myllymäki (University of Helsinki)
Guillaume Bouchard (Xerox Research Centre Europe)
Opponent:
Jan Larsen, (Technical University of Denmark)

Abstract:
Statistical inference forms the backbone of modern science. It is often viewed as giving an objective validation for hypotheses or models. Perhaps for this reason the theory of statistical inference is often derived with the assumption that the "truth" is within the model family. However, in many real-world applications the applied statistical models are incorrect. A more appropriate probabilistic model may be computationally too complex, or the problem to be modelled may be so new that there is little prior information to be incorporated. However, in statistical theory the theoretical and practical implications of the incorrectness of the model family are to a large extent unexplored.

This thesis focusses on conditional statistical inference, that is, modeling of classes of future observations given observed data, under the assumption that the model is incorrect. Conditional inference or prediction is one of the main application areas of statistical models which is still lacking a conclusive theoretical justification of Bayesian inference. The main result of the thesis is an axiomatic derivation where, given an incorrect model and assuming that the utility is conditional likelihood, a discriminative posterior yields a distribution on model parameters which best agrees with the utility. The devised discriminative posterior outperforms the classical Bayesian joint likelihood-based approach in conditional inference. Additionally, a theoretically justified expectation maximization-type algorithm is presented for obtaining conditional maximum likelihood point estimates for conditional inference tasks. The convergence of the algorithm is shown to be more stable than in earlier partly heuristic variants.

The practical application field of the thesis is inference of relevance from eye movement signals in an information retrieval setup. It is shown that relevance can be predicted to some extent, and that this information can be exploited in a new kind of task, proactive information retrieval. Besides making it possible to design new kinds of engineering applications, statistical modeling of eye tracking data can also be applied in basic psychological research to make hypotheses of cognitive processes affecting eye movements, which is the second application area of the thesis.

Input variable selection methods for construction of interpretable regression models

Jarkko Tikka

Dissertation for the degree of Doctor of Science in Technology on 12 December 2008.

External examiners:
Michel Verleysen (Université catholique de Louvain)
Patrik. O. Hoyer (University of Helsinki)
Opponent:
Colin Fyfe (University of the West of Scotland, United Kingdom)

Abstract:
Large data sets are collected and analyzed in a variety of research problems. Modern computers allow to measure ever increasing numbers of samples and variables. Automated methods are required for the analysis, since traditional manual approaches are impractical due to the growing amount of data. In the present thesis, numerous computational methods that are based on observed data with subject to modelling assumptions are presented for producing useful knowledge from the data generating system.

Input variable selection methods in both linear and nonlinear function approximation problems are proposed. Variable selection has gained more and more attention in many applications, because it assists in interpretation of the underlying phenomenon. The selected variables highlight the most relevant characteristics of the problem. In addition, the rejection of irrelevant inputs may reduce the training time and improve the prediction accuracy of the model.

Linear models play an important role in data analysis, since they are computationally efficient and they form the basis for many more complicated models. In this work, the estimation of several response variables simultaneously using the linear combinations of the same subset of inputs is especially considered. Input selection methods that are originally designed for a single response variable are extended to the case of multiple responses. The assumption of linearity is not, however, adequate in all problems. Hence, artificial neural networks are applied in the modeling of unknown nonlinear dependencies between the inputs and the response.

The first set of methods includes efficient stepwise selection strategies that assess usefulness of the inputs in the model. Alternatively, the problem of input selection is formulated as an optimization problem. An objective function is minimized with respect to sparsity constraints that encourage selection of the inputs. The trade-off between the prediction accuracy and the number of input variables is adjusted by continuous-valued sparsity parameters.

Results from extensive experiments on both simulated functions and real benchmark data sets are reported. In comparisons with existing variable selection strategies, the proposed methods typically improve the results either by reducing the prediction error or decreasing the number of selected inputs or with respect to both of the previous criteria. The constructed sparse models are also found to produce more accurate predictions than the models including all the input variables.

Advances in independent component analysis and nonnegative matrix factorization

Zhijian Yuan

Dissertation for the degree of Doctor of Science in Technology on 24 April, 2009.

External examiners:
Andrzej Cichocki (RIKEN, Japan)
Patrick O. Hoyer (University of Helsinki)
Opponent:
Fabian Theis (Helmholtz Zentrum München, Germany)

Abstract:
A fundamental problem in machine learning research, as well as in many other disciplines, is finding a suitable representation of multivariate data, i.e. random vectors. For reasons of computational and conceptual simplicity, the representation is often sought as a linear transformation of the original data. In other words, each component of the representation is a linear combination of the original variables. Well-known linear transformation methods include principal component analysis (PCA), factor analysis, and projection pursuit. In this thesis, we consider two popular and widely used techniques: independent component analysis (ICA) and nonnegative matrix factorization (NMF).

ICA is a statistical method in which the goal is to find a linear representation of nongaussian data so that the components are statistically independent, or as independent as possible. Such a representation seems to capture the essential structure of the data in many applications, including feature extraction and signal separation. Starting from ICA, several methods of estimating the latent structure in different problem settings are derived and presented in this thesis. FastICA as one of most efficient and popular ICA algorithms has been reviewed and discussed. Its local and global convergence and statistical behavior have been further studied. A nonnegative FastICA algorithm is also given in this thesis.

Nonnegative matrix factorization is a recently developed technique for finding parts-based, linear representations of non-negative data. It is a method for dimensionality reduction that respects the nonnegativity of the input data while constructing a low-dimensional approximation. The non-negativity constraints make the representation purely additive (allowing no subtractions), in contrast to many other linear representations such as principal component analysis and independent component analysis. A literature survey of Nonnegative matrix factorization is given in this thesis, and a novel method called Projective Nonnegative matrix factorization (P-NMF) and its applications are provided.

Advances in unlimited-vocabulary speech recognition for morphologically rich languages

Teemu Hirsimäki

Dissertation for the degree of Doctor of Science in Technology on 6 August, 2009.

External examiners:
Anssi Klapuri (Tampere University of Technology)
Dilek Hakkani-Tür (International Computer Science Institute, Berkeley, USA)
Opponent:
Steve Renals (University of Edinburgh, United Kingdom)

Abstract:
Automatic speech recognition systems are devices or computer programs that convert human speech into text or make actions based on what is said to the system. Typical applications include dictation, automatic transcription of large audio or video databases, speech-controlled user interfaces, and automated telephone services, for example. If the recognition system is not limited to a certain topic and vocabulary, covering the words in the target languages as well as possible while maintaining a high recognition accuracy becomes an issue.

The conventional way to model the target language, especially in English recognition systems, is to limit the recognition to the most common words of the language. A vocabulary of 60 000 words is usually enough to cover the language adequately for arbitrary topics. On the other hand, in morphologically rich languages, such as Finnish, Estonian and Turkish, long words can be formed by inflecting and compounding, which makes it difficult to cover the language adequately by vocabulary-based approaches.

This thesis deals with methods that can be used to build efficient speech recognition systems for morphologically rich languages. Before training the statistical n-gram language models on a large text corpus, the words in the corpus are automatically segmented into smaller fragments, referred to as morphs. The morphs are then used as modelling units of the n-gram models instead of whole words. This makes it possible to train the model on the whole text corpus without limiting the vocabulary and enables the model to create even unseen words by joining morphs together. Since the segmentation algorithm is unsupervised and data-driven, it can be readily used for many languages.

Speech recognition experiments are made on various Finnish recognition tasks and some of the experiments are also repeated on an Estonian task. It is shown that the morph-based language models reduce recognition errors when compared to word-based models. It seems to be important, however, that the n-gram models are allowed to use long morph contexts, especially if the morphs used by the model are short. This can be achieved by using growing and pruning algorithms to train variable-length n-gram models. The thesis also presents data structures that can be used for representing the variable-length n-gram models efficiently in recognition systems.

By analysing the recognition errors made by Finnish recognition systems it is found out that speaker adaptive training and discriminative training methods help to reduce errors in different situations. The errors are also analysed according to word frequencies and manually defined error classes.

Theses

Master of Science in Technology

2008

Antson, Janne
Modernization of the measurement related data communication network at laboratory of physics and a study on implementing an automated monitoring system for measurement equipment based on the network (Fysiikan laboratorion mittaustietoverkon ajantasaistaminen ja selvitys mittauslaitteistojen automaattisen valvontajärjestelmän toteuttamisesta sen avulla)

Defée, Karolina
Usean tason dataa hyödyntävä laskentahilan tietoturvajärjestelmän prototyyppi; a prototype for a grid intrusion detection system incorporating multilevel data

Dobrinkat, Marcus
Domain adaptation in statistical machine translation systems via user feedback

Engström, Sam
Unsupervised learning of morphology in information retrieval (Ohjaamaton morfologian oppiminen tiedonhaussa)

Ermolov, Andrey
Analysis of differences between metabolic time series with hidden Markov models

Kurimo, Eero
Motion blur and signal noise in low light imaging

Kuusisto, Markus
Tekstuuriin ja rekonstruktioon perustuva liikkeentunnistusmenetelmä videotallennusjärjestelmiä varten

Mäkelä, Rami
Converting the tapio paper machine analyzer (pma) into a digital signal processing system

Niinistö, Ari
Survey of affordable parallel computing systems; katsaus edullisiin rinnakkaislaskentajärjestelmiin

Ollikainen, Marja
Matching medical documents to users; lääketieteellisten dokumenttien sovitus käyttäjille

Parkkinen, Juuso
Generative Probabilistic Models of Biological and Social Network Data .

Pitkänen, Sampo
Optimal reception of 64 quadrature amplitude modulation in high-speed downlink packet access (64-QAM-signaalin optimoitu vastaanottomenetelmä HSDPA:ssa)

Ramkumar, Pavan
Modeling the dynamics of human neuromagnetic brain rhythms

2009

De Alba Rivera, Luis Gabriel
Modeling and profiling people's way of living: A data mining approach to a health survey.

Eirola, Emil
Variable Selection with the Delta Test in Theory and Practice.

Itkonen, Sami
Building Business Intelligence Infrastructure.

Kallasjoki, Heikki
Methods for spectral envelope estimation in noise robust speech recognition.

Kantonen, Tuomas Kalevi
Augmented Collaboration in Mixed Environments.

Kaunisto, Antti
Video quality comparison of MPEG-1/2, MPEG-4 and H.264 codecs in the G-cluster Game-on-Demand Environment.

Keraudy, Stevan
Histogram equalization for noise robust speech recognition.

Kozma, László
Proactive interface for image retrieval.

Lampi, Golan
Self-Organizing Maps in Decision Support: a Decision Support System Prototype.

Luttinen, Jaakko
Gaussian-process factor analysis for modeling spatiotemporal data.

Mäkelä, Tuomas
Dental x-ray image stitching algorithm.

Nybo, Kristian
Graph visualization using latent variable models.

Rossi, Pekka
Life Cycle Analysis of Convective Cells through Image Processing and Data Fusion.

Ruokolainen, Teemu
Topic adaptation for speech recognition in multimodal environment.

Sovilj, Dušan
Fast variable selection using delta test.

Suvitaival, Tommi
Bayesian Two-Way Analysis of High-Dimensional Collinear Metabolomics Data.

Toiviainen, Maunu
Near-infrated spectroscopy of solids; Chemometrics and signal processing.

Venesmaa, Klaus
Influence of standards IEEE-754 and IEEE-854 and floating point's precision to data mining and machine learning methods (Standardien IEEE-754 ja IEEE-854 sekä liukuluvun tarkkuuden vaikutus tiedon louhinnan ja koneoppimisen menetelmiin).

Research Projects

Chapter 1

Introduction

The Centre of Excellence called the Adaptive Informatics Research Centre (AIRC) started in January 2006 in the Laboratory of Computer and Information Science at Helsinki University of Technology. It followed the tradition of the Neural Networks Research Centre (NNRC), operative from 1994 to 2005, also under the national Centre of Excellence status.

The core function and strength of our research centre is the ability to analyze and process extensive data sets coming from a number of application fields using our own innovative and generic methods. Our research has concentrated on neurocomputing and statistical machine learning algorithms, with a number of applications. In the algorithmic research, we have attained a world class status especially in such unsupervised machine learning methods as the Self-Organizing Map and Independent Component Analysis.

Building on this solid methodological foundation, we have started to apply the knowledge, expertise and tools to advance knowledge in other domains and disciplines. In the AIRC, we take a goal-oriented, ambitious, and interdisciplinary approach in targeting at the adaptive informatics problem. By adaptive informatics we mean a field of research where automated learning algorithms are used to discover the relevant informative concepts, components, and their mutual relations from large amounts of data. Access to the ever-increasing amounts of available data and its transformation to forms intelligible for the human user is one of the grand challenges in the near future.

The AIRC Centre of Excellence focuses on several adaptive informatics problems. One is the efficient retrieval and processing techniques for text, digital audio and video, and numerical data such as biological and medical measurements, which will create valuable information sources. Another problem area are advanced multimodal natural interfaces. We are building systems that process multimodal contextual information including spoken and written language, images, videos, and explicit and implicit user feedback. Automated semantic processing of such information will facilitate cost-effective knowledge acquisition and knowledge translation without the need to build the descriptions manually. Yet another problem, which we approach together with experts in brain science and molecular biology, is to develop and apply our algorithmic methods to problems in neuroinformatics and bioinformatics.

The Adaptive Informatics methodology that we focus on is to build empirical models of the data by using automated machine learning techniques, in order to make the information usable. The deep expertise on the algorithmic methods, gained over the years, is used to build realistic solutions, starting from the problem requirements. The application domains have been chosen because of our acquired knowledge in some of their core problems, because of their strategic importance in the near future, and because of their mutual interrelations. The algorithms are based on our own core expertise. Future research will continue to be novel, innovative, as well as inter- and multi-disciplinary, with a specific

focus on shared research activities that will have a significant societal impact.

The AIRC Centre of Excellence consists of five interrelated research groups: Algorithms and Methods, Bioinformatics and Neuroinformatics, Multimodal Interfaces, Computational Cognitive Systems, and Adaptive Informatics Applications (see Figure 1).

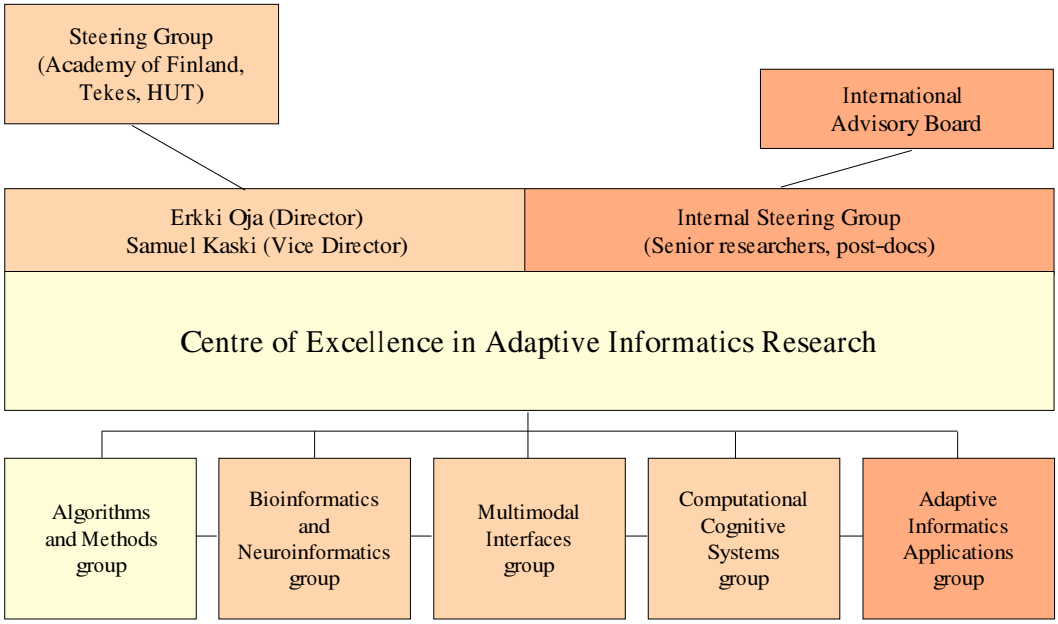


Figure 1.1: *The organization of the AIRC Centre of Excellence*

The Algorithms and Methods group conducts basic algorithmic research in adaptive informatics that relies heavily on computer science, mathematics and statistics, and is partly motivated by the research problems of other groups. In contrast, the groups of Bioinformatics and Neuroinformatics, Multimodal Interfaces and Computational Cognitive Systems form an interdisciplinary research network with shared research interests in life and human sciences. The group of Adaptive Informatics Applications brings the research results into practice together with collaborating enterprises. This inter- and multi-disciplinary diversity facilitates a rich exchange of ideas, knowledge and expertise both within and between research groups. The ideas generated in one research group spark innovative ideas and research methods in other groups. This ability to pool knowledge and resources between groups reduces duplication, saves time, and generates more powerful research methods and results. Altogether, it makes the Centre of Excellence a coherent whole.

Each group has a wide range of national and international collaborators both in Academia and industry. Researcher training, graduate studies, and promotion of creative research is strongly emphasized, following the successful existing traditions.

The present Biennial Report 2008 - 2009 details the individual research projects of the five groups during the middle two years of the six-year period of the AIRC. Additional information including demos etc. is available from our Web pages, `www.cis.hut.fi/research`.

Algorithms and Methods

Chapter 2

Bayesian learning of latent variable models

Juha Karhunen, Antti Honkela, Tapani Raiko, Alexander Ilin, Koen Van Leemput, Jaakko Luttinen, Matti Tornio, Markus Harva

2.1 Bayesian modeling and variational learning

Unsupervised learning methods are often based on a generative approach where the goal is to find a latent variable model which explains how the observations were generated. It is assumed that there exist certain latent variables (also called in different contexts source signals, factors, or hidden variables) which have generated the observed data through an unknown mapping. The goal of generative learning is to identify both the latent variables and the unknown generative mapping.

The success of a specific model depends on how well it captures the structure of the phenomena underlying the observations. Various linear models have been popular, because their mathematical treatment is fairly easy. However, in many realistic cases the observations have been generated by a nonlinear process. Unsupervised learning of a nonlinear model is a challenging task, because it is typically computationally much more demanding than for linear models, and flexible models require strong regularization for avoiding overfitting.

In Bayesian data analysis and estimation methods, all the uncertain quantities are modeled in terms of their joint probability distribution. The key principle is to construct the joint posterior distribution for all the unknown quantities in a model, given the data sample. This posterior distribution contains all the relevant information on the parameters to be estimated in parametric models, or the predictions in non-parametric prediction or classification tasks [1, 2].

Denote by $\mathcal{H}$ the particular model under consideration, and by $\boldsymbol{\theta}$ the set of model parameters that we wish to infer from a given data set X. The posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$ of the parameters given the data X and the model $\mathcal{H}$ can be computed from the Bayes' rule

$$p(\boldsymbol{\theta}|X, \mathcal{H}) = \frac{p(X|\boldsymbol{\theta}, \mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(X|\mathcal{H})} \tag{2.1}$$

Here $p(X|\boldsymbol{\theta}, \mathcal{H})$ is the likelihood of the parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{H})$ is the prior pdf of the parameters, and $p(X|\mathcal{H})$ is a normalizing constant. The term $\mathcal{H}$ denotes all the assumptions made in defining the model, such as the choice of a particular model class and structure, specific noise model, etc.

The parameters $\boldsymbol{\theta}$ of a particular model $\mathcal{H}_i$ are often estimated by seeking the peak value of a probability distribution. The non-Bayesian maximum likelihood (ML) method uses to this end the distribution $p(X|\boldsymbol{\theta}, \mathcal{H})$ of the data, and the Bayesian maximum a posteriori (MAP) method finds the parameter values that maximize the posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$. However, using point estimates provided by the ML or MAP methods is often problematic, because the model order estimation and overfitting (choosing too complicated a model for the given data) are severe problems [1, 2].

Instead of searching for some point estimates, the correct Bayesian procedure is to use all possible models to evaluate predictions and weight them by the respective posterior probabilities of the models. This means that the predictions will be sensitive to regions where the probability mass is large instead of being sensitive to high values of the probability density [3, 2]. This procedure optimally solves the issues related to the model complexity and choice of a specific model $\mathcal{H}_i$ among several candidates. In practice, however, the differences between the probabilities of candidate model structures are often very large, and hence it is sufficient to select the most probable model and use the estimates or predictions given by it.

A problem with fully Bayesian estimation is that the posterior distribution (2.1) has a highly complicated form except for in the simplest problems. Therefore it is too difficult to handle exactly, and some approximative method must be used. Variational methods

form a class of approximations where the exact posterior is approximated with a simpler distribution [4, 2]. In a method commonly known as *Variational Bayes (VB)* [1, 3, 2] the misfit of the approximation is measured by the Kullback-Leibler (KL) divergence between two probability distributions $q(v)$ and $p(v)$. The KL divergence is defined by

$$D(q \parallel p) = \int q(v) \ln \frac{q(v)}{p(v)} dv \tag{2.2}$$

which measures the difference in the probability mass between the densities $q(v)$ and $p(v)$.

A key idea in the VB method is to minimize the misfit between the actual posterior pdf and its parametric approximation using the KL divergence. The approximating density is often taken a diagonal multivariate Gaussian density, because the computations become then tractable. Even this crude approximation is adequate for finding the region where the mass of the actual posterior density is concentrated. The mean values of the Gaussian approximation provide reasonably good point estimates of the unknown parameters, and the respective variances measure the reliability of these estimates.

A main motivation of using VB is that it avoids overfitting which would be a difficult problem if ML or MAP estimates were used. VB method allows one to select a model having appropriate complexity, making often possible to infer the correct number of latent variables or sources. It has provided good estimation results in the very difficult unsupervised (blind) learning problems that we have considered.

Variational Bayes is closely related to information theoretic approaches which minimize the description length of the data, because the description length is defined to be the negative logarithm of the probability. Minimal description length thus means maximal probability. In the probabilistic framework, we try to find the latent variables or sources and the nonlinear mapping which most probably correspond to the observed data. In the information theoretic framework, this corresponds to finding the latent variables or sources and the mapping that can generate the observed data and have the minimum total complexity. This information theoretic view also provides insights to many aspects of learning and helps to explain several common problems [5].

In the following subsections, we first discuss a natural conjugate gradient algorithm which speeds up learning remarkably compared with compared alternative popular algorithms. After this we consider variational Bayesian learning of nonlinear state-space models, which are applied to model predictive control. This is followed by extensions of probabilistic principal component analysis (PCA) to binary PCA, missing values and achieving robustness in the presence of outliers. We then consider time series modeling in bioinformatics to learn gene regulatory relationships from time series expression data, as well as climate data analysis using Gaussian processes. We have also applied Bayesian methods to the astronomical data analysis problem of estimating time delays in gravitational lensing, as well as to medical image computing, focusing there on model-based segmentation and registration of magnetic resonance images of the brain. In most of these topics, we used variational approximations.

2.2 Algorithmic improvements for variational inference

Natural conjugate gradient

Variational methods for approximate inference in machine learning often adapt a parametric probability distribution to optimize a given objective function. This view is especially useful when applying variational Bayes (VB) to models outside the conjugate-exponential family. For them, variational Bayesian expectation maximization (VB EM) algorithms are not easily available, and gradient-based methods are often used as alternatives.

In previous machine learning algorithms based on natural gradients [6], the aim has been to use maximum likelihood to directly update the model parameters $\boldsymbol{\theta}$ taking into account the geometry imposed by the predictive distribution for data $p(\boldsymbol{X}|\boldsymbol{\theta})$. The resulting geometry is often very complicated as the effects of different parameters cannot be separated and the Fisher information matrix is relatively dense.

Recently, in [7], we propose using natural gradients for free energy minimisation in variational Bayesian learning using the simpler geometry of the approximating distributions $q(\boldsymbol{\theta}|\boldsymbol{\xi})$. Because the approximations are often chosen to minimize dependencies between different parameters $\boldsymbol{\theta}$, the resulting Fisher information matrix with respect to the variational parameters $\boldsymbol{\xi}$ will be mostly diagonal and hence easy to invert.

While taking into account the structure of the approximation, plain natural gradient in this case ignores the structure of the model and the global geometry of the parameters $\boldsymbol{\theta}$. This can be addressed by using conjugate gradients. Combining the natural gradient search direction with a conjugate gradient method yields our proposed *natural conjugate gradient (NCG)* method, which can also be seen as an approximation to the fully Riemannian conjugate gradient method.

The NCG algorithm was compared against conjugate gradient (CG) and natural gradient (NG) algorithms in learning a nonlinear state-space model [8]. The results for a number of datasets ranging from 200 to 500 samples of 21 dimensional speech spectrograms can be seen in Figure 2.1. The plain CG and NG methods were clearly slower than others and the maximum runtime of 24 hours was reached by most CG and some NG runs. NCG was clearly the fastest algorithm with the older heuristic method of [8] between these extremes. The results with a larger data set are very similar with NCG outperforming all alternatives by a factor of more than 10.

The experiments in [7] show that the natural conjugate gradient method outperforms both conjugate gradient and natural gradient methods by a large margin. Considering univariate Gaussian distributions, the regular gradient is too strong for model variables with small posterior variance and too weak for variables with large posterior variance. The posterior variance of latent variables is often much larger than the posterior variance of model parameters and the natural gradient takes this into account in a very natural manner.

Transformation of latent variables

Variational methods have been used for learning linear latent variable models in which observed data vectors $\mathbf{x}(t)$ are modeled as linear combination of latent variables $\mathbf{s}(t)$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \boldsymbol{\mu} + \mathbf{n}(t), \quad t = 1, \ldots, N. \tag{2.3}$$

The latent variables are assigned some prior distributions, such as zero-mean Gaussian priors with uncorrelated components in the basic factor analysis model. When VB learning is used, the true posterior probability density function (pdf) of the unknown variables is

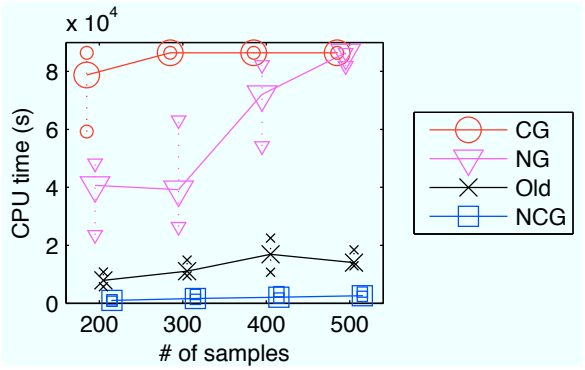


Figure 2.1: Convergence speed of the natural conjugate gradient (NCG), the natural gradient (NG) and the conjugate gradient (CG) methods as well as the heuristic algorithm (Old) with different data sizes. The lines show median times with 25 % and 75 % quantiles shown by the smaller marks. The times were limited to at most 24 hours, which was reached by a number of simulations.

approximated using a tractable pdf factorized as follows:

$$p\left(\boldsymbol{\mu}, \mathbf{A}, \mathbf{s}(1), ..., \mathbf{s}(N) \mid \{\mathbf{x}(t)\}\right) \approx q(\boldsymbol{\mu})q(\mathbf{A})q(\mathbf{s}(1)) \ldots q(\mathbf{s}(N)) \,.$$

This form of the posterior approximation q ignores the strong correlations present between the variables, which often causes slow convergence of VB learning.

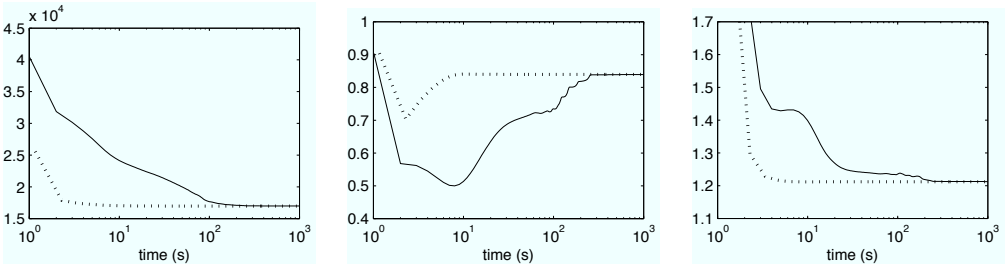
Parameter-expanded VB (PX-VB) methods were recently proposed to address the slow convergence problem [9]. The general idea is to use auxiliary parameters in the original model to reduce the effect of strong couplings between different variables. The auxiliary parameters are optimized during learning, which corresponds to *joint* optimization of different components of the variational approximation of the true posterior. In this way strong functional couplings between the components are reduced and faster convergence is facilitated. One of the main challenges for applying the PX-VB methodology is to use proper reparameterization of the original model.

In our recent conference paper [10], we present a similar idea in the context of VB learning of factor analysis models. There we use auxiliary parameters $\mathbf{b}$ and $\mathbf{R}$ which translate and rotate the latent variables:

$$\begin{aligned} \mathbf{s}(t) &\leftarrow \mathbf{s}(t) - \mathbf{b} & \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} + \mathbf{A}\mathbf{b} \\ \mathbf{s}(t) &\leftarrow \mathbf{R}\mathbf{s}(t) & \mathbf{A} &\leftarrow \mathbf{A}\mathbf{R}^{-1} \,. \end{aligned}$$

The optimal parameters $\mathbf{b}$ and $\mathbf{R}$ which minimize the misfit between the posterior pdf and its approximation can then be computed analytically. This corresponds to joint optimization of factors $q(\mathbf{s}(t))$. In our paper, we show that the proposed transformations essentially perform centering and whitening of the hidden factors taking into account their posterior uncertainties.

We tested the effect of the proposed transformations by applying the VB PCA model to an artificial dataset consisting of $N = 200$ samples of normally distributed 50-dimensional vectors $\mathbf{x}(t)$. Figure 2.2 shows the minimized VB cost and the root mean squared error (RMSE) computed on the training and test sets during learning. The curves indicate that the method first overfits providing a solution with an unreasonably small RMSE. Later, learning proceeds toward a better solution yielding smaller test RMSE. Note that using



(a) VB cost (b) Training RMSE (c) Test RMSE

Figure 2.2: Convergence of VB PCA tested on artificial data. The dotted and solid curves represent the results with and without the proposed transformations, respectively.

the proposed transformations reduced the overfitting effect at the beginning of learning, which led to faster convergence to the optimal solution.

2.3 Nonlinear state-space models for model-predictive control

In many cases, measurements originate from a dynamical system and form a time series. In such instances, it is often useful to model the dynamics in addition to the instantaneous observations. We have used rather general nonlinear models for both the data (observations) and dynamics of the sources (latent variables) [8]. This results in a state-space model where the sources can be interpreted as the internal state of the underlying generative process.

The general form of our nonlinear model for the generative mapping from the source (latent variable) vector $\mathbf{s}(t)$ to the data (observation) vector $\mathbf{x}(t)$ at time t is

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t)\,. \tag{2.4}$$

The dynamics of the sources can be modelled by another nonlinear mapping, which leads to a source model [8]

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1), \boldsymbol{\theta}_g) + \mathbf{m}(t)\,, \tag{2.5}$$

where $\mathbf{s}(t)$ are the sources (states) at time t, $\mathbf{m}$ is the Gaussian noise, and $\mathbf{g}(\cdot)$ is a vector containing as its elements the nonlinear functions modelling the dynamics.

The nonlinear functions are modelled by MLP networks. Since the states in dynamical systems are often slowly changing, the MLP network for mapping $\mathbf{g}$ models the change in the value of the source:

$$\mathbf{g}(\mathbf{s}(\mathbf{t}-\mathbf{1})) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d}\,. \tag{2.6}$$

The dynamic mapping $\mathbf{g}$ is thus parameterized by the matrices $\mathbf{C}$ and $\mathbf{D}$ and bias vectors $\mathbf{c}$ and $\mathbf{d}$.

Estimation of the arising state-space model is rather involved, and it is discussed in detail in our earlier paper [8]. An important advantage of the proposed nonlinear state-space method (NSSM) is its ability to learn a high-dimensional latent source space. We have also reasonably solved computational and over-fitting problems which have been major obstacles in developing this kind of unsupervised methods thus far. Potential applications for our method include prediction and process monitoring, control and identification. MATLAB software package is available under the name nonlinear dynamical factor analysis on the home page of our Bayes group [11].

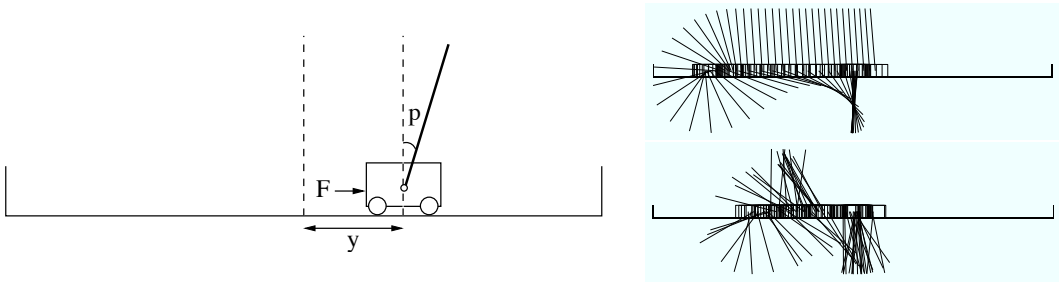


Figure 2.3: Left: The cart-pole system. The goal is to swing the pole to an upward position and stabilize it without hitting the walls. The cart can be controlled by applying a force to it. Top left: The pole is successfully swinged up by moving first to the left and then right. Bottom right: Our controller works quite reliably even in the presence of serious observation noise.

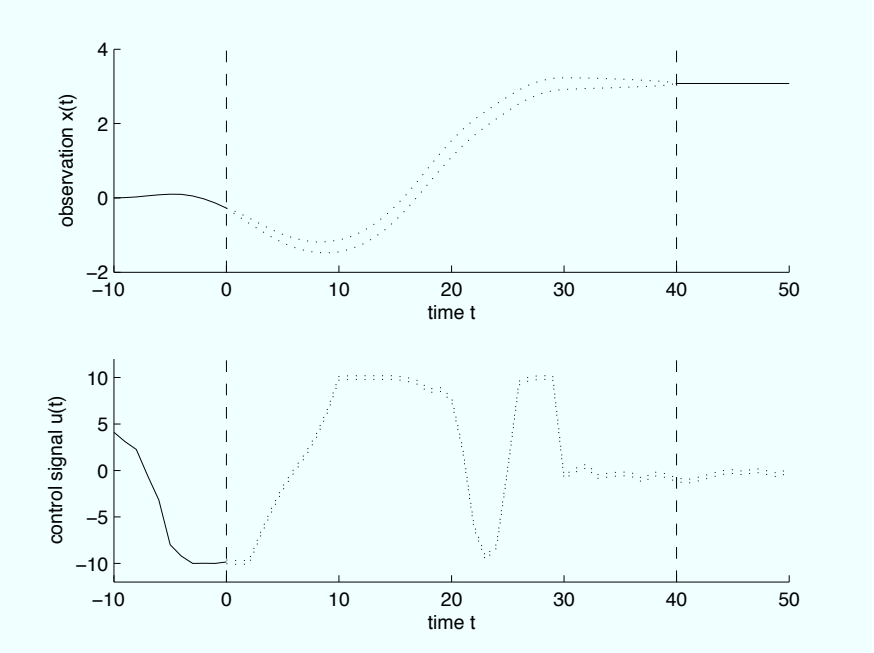


Figure 2.4: Optimistic inference control is a novel way of doing model predictive control, where we assume that the goal state has been reached after some window of uncertainty. The hidden states, observations and control signals are inferred using Bayesian inference methods. This approach bridges the gap between model-predictive control and Bayesian inference and thus algorithic developments on one side can be applied on the other side. The inferred observations and control signals are plotted with confidence intervals. The current time is $t_0 = 0$ and after time $t_0 + T_c = 40$, the observation $\mathbf{x}(t)$ is assumed to be at the desired level $\mathbf{r}(t)$.

In [15], we studied such a system combining variational Bayesian learning of an unknown dynamical system with nonlinear model-predictive control. For being able to control the dynamical system, control inputs are added to the nonlinear state-space model as part of the hidden state. Then we can use stochastic nonlinear model-predictive control, which is based on optimising control signals based on maximising a utility function. Figure 2.3 shows a simulation with an alternative method for model-predictive control.

The results with a simulated cart-pole swing-up task confirm that selecting actions based on a state-space model instead of the observation directly has many benefits: First, it is more resistant to noise because it implicitly involves filtering. Second, the observations (without history) do not always carry enough information about the system state. Third, when nonlinear dynamics are modelled by a function approximator such as an multilayer perceptron network, a state-space model can find such a representation of the state that it is more suitable for the approximation and thus more predictable.

2.4 Extensions of probabilistic PCA

PCA of large-scale datasets with many missing values

Principal component analysis (PCA) is a classical data analysis technique. Some algorithms for PCA scale better than others to problems with high dimensionality. They also differ in the ability to handle missing values in the data. In our recent papers [16, 17], a case is studied where the data are high-dimensional and a majority of the values are missing. In the case of very sparse data, overfitting becomes a severe problem even in simple linear models such as PCA. Regularization can be provided using the Bayesian approach by introducing prior for the model parameters. The PCA model can then be identified using, for example, maximum a posteriori estimates (MAPPCA) or variational Bayesian (VBPCA) learning.

In [16, 17], we study different approaches to PCA for incomplete data. We show that faster convergence can be achieved using the following rule for the model parameters:

$$\theta_i \leftarrow \theta_i - \gamma \left(\frac{\partial^2 C}{\partial \theta_i^2} \right)^{-\alpha} \frac{\partial C}{\partial \theta_i} ,$$

where α is a control parameter that allows the learning algorithm to vary from the standard gradient descent ($\alpha = 0$) to the diagonal Newton's method ($\alpha = 1$). These learning rules can be used for standard PCA learning and extended to MAPPCA and VBPCA.

The algorithms were tested on the Netflix problem (`http://www.netflixprize.com/`), which is a task of predicting preferences (or producing personal recommendations) by using other people's preferences. The Netflix problem consists of movie ratings given by 480189 customers to 17770 movies. There are 100480507 ratings from 1 to 5 given, and the task is to predict 2817131 other ratings among the same group of customers and movies. 1408395 of the ratings are reserved for validation. Thus, 98.8% of the values are missing.

We used different variants of PCA in order to predict the test ratings in the Netflix data set. The obtained results are shown in Figure 2.5. The best accuracy was obtained using VB PCA with a simplified form of the posterior approximation (VBPCAd in Figure 2.5). That method was also able to provide reasonable estimates of the uncertainties of the predictions.

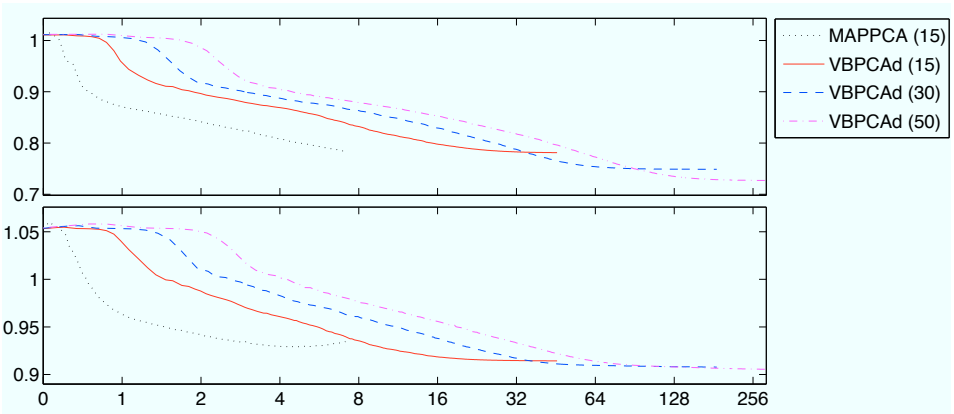


Figure 2.5: Root mean squared errors for the Netflix data (y-axis) plotted against the processor time in hours. The upper plot shows the training error while the lower plot shows the error for the probing data provided by Netflix. The time scale is linear from 0 to 1 and logarithmic above 1.

Binary PCA for collaborative filtering

In [18], we proposed an algorithm for binary principal component analysis that scales well to very high dimensional and very sparse data. Binary PCA finds components from data assuming Bernoulli distributions for the observations. The probabilistic approach allows for straightforward treatment of missing values.

We applied the proposed method to the same collaborative filtering problem prepared by Netflix. The collected ratings can be represented in the form of a matrix $\mathbf{X}$ in which each column contains ratings given by one user and each row contains ratings given to one movie. As a preprocessing step, the ratings were encoded with binary values, according to the following scheme:

$$\begin{aligned} 1 &\rightarrow 0000 \\ 2 &\rightarrow 0001 \\ 3 &\rightarrow 0011 \\ 4 &\rightarrow 0111 \\ 5 &\rightarrow 1111 \end{aligned}$$

With this scheme, each element in the data tells whether a rating is greater or smaller than a particular threshold.

We model the probability of each element x_{ij} of $\mathbf{X}$ to be 1 using the following formula:

$$P(x_{ij} = 1) = \sigma(\mathbf{a}_i^T \mathbf{s}_j) \tag{2.7}$$

where $\mathbf{a}_i$ and $\mathbf{s}_j$ are parameter vectors (both contain c elements) corresponding to the i-th movie and j-th user, respectively. The parameters $\mathbf{a}_i$ and $\mathbf{s}_j$ are assigned Gaussian priors and they are estimated from on the available ratings using the MAP method.

The results with the proposed binary PCA algorithm are slighlty worse than the ones obtained with PCA. However, by blending the two approaches, we were able to improve our previously best results obtained with PCA alone [16, 17]. Figure 2.6 shows the predictions of binary PCA against traditional PCA on a smaller MovieLens data set. The difference between the predictions suggests that the two methods model the data differently and blending them can improve the overall prediction performance.

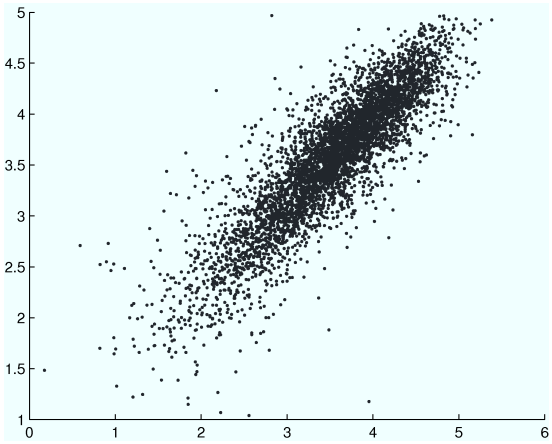


Figure 2.6: Predictions on a test set from the MovieLens data using PCA (x-axis) and the binary PCA model (y-axis). Note that PCA gives predictions outside the allowed range 1 to 5, whereas the predictions of binary PCA fall between 1 and 5 by construct.

Robust PCA for incomplete data

Standard PCA is known to be sensitive to outliers in the data because it is based on minimisation of a quadratic criterion such as the mean-square representation error. Thus, corrupted or atypical observations may cause the failure of PCA, especially for data sets with missing values. A standard way to cope with this problem is replacing the quadratic cost function of PCA a function which grows more slowly.

In [19], we present a new robust PCA model based on the Student-t distribution and show how it can be identified for data sets with missing values. We make the assumption that the outliers can arise independently in each sensor (i.e. for each dimension of a data vector). This assumption is different to the previously introduced techniques [21] and it turns out to be important for modeling incomplete data sets. The proposed model can improve the quality of the principal subspace estimation and provide better reconstructions of missing values. The model can also be used to remove outliers by estimating the true values of their corrupted components from the uncorrupted ones.

We tested the robust PCA model on the Helsinki Testbed data set which at the moment of our studies contained many atypical measurements and missing values. The model was used to estimate four principal components of the temperature measurements from 79 stations in Southern Finland. Figure 2.7 presents the reconstruction of the data using our robust PCA model for four different stations. The reconstructions look very reasonable with most of the outliers being removed.

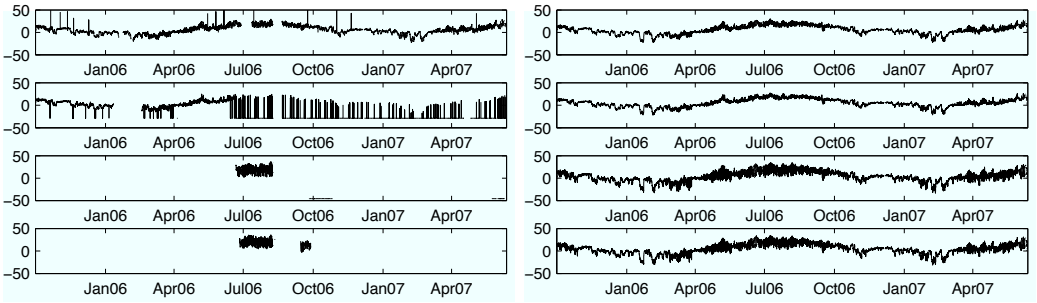


Figure 2.7: Four example signals from the Helsinki Testbed dataset and their reconstructions using the proposed robust PCA.

2.5 Time-series modelling in bioinformatics

Bayesian methods are well-suited for analysis of molecular biology data as the data sets practically always consist of very few samples with a high noise level. We have studied models of gene transcription regulation based on time series gene expression data in collaboration with the Machine Learning and Optimisation group at the University of Manchester. This is a very challenging modelling task as the time series are very short, typically at most a dozen time points.

In [22], we have developed a method of modelling single input motif systems, where a single transcription factor regulates a number of genes. This is achieved by imposing a Gaussian process prior on the latent regulator (transcription factor protein) activity, which under a linear ODE transcription model leads to a joint Gaussian process model for all observable gene expression values. The model can further be extended by incorporating the transcription factor expression levels through a translation model. It is also possible to consider nonlinear models by using approximate inference. A sample model of p53 activation is illustrated in Fig. 2.8.

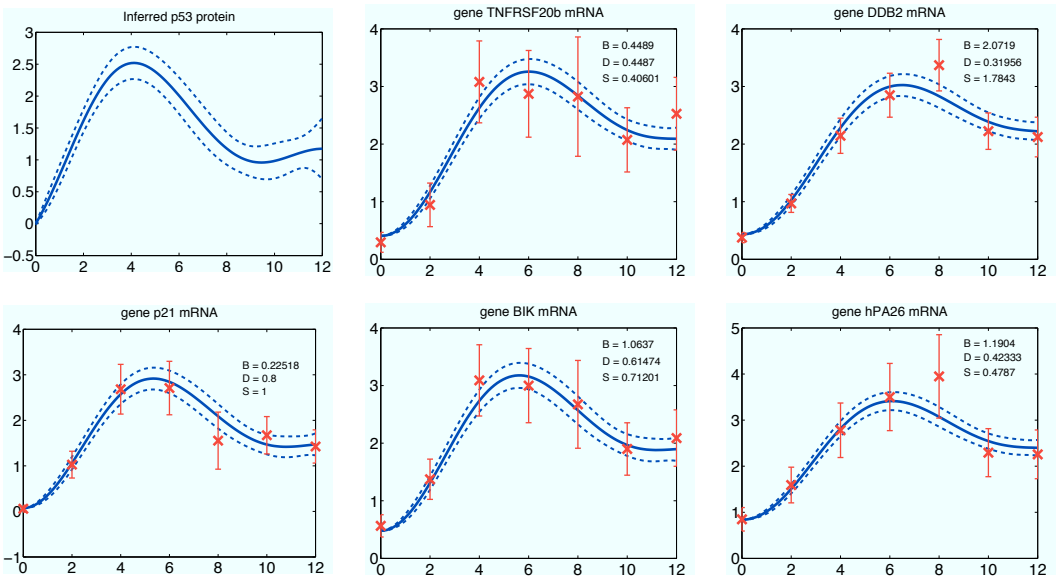


Figure 2.8: An inferred model of transcription factor p53 activation based on five known target genes. Red marks denote observed gene expression values while blue curves are inferred by the model along with 2 standard deviation error bars.

We have applied the model to genome-wide ranking of potential target genes of transcription factors. In experiments with key regulators of *Drosophila* mesoderm and muscle development, this has lead to extremely promising results in terms of enrichment of differential expression in loss-of-function mutants as well as ChIP-chip binding near the predicted target genes [23].

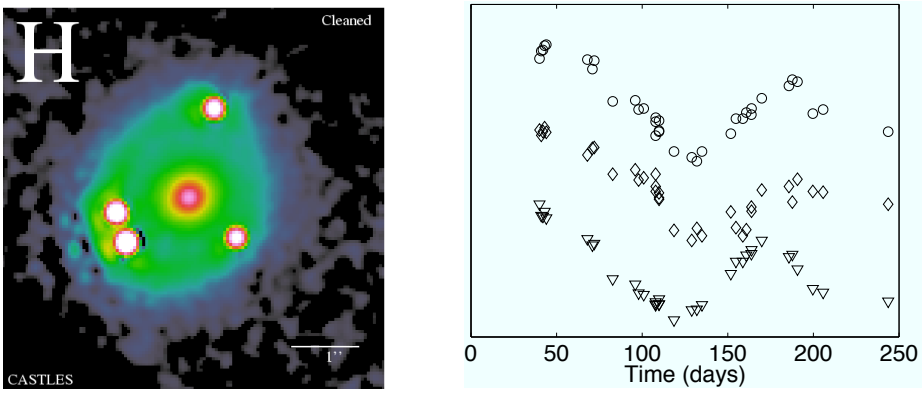


Figure 2.9: Left: The four images of PG1115+080. Right: The corresponding intensity measurements (the two images closest to each other are merged).

2.6 Estimation of time delays in gravitational lensing in astronomy

Most of the research topics contained in Markus Harva's doctoral thesis [24] which appeared in 2008 have already been described in our earlier biennial reports under their chapters on Bayesian learning of latent variable models. However, the journal paper [27] on estimation of time delays in gravitational lensing was published in 2008, and therefore we discuss that work here.

Gravitational lensing occurs when the light coming from a distant bright source is bent by the gravitational potential of an intermediate galaxy such that several images of the source are observed (see the left panel of Figure 2.9 for an example system). Relativistic effects and the different lengths of the paths affect the time it takes for the photons originating from the source to travel to the observer. This is perceived as a delay in the intensity variations between the images (see the right panel of Figure 2.9). The significance of estimating the delays in such systems stems from the early observation that they can be used in determining important cosmological quantities [25].

The delay estimation problem is difficult for various reasons. The main challenge is the uneven sampling rate, as the sampling times are determined by factors one cannot control such as observing conditions and scheduling. The signal-to-noise ratio in the observations is often poor too, although this varies somewhat between datasets. Classical delay estimation methods usually rely on the cross-correlation function which is easy to evaluate between regularly sampled signals. The obvious way to attack the problem with unevenly sampled signals would then be to interpolate them appropriately to obtain evenly sampled signals and then apply the cross correlation method. But with all the gaps and the noise in the data, the interpolation can introduce spurious features to the data which make the cross-correlation analysis go awry [26].

In [27], a method for estimating the delay between irregularly sampled signals is presented. Since interpolation on the gappy and noisy data can be venturesome, that is avoided. Instead the two observed signals, $x_1(t)$ and $x_2(t)$, are postulated to have been emitted from the same latent source signal $s(t)$, the observation times being determined by the actual sampling times and the delay. The source is then assumed to follow the Wiener process: $s(t_{i+1}) - s(t_i) \sim N(0, [(t_{i+1} - t_i)\,\sigma]^2)$. This prior encodes the notion of "slow variability" into the model which is an assumption implicitly present in many of the other methods as well. The model is estimated using exact marginalization, which leads

to a specific type of Kalman-filter, combined with the Metropolis-Hastings algorithm.

We have used the proposed method to determine the delays in several gravitational lensing systems. Controlled comparisons against other methods cannot, however, be done with real data as the true delays are unknown to us. Instead, artificial data, where the ground truth is known, must be used. Figure 2.10 shows the performance of several methods in an artificial setting.

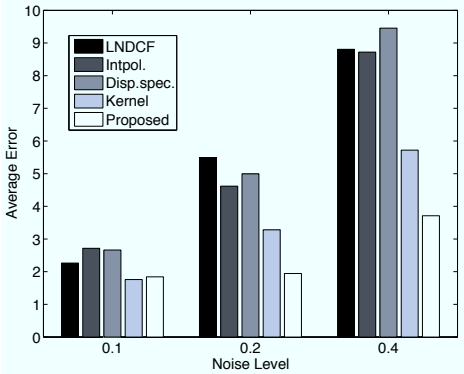


Figure 2.10: Average errors of the methods for three groups of datasets.

2.7 Automated segmentation of brain MR images

Many studies in basis neuroscience and neurological and psychiatric diseases benefit from fully-automated techniques that are able to reliably assign a neuroanatomical label to each voxel in magnetic resonance (MR) images of the brain. In order to cope with the complex anatomy of the human brain, the large overlap in intensity characteristics between structures of interest, and the dependency of MR intensities on the acquisition sequence used, state-of-the-art brain MR labeling techniques rely on prior information extracted from a collection of manually labeled training datasets. Typically, this prior information is represented in the form of *probabilistic atlases*, constructed by first aligning the training datasets together using linear spatial transformations, and then calculating the probability of each voxel being occupied by a particular structure as the relative frequency that structure occurred at that voxel across the training datasets.

While these "average" atlases are intuitive and straightforward to compute, they are not necessarily the best way to extract population-wise statistics from the training data. Atlases built from a limited number of training images tend to generalize poorly to subjects not included in the training database, necessitating heuristic approaches such as spatially blurring atlases used in automated segmentation algorithms. Another problem is that such atlases do not include non-linear deformations aligning corresponding structures across subjects, although this would be a natural way to model anatomical variations.

In [31], we took a critical look at the generative model implicitly underlying probabilistic brain atlases, and proposed to generalize it using tetrahedral mesh-based representations endowed with explicit deformation models. We demonstrated how Bayesian inference can be used to automatically learn the optimal properties of the resulting atlases from a set of manual example segmentations in MR images of training subjects. The learning involves maximizing the probability with which an atlas model would generate the example segmentations, or, equivalently, minimizing the number of bits needed to encode them. This procedure automatically yields sparse atlas representations that explicitly avoid overfitting to the training data, and are therefore better at predicting the neuroanatomy in new subjects than conventional probabilistic atlases [31]. An example of an optimal mesh-based atlas, built from manual annotations of 36 neuroanatomical structures in four individuals, is shown in figure 2.11.

In subsequent work aiming at automatically delineating the subregions of the hippocampus from very high resolution MR images [32, 36, 35], we supplemented the prior distribution provided by a mesh-based atlas, which models the generation of images where each voxel is assigned a unique neuroanatomical label, with a parametric likelihood distribution that predicts how such label images translate into MR images, where each voxel has an intensity. Together these distributions form a complete generative model of MR images that we then used to obtain fully automated structural measurements in a Bayesian fashion, using concepts from our earlier work [28, 29]. In particular, we estimated how the position of the nodes of the atlas mesh are optimally warped onto an image under study, while simultaneously inferring the parameters of the likelihood distribution. Figure 2.12 shows an example of a fully-automated segmentation of the subregions of the hippocampus computed using this approach.

Additional joint work in brain MR analysis we contributed to during the years 2008-2009 include group-wise segmentation of collections of images for which no manual training data is available [38, 41], non-parametric Bayesian whole-brain parcellation [39, 40] and information theoretical image alignment [37], as well as a number of clinical research papers [30, 33, 34].

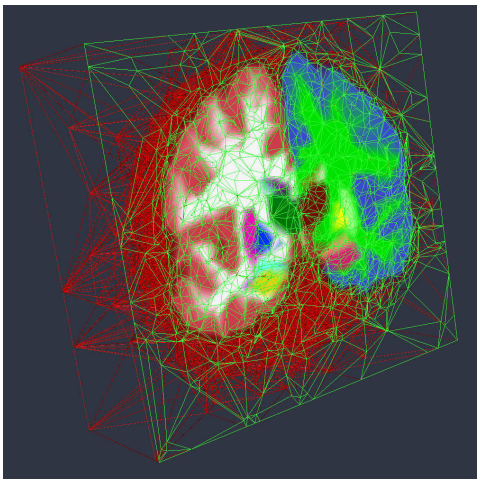


Figure 2.11: Optimal tetrahedral mesh-based atlas built from manual annotations of 36 neuroanatomical structures in 4 subjects. The prior probabilities for the different structures have been color-coded for visualization purposes.

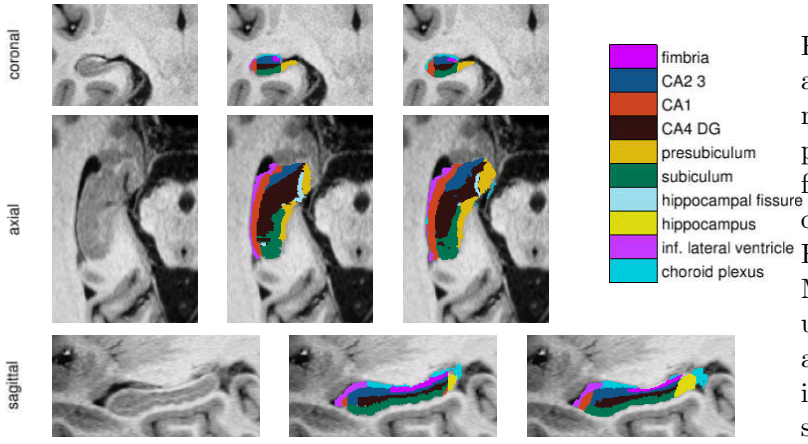


Figure 2.12: Fully automated segmentation of hippocampal subfields from ultra-high resolution MR scans. From left to right: MR data, manual delineations, and corresponding automated segmentations.

References

[1] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[2] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] H. Lappalainen and J. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, Springer, 2000, pages 75–92.

[4] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan, editor, *Learning in Graphical Models*, MIT Press, 1999, pages 105–161.

[5] A. Honkela and H. Valpola. Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Transactions on Neural Networks*, 15(4):267–282, 2004.

[6] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[7] A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In *Proc. of the 14th Int. Conf. on Neural Information Processing (ICONIP 2007)*, Kitakyushu, Japan, November 2007.

[8] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[9] Y. Qi, T. S. Jaakkola. Parameter expanded variational Bayesian methods. In *Advances in Neural Information Processing Systems 19*, pp. 1097–1104, Cambridge, MA, 2007.

[10] J. Luttinen, A. Ilin, and Tapani Raiko. Transformations for variational factor analysis to speed up learning. In *Proc. of the 14th European Symposium on Artificial Neural Networks (ESANN 2009)*, pp. 77–82, Bruges, Belgium, April 2009.

[11] Home page of our Bayes group: `http://www.cis.hut.fi/projects/bayes/`.

[12] A. Trapletti, *On Neural Networks as Statistical Time Series Models*. PhD Thesis, Technische Universität Wien, 2000.

[13] M. Tornio, A. Honkela, and J. Karhunen. Time series prediction with variational Bayesian nonlinear state-space models. In *Proc. European Symp. on Time Series Prediction (ESTSP'07)*, pages 11–19, Espoo, Finland, February 2007.

[14] T. Raiko, M. Tornio, A. Honkela, and J. Karhunen. State inference in variational Bayesian nonlinear state-space models. In *Proc. of the 6th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA 2006)*, pages 222–229, Charleston, South Carolina, USA, March 2006.

[15] T. Raiko and M. Tornio. Variational Bayesian learning of nonlinear hidden state-space models for model predictive control. In *Neurocomputing*, volume 72, issues 16–18, pages 3704–3712, October, 2009.

[16] T. Raiko, A. Ilin, and J. Karhunen. Principal component analysis for sparse high-dimensional data. In *Proc. of the 14th Int. Conf. on Neural Information Processing (ICONIP 2007)*, Kitakyushu, Japan, pp. 566–575, 2008.

[17] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. Tech. report TKK-ICS-R6, Helsinki University of Technology, TKK reports in information and computer science, Espoo, Finland, 2008.

[18] L. Kozma, A. Ilin, and Tapani Raiko. Binary principal component analysis in the Netflix collaborative filtering task. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Grenoble, France, September 2009.

[19] J. Luttinen, A. Ilin, and Juha Karhunen. Bayesian robust PCA for incomplete data. In *Proc. of the 8th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2009)*, pp. 66–73, Paraty, Brazil, March 2009.

[20] J. Zhao, Q. Jiang. Probabilistic PCA for t distributions. *Neurocomputing*, 69:2217–2226, 2006.

[21] C. Archambeau, N. Delannay, M. Verleysen. Robust probabilistic projections. In *Proc. of the 23rd International Conference on Machine Learning (ICML 2006)*, pp. 33-40, New York, NY, USA, 2006.

[22] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics* 24(16):i70–i75, 2008.

[23] A. Honkela et al. A model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A*, 2010. doi:10.1073/pnas.0914285107

[24] M. Harva. Algorithms for approximate Bayesian inference with applications to astronomical data analysis. *TKK Dissertations in Information and Computer Science*, TKK-ICS-D3, Espoo, Finland, 2008. Available at `http://lib.tkk.fi/Diss/2008/isbn9789512293483/`.

[25] S. Refsdal. On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect. *Monthly Notices of the Royal Astronomical Society*, 128:307–310, 1964.

[26] J. C. Cuevas-Tello, P. Tino, and S. Raychaudhury. How accurate are the time delay estimates in gravitational lensing? *Astronomy & Astrophysics*, 454:695–706, 2006.

[27] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. *Neurocomputing*, 72(1-3):32–38, 2008.

[28] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated Model-based Bias Field Correction of MR Images of the Brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896, 1999

[29] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated Model-based Tissue Classification of MR Images of the Brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, 1999

[30] T. Autti, M. Mannerkoski, J. Hämäläinen, K. Van Leemput, and L. Åberg. JNCL patients show marked brain volume alterations on longitudinal MRI in adolescence. *Journal of Neurology*, 255(8):1226–1230, 2008

[31] K. Van Leemput Encoding Probabilistic Brain Atlases Using Bayesian Inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837, 2009

[32] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L.L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl. Automated Segmentation of Hippocampal Subfields from Ultra-High Resolution In Vivo MRI. *Hippocampus*, 19(6):549–557, 2009

[33] B. Fischl, A. A. Stevens, N. Rajendran, B. T. T. Yeo, D. N. Greve, K. Van Leemput, J. Polimeni, S. Kakunoori, R. L. Buckner, J. L. Pacheco, D. H. Salat, J. Melcher, M. P. Frosch, B. T. Hyman, P. E. Grant, B. R. Rosen, A. J. W. van der Kouwe, G. C. Wiggins, L. L. Wald, J. C. Augustinack. Predicting the Location of Entorhinal Cortex from MRI. *NeuroImage*, 47(1):8–17, 2009

[34] M. K. Mannerkoski, H. J. Heiskala, K. Van Leemput, L. E. Åberg, R. Raininko, J. Hämäläinen, and T. H. Autti. Children and adolescents with learning and intellectual disabilities and familial need for full-time special-education show regional brain alterations: A voxel-based morphometry study. *Pediatric Research*, 66(3):306-0311, 2009

[35] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L. L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl. Model-Based Segmentation of Hippocampal Subfields in Ultra-High Resolution In Vivo MRI. *Proceedings of the MICCAI 2008 Workshop on the Computational Anatomy and Physiology of the Hippocampus (CAPH'08)*, pp. 46–55, September 6, 2008, New York, USA

[36] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L. L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl. Model-Based Segmentation of Hippocampal Subfields in Ultra-High Resolution In Vivo MRI. *Lecture Notes in Computer Science*, 5241:235–243, 2008

[37] M. R. Sabuncu, B. T. T. Yeo, T. Vercauteren, K. Van Leemput, and P. Golland. Asymmetric image-template registration. *Lecture Notes in Computer Science*, 5761:565–573, 2009

[38] T. Riklin Raviv, K. Van Leemput, W. M. Wells, and P. Golland. Joint Segmentation of Image Ensembles via Latent Atlases. *Lecture Notes in Computer Science*, 5761:272–280, 2009

[39] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. Supervised Nonparametric Image Parcellation. *Lecture Notes in Computer Science*, 5762:1075–1083, 2009

[40] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. Nonparametric Mixture Models for Supervised Image Parcellation. *Proceedings of the MICCAI 2009 Workshop on Probabilistic Models for Medical Image Analysis (PMMIA 2009)*, pp. 301-313, September 20, 2009, London, UK

[41] T. Riklin Raviv, B. Menze, K. Van Leemput, B. Stieltjes, M. A. Weber, N. Ayache, W. M. Wells, and P. Golland. Joint Segmentation via Patient-Specific Latent Anatomy Model. *Proceedings of the MICCAI 2009 Workshop on Probabilistic Models for Medical Image Analysis (PMMIA 2009)*, pp. 244-255, September 20, 2009, London, UK

Chapter 3

Blind and semi-blind source separation

Erkki Oja, Alexander Ilin, Zhirong Yang, Zhijian Yuan, Jaakko Luttinen

3.1 Introduction

Erkki Oja

What is Blind and Semi-blind Source Separation? Blind source separation (BSS) is a class of computational data analysis techniques for revealing hidden factors, that underlie sets of measurements or signals. BSS assumes a statistical model whereby the observed multivariate data, typically given as a large database of samples, are assumed to be linear or nonlinear mixtures of some unknown latent variables. The mixing coefficients are also unknown.

By BSS, these latent variables, also called sources or factors, can be found. Thus BSS can be seen as an extension to the classical methods of Principal Component Analysis and Factor Analysis. BSS is a much richer class of techniques, however, capable of finding the sources when the classical methods, implicitly or explicitly based on Gaussian models, fail completely.

In many cases, the measurements are given as a set of parallel signals or time series. Typical examples are mixtures of simultaneous sounds or human voices that have been picked up by several microphones, brain signal measurements from multiple EEG sensors, several radio signals arriving at a portable phone, or multiple parallel time series obtained from some industrial process.

Perhaps the best known single methodology in BSS is Independent Component Analysis (ICA), in which the latent variables are nongaussian and mutually independent. However, also other criteria than independence can be used for finding the sources. One such simple criterion is the non-negativity of the sources. Sometimes more prior information about the sources is available or is induced into the model, such as the form of their probability densities, their spectral contents, etc. Then the term "blind" is often replaced by "semiblind".

Our earlier contributions in ICA research. In our ICA research group, the research stems from some early work on on-line PCA, nonlinear PCA, and separation, that we were involved with in the 80's and early 90's. Since mid-90's, our ICA group grew considerably. This earlier work has been reported in the previous Triennial and Biennial reports of our laboratory from 1994 to 2007 [1]. A notable achievement from that period was the textbook "Independent Component Analysis" by A. Hyvärinen, J. Karhunen, and E. Oja [2]. It has been very well received in the research community; according to the latest publisher's report, over 5200 copies had been sold by August, 2009. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. In 2005, the Japanese translation of the book appeared (Tokyo Denki University Press), and in 2007, the Chinese translation (Publishing House of Electronics Industry).

Another tangible contribution has been the public domain FastICA software package [3]. This is one of the few most popular ICA algorithms used by the practitioners and a standard benchmark in algorithmic comparisons in ICA literature.

In the reporting period 2008 - 2009, ICA/BSS research stayed as one of the core projects in the laboratory, with the pure ICA theory waning and being replaced by several new directions in blind and semiblind source separation. In this Chapter, we present two such novel directions.

Chapter 3 starts by introducing some theoretical advances on Nonnegative Matrix Factorization undertaken during the reporting period, especially the new Projective Nonnegative Matrix Factorization (PNMF) principle, which is a principled way to perform approximate nonnegative Principal Component Analysis. Then the Gaussian-process fac-

tor analysis (GPFA) method, a semi-blind source separation principle, is applied to climate data analysis. Climate research is an interesting and potentially very useful application for large-scale semiblind models, that will be under intensive research in our group in the near future.

Another way to formulate the BSS problem is Bayesian analysis. This is covered in the separate Chapter 2.

References

[1] Triennial and Biennial reports of CIS and AIRC. `http://www.cis.hut.fi/research/reports/`.

[2] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. J. Wiley, 2001.

[3] The FastICA software package. `http://www.cis.hut.fi/projects/ica/fastica/`.

3.2 Non-negative projections

Zhirong Yang, Zhijian Yuan, and Erkki Oja

Projecting high-dimensional input data into a lower-dimensional subspace is a fundamental research topic in signal processing, machine learning and pattern recognition. Non-negative projections are desirable in many real-world applications where the original data are non-negative, consisting for example of digital images or various spectra. It was pointed out by Lee and Seung [3] that the positivity or non-negativity of a linear expansion is a very powerful constraint, that seems to lead to sparse representations for the data. Their method, *non-negative matrix factorization (NMF)*, minimizes the difference between the data matrix $\mathbf{X}$ and its non-negative decomposition $\mathbf{WH}$. The difference can be measured by the Frobenius matrix norm or the Kullback-Leibler divergence.

Yuan and Oja [7] proposed the *projective non-negative matrix factorization (PNMF)* method which replaces $\mathbf{H}$ in NMF with $\mathbf{W}^T\mathbf{X}$, thus the data matrix $\mathbf{X}$ is approximated as

$$\mathbf{X} \approx \mathbf{W}\mathbf{W}^T\mathbf{X}.$$

The nonnegative matrix $\mathbf{W}$ is assumed to have a much lower rank than the data matrix itself. This actually combines the objective of principal component analysis (PCA) with the non-negativity constraint. The PNMF algorithm has been applied e.g. to facial image processing, and the empirical results indicate that PNMF is able to produce more spatially localized, part-based representations of visual patterns.

Recently, we have extended and completed the preliminary work with the following new contributions [5]: (1) formal convergence analysis of the original PNMF algorithms, (2) PNMF with the orthonormality constraint, (3) nonlinear extension of PNMF, (4) comparison of PNMF with two classical and two recent algorithms [6, 2] for clustering, (5) a new application of PNMF for recovering the projection matrix in a nonnegative mixture model, (6) comparison of PNMF with the approach of discretizing eigenvectors, and (7) theoretical justification of moving a term in the generic multiplicative update rule. Our in-depth analysis shows that the PNMF replacement has positive consequences in sparseness of the approximation, orthogonality of the factorizing matrix, decreased computational complexity in learning, close equivalence to clustering, generalization of the approximation to new data without heavy re-computations, and easy extension to a nonlinear kernel method with wide applications for optimization problems. Figure 3.1 demonstrates the advantage of PNMF over two other methods for the Nonnegative Kernel Principal Component Analysis problem.

Furthermore, we have proposed a more general method called α-PNMF [4], using α-divergence instead of Kullback-Leibler divergence as the error measure in PNMF. We have derived the multiplicative update rules for the new learning objective. The convergence of the iterative updates is proven using the Lagrangian approach. Experiments have been conducted, in which the new algorithm outperforms α-NMF [1] for extracting sparse and localized part-based representations of facial images.

Our method can also achieve better clustering results than α-NMF and ordinary PNMF for a variety of datasets. Table 3.1 shows the resulting clustering purities on six datasets.

References

[1] Andrzej Cichocki, Hyekyoung Lee, Yong-Deok Kim, and Seugjin Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29:1433–1440,

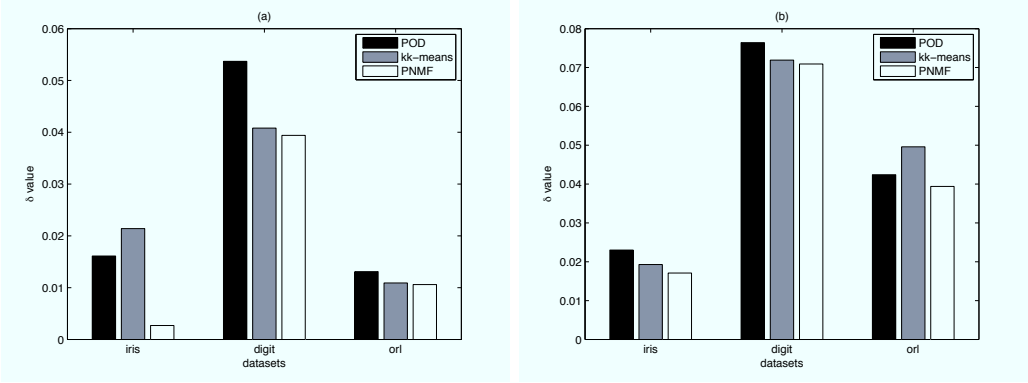


Figure 3.1: Comparison of POD, KK-means, and PNMF with (a) linear and (b) RBF kernels for the Nonnegative Kernel Principal Component Analysis problem. Smaller δ-values are better objectives relative to the KPCA solution

Table 3.1: Clustering purities using α-NMF, PNMF and α-PNMF. The best result for each dataset is highlighted with boldface font.

	α-NMF			PNMF	α-PNMF		
datasets	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$	-	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
Iris	0.83	0.85	0.84	0.95	0.95	0.95	**0.97**
Ecoli5	0.62	0.65	0.67	0.72	0.72	0.72	**0.73**
WDBC	0.70	0.70	0.72	0.87	0.86	0.87	**0.88**
Pima	0.65	0.65	0.65	0.65	0.67	0.65	**0.67**
AMLALL	0.95	0.92	0.92	0.95	**0.97**	0.95	0.92
ORL	0.47	0.47	0.47	0.75	0.76	0.75	**0.80**

2008.

[2] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. Kernel kmeans, spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, Seattle, WA, USA, 2004.

[3] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[4] Zhirong Yang and Erkki Oja. Projective nonnegative matrix factorization with α-divergence. In *Proceedings of 19th International Conference on Artificial Neural Networks (ICANN)*, pages 20–29, Limassol, Cyprus, 2009. Springer.

[5] Zhirong Yang and Erkki Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transaction on Neural Networks*, 2010. In press.

[6] Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pages 313–319, 2003.

[7] Zhijian Yuan and Erkki Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proc. of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, pages 333–342, Joensuu, Finland, June 2005.

3.3 Reconstruction of historical climate data by Gaussian-process factor analysis

Alexander Ilin and Jaakko Luttinen

Studying natural variability of climate is a topic of intensive research in climatology. In our earlier research, we have extended the classical technique of rotated Principal Components, or Empirical Orthogonal Functions, by introducing the concept of "interesting structure" for massive sets of spatio-temporal climate measurements. In our case, the goal of exploratory analysis is to find signals with some specific structures of interest. They may for example manifest themselves mostly in specific variables, which exhibit prominent variability in a specific timescale etc. An example of such analysis can be extracting clear trends or quasi-oscillations from climate records. The procedure for obtaining suitable rotations of EOFs can be based on the general algorithmic structure of denoising source separation (DSS) [1].

However, understanding long-term variability of climate faces the problem of the scarcity of climate observations in the past. Thus, reconstruction of historical climate becomes an important problem.

The standard methods of statistical reconstruction are ad hoc adjustments of PCA for incomplete data making such additional assumptions as temporal and spatial smoothness of the observed climate variables. These assumptions were used, for example, in [2] to reconstruct the global sea surface temperatures (SST) in the 1856–1991 period from the MOHSST5 data set (which is largely based on the measurements made from merchant ships). The method presented there uses additional information about the quality of the data and this uncertainty information is derived from the number of different sources which were used to compute each data sample.

In our recent papers [3, 4], we use the Bayesian framework to perform statistical reconstructions of spatio-temporal data. In [3], we adopt the basic variational Bayesian PCA model and use additional uncertainty information to improve the reconstruction performance.

In [4], we present a more advanced probabilistic model called *Gaussian-process factor analysis (GPFA)*. The method is based on standard matrix factorization:

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \text{noise} = \sum_{d=1}^{D} \mathbf{w}_{:d}\mathbf{x}_{d:}^{\mathrm{T}} + \text{noise},$$

where $\mathbf{Y}$ is a data matrix in which each row contains measurements in one spatial location and each column corresponds to one time instance. Each $\mathbf{x}_{d:}^{T}$ is a row vector representing the time series of one of the D factors, whereas $\mathbf{w}_{:d}$ is a column vector of loadings which are spatially distributed. Matrix $\mathbf{Y}$ may contain missing values and the samples can be unevenly distributed in space and time.

We assume that both factors $\mathbf{x}_{d:}$ and corresponding loadings $\mathbf{w}_{:d}$ have prominent structures that we model using the tool of Gaussian processes [5]. The model is identified in the framework of variational Bayesian learning and high computational cost of GP modeling is reduced by using sparse approximations derived in the variational methodology.

In the experiments reported in [4], we show that GPFA can provide better reconstructions of global SST set compared to variational Bayesian PCA. Figure 3.2 shows the spatial and temporal patterns of the four most dominant principal components found by GPFA from the MOHSST5 data set. The obtained test reconstruction errors were 0.5714 for GPFA and 0.6180 for VBPCA, which can be seen as a significant improvement.

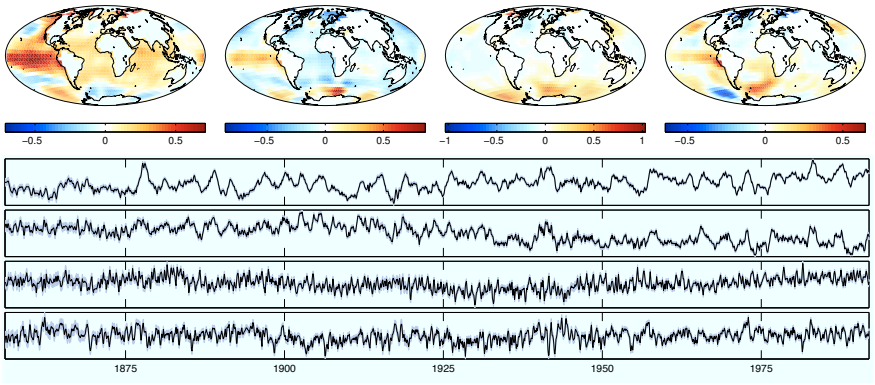


Figure 3.2: The spatial and temporal patterns of the four most dominating principal components estimated by GPFA from the MOHSST5 dataset. The solid lines and gray color in the time series show the mean and two standard deviations of the posterior distribution.

References

[1] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

[2] A. Kaplan, M. Cane, Y. Kushnir, M. Blumenthal, B. Rajagopalan. Analysis of global sea surface temperatures 1856–1991. *Journal of Geophysical Research*, 103:18567–18589, 1998.

[3] A. Ilin and A. Kaplan. Bayesian PCA for reconstruction of historical sea surface temperatures. In *Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN 2009)*, pp. 1322–1327, Atlanta, USA, June 2009.

[4] J. Luttinen and A. Ilin. Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Advances in Neural Information Processing Systems (NIPS) 22*, Vancouver, Canada, Dec. 2009.

[5] C. E. Rasmussen, C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

Chapter 4

Multi-source machine learning

Samuel Kaski, Arto Klami, Gayle Leen, Jaakko Peltonen, Kai Puolamäki, Janne Sinkkonen, Zhirong Yang, Helena Aidos, Ilkka Huopaniemi, Kristian Nybo, Juuso Parkkinen, Eerika Savia, Tommi Suvitaival, Abhishek Tripathi

4.1 Introduction

We develop statistical machine learning methods for extracting useful regularities from large, high-dimensional data sets. In practical computational data analysis tasks a common problem is lack of sufficient amount of representative data. Modeling requires either data or prior knowledge which by definition does not exist in knowledge discovery or data mining tasks. If there was enough data, modern statistical machine learning toolboxes would contain powerful approaches to building flexible models that do not make strong assumptions about data, but flexible models are naturally weak given little data.

In many applications, for instance in molecular biology and neuroinformatics, there is data available in public or special-purpose databanks, but the problem is that not everything is relevant. We are developing new machine learning methods capable of learning from *multiple data sources* containing only *partially relevant* data, and generalizing to new contexts. The methods extend and generalize the current approaches called multi-view, multi-way and multi-task learning, on structured and unstructured domains.

Moreover, we have developed new principles and methods for the task of *visualizing* high-dimensional data; this task is central in any knowledge discovery process.

4.2 Multi-view learning

Multi-view learning tells how several data sources, or views, can be combined to extract more relevant information. We focus on unsupervised settings, where the relevance comes from statistical dependencies between multiple views of the same objects. For example, a collection of images with captions can be represented with two views, one capturing the contents of the image while the other describes the caption. Dependencies between these representations reveal more information on the intended content, or semantics, of the images than either view alone.

We have developed new theory for decomposing variation in multiple views into source-specific and shared components [1], building on Bayesian latent-variable models that capture the dependencies by assigning flexible source-specific models for describing the noise in each of the views. The same basic formulation extends to various practical models. A prime example is [2] that applies hierarchical non-parametric Bayesian models for making the source-specific parts extremely flexible, and builds a hierarchical grouping of human genes based on both mRNA and protein expression. The model, illustrated in Figure 4.1, reveals processes that could not be found by looking at either view alone.

Besides advanced Bayesian solutions, we have also developed novel multi-view algorithms for application purposes. [4] aimed at creating an easy-to-use data integration tool for bioinformatics applications and was accompanied by an open-source software package, while [3] introduces a fast algorithm for maximizing mutual information of linear projections, applied to brain imaging data.

Going beyond standard multi-view learning, we have also developed novel solutions for applications without co-occurring data. Traditional multi-view learning can only be applied for cases with clear one-to-one co-occurrence between the views. We showed in [5] that the co-occurrence itself can be learned by maximizing statistical dependency between two views with no known co-occurrence. In brief, the idea is to order the samples of one of the views so that the dependency between the views is maximal. This, in turn, requires efficient means for measuring the dependency, provided by classical data integration tools like canonical correlation analysis (CCA). A simple iterative algorithm alternating between optimizing the ordering (solved through a linear assignment problem) and finding a representation that maximally captures the dependency (solved through CCA) finds the co-occurrences with high accuracy. We have applied the algorithm for aligning probeset of various microarray brands, matching metabolite identities of different species or measurement batches, and aligning sentences of bi-lingual corpora.

References

[1] Arto Klami and Samuel Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72:39–46, 2008.

[2] Simon Rogers, Arto Klami, Janne Sinkkonen, Mark Girolami, and Samuel Kaski. Infinite Factorization of Multiple Non-parametric Views. *Machine Learning*, 79(1–2):201–226, 2010.

[3] Eerika Savia, Arto Klami, and Samuel Kaski. Fast dependent components for fMRI analysis. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1737–1740, 2009.

[4] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

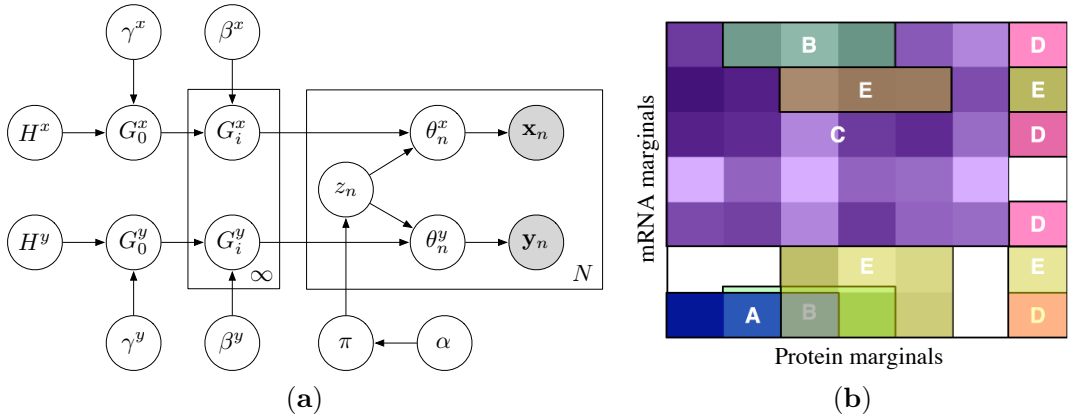


Figure 4.1: (**a**): Illustration of the hierarchical Dirichlet process model for cluster analysis of coupled data sources. (**b**): Application of the model on coupled analysis of mRNA and protein concentrations. Both marginals correspond to clusters of genes, automatically detected by the model, and the color-codes and letters indicate higher-level processes obtained by simultaneous clustering of the contingency table of cluster assigments.

[5] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Using dependencies to pair samples for multi-view learning. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1561–1564, 2009.

4.3 Multi-task learning

We have introduced two new multi-task learning setups, suitable for different scenarios, and solutions for them: *relevant subtask learning* and *paired multi-task learning*.

Relevant subtask learning

It is all too common in classification tasks that there is too little training data to estimate sufficiently powerful models. The problem is particularly hard for the high-dimensional data in genome-wide studies of modern bioinformatics, but appears also in image classification from few examples, finding of relevant texts, etc.

After realizing that the world is full of other data sets, the problem becomes how to simultaneously learn from a small data set and retrieve useful information from the other data sets. We have recently introduced a learning problem called *relevant subtask learning*, a variant of multi-task learning, which aims to solve the small-data problem by intelligently making use of other, potentially related "background" data sets.

Such potentially related "background" data sets are available for instance in bioinformatics, where there are databases full of data measured for different tasks, conditions or contexts; for texts there is the web. Such data sets are *partially relevant*: they do not come from the exact same distribution as future test data, but their distributions may still contain some useful part. Our research problem is, *can we use the partially relevant data sets to build a better classifier for the test data?*

Learning from one of the data sets is called a "task". Our scenario is then a special kind of *multi-task learning* problem. However, in contrast to typical multi-task learning, our problem is fundamentally asymmetric and more structured; test data fits one task, the "*task-of-interest*," and other tasks may contain *subtasks* relevant for the task-of-interest, but no other task needs to be wholly relevant.

We have introduced a method that uses logistic regression classifiers. The key is to assume that each data set is a mixture of relevant and irrelevant samples. By fitting this model to all data sets, the common model for relevant samples learns from all tasks. To fit the model, we have used both simple maximum likelihood fitting [1] and more advanced variational Bayesian inference [3]. We model the irrelevant part with a sufficiently flexible model such that irrelevant samples cannot distort the model for relevant data. A sample application is a news recommender for one user, where classifications from other users are available (Fig. 4.2). The relevant subtask learner outperforms a comparable standard multi-task learning model [4].

The generalization error of relevant subtask learning has been analyzed theoretically in [5] in a slightly different setting, where the task is density estimation and supplementary tasks are assumed to be mixtures of a shared interesting density and a non-interesting task-specific density. Relevant subtask learning has smaller generalization error than learning from the task-of-interest alone or from a supplementary task alone.

Paired Multi-task Learning

When faced with an abundance of tasks containing potentially relevant information to a desired learning task, we ask: how can we decide which tasks are relevant? And what is the relationship between the different tasks? Knowledge about the task relationships and problem structure can then be exploited in jointly learning multiple tasks. By sharing statistical strength between different tasks, this *multitask learning* set-up can overcome potential problems when there is little data for a single task.

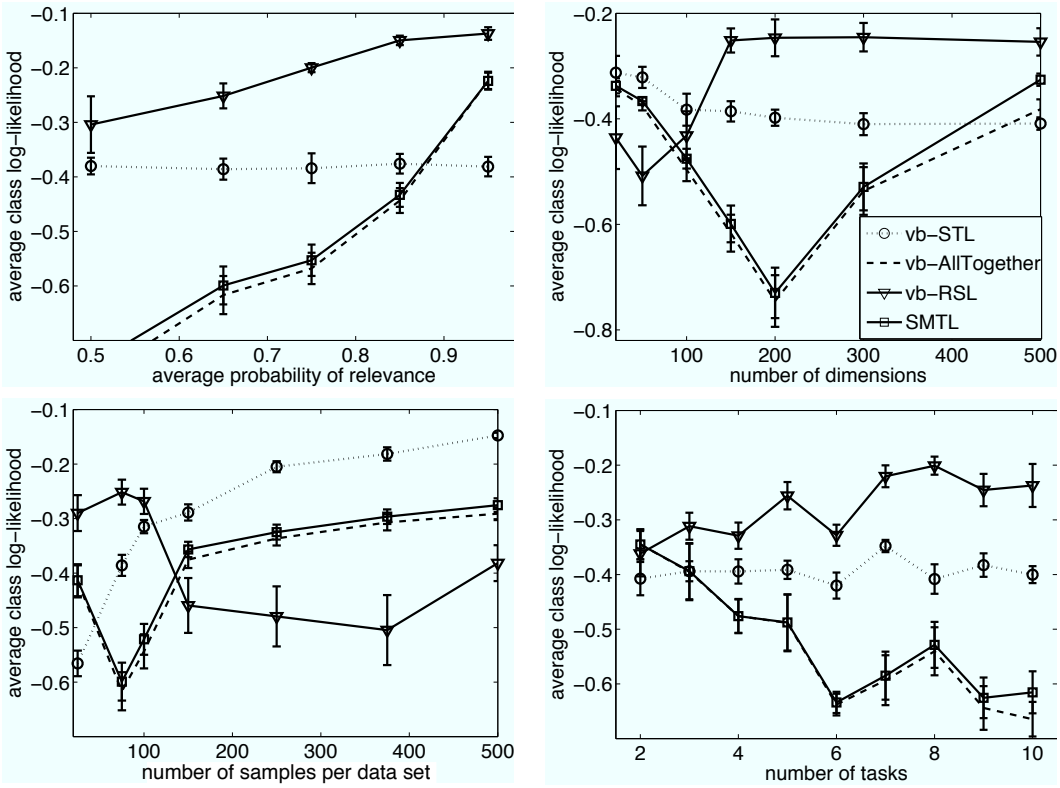


Figure 4.2: Comparison of multi-task learning approaches on news article data. The task was to predict relevance of news articles to a specific reader (the reader-of-interest), using articles rated by other readers as additional sources of information. Results are shown as a function of several design parameters: the proportion of relevant samples (**t**op left), data dimensionality (**t**op right), the number of samples per data set (**b**ottom left) and the number of tasks (data sets; **b**ottom right). Relevant subtask learning (vb-RSL) outperforms a multi-task method that clusters tasks (SMTL; [4]) and to two naive methods ("vb-STL" and "vb-AllTogether") when there are many dimensions but few samples per data set (less than 100), which is a realistic scenario.

We address a specific problem in bioinformatics: learning to choose control samples for use in a differential gene expression experiment in cancer (case vs control). Gene expression measurements are likely to contain bias due to factors such as patient-specific and laboratory-specific effects, and typically there are only a small number of samples available for each experimental condition. These factors make it problematic to select a set of pairs of control and tumor tissue (case) samples, such that the differential gene expression of the case samples is solely due to cancer-specific variation. However, there is potentially a huge amount of useful information about cancer, and the relationship to control tissue contained in publicly available gene expression databases. If two cancer types are similar, then it is likely that they will use similar control samples.

The suitable controls for each experiment form a group of controls. These groups are considered as classes / control tissues, and the task is to classify each case sample to one of these classes. We formulate this as a multi-task learning problem in [2] where we have *a set of primary tasks* (choosing the control class for each sample for a cancer type) which we want to learn, and *a set of auxiliary tasks* (choosing the control class for each control sample). This follows a paired structure, such that each primary task is paired

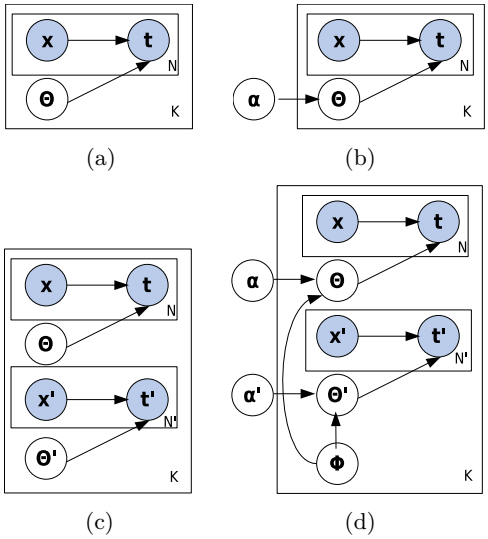


Figure 4.3: Schematic illustration of statistical strength sharing in multitask learning scenarios. Learning a set of K tasks as in (a) amounts to finding different parameterisations $\theta_i, i = 1, ..., K$ for the tasks. If the tasks are assumed to be related, multitask learning approaches assume some shared structure across all K tasks through a common parameterisation via α (b). We consider the situation where there are K pairs of tasks (c), and propose the structure in (d) to share information between the tasks. There is shared structure within each task set's parameterisation θ, θ' through α, α' and across each of the K pairs through ϕ.

with a corresponding auxiliary task. We transfer information about the *relatedness of the auxiliary set of tasks* to the set of primary tasks (see Figure 4.3 and its caption for more details). We formulate the model using the Gaussian process framework; the task functions are given Gaussian process priors and the task structure is modeled through the parameterisation of the covariance functions. For each set of tasks, the task functions are assumed to come from a linear combination of an underlying set of latent functions. This linear combination, which models the inter-task similarity in each set, is constrained to be the same for both the primary and auxiliary task set.

In learning the classification, we use knowledge about the relationships between the case and the control samples. This pairing is transferred to new pairs, such that our model can infer a suitable control sample for a new case sample (see Figure 4.4).

References

[1] Samuel Kaski and Jaakko Peltonen. Learning from relevant tasks only. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 608–615. Springer-Verlag, Berlin, Germany, 2007.

[2] Gayle Leen, David R. Hardoon and Samuel Kaski. Automatic Choice of Control Measurements, In *Advances in Machine Learning (Proc. ACML'09, The 1st Asian Conference on Machine Learning)*,5828:206–219. Springer, 2009.

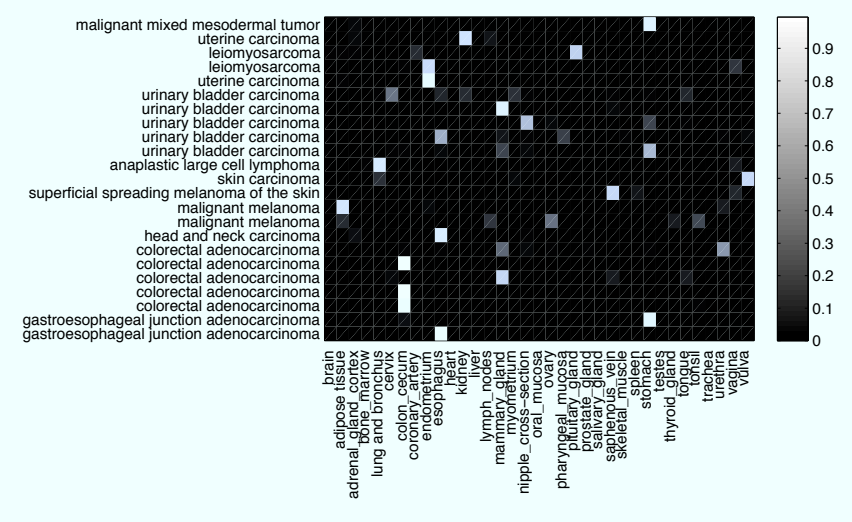


Figure 4.4: Visualization of the probability distribution over the control classes (x axis) for some tumor samples (y axis) with unknown control classes

[3] Jaakko Peltonen, Yusuf Yaslan, and Samuel Kaski. Relevant subtask learning by constrained mixture models. *Intelligent Data Analysis*, to appear.

[4] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research*, 8: 35–63, 2007.

[5] Keisuke Yamazaki and Samuel Kaski. An Analysis of Generalization Error in Relevant Subtask Learning. In Mario Köppen, Nikola Kasabov, and George Cooghill, editors, *Advances in Neuro-Information Processing, 15th International Conference, ICONIP 2008*, pages 629–637. Springer-Verlag, Berlin Heidelberg, 2009.

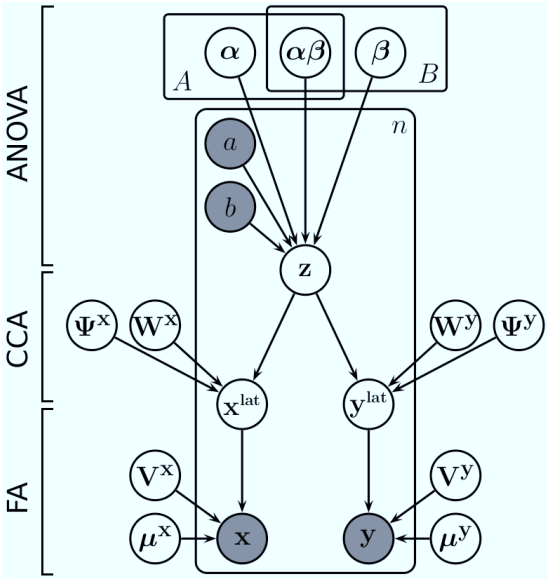


Figure 4.5: Plate diagram of the graphical model for Multi-Way, Multi-View Learning

4.4 Multi-way learning

Finding effects of one or multiple known covariates from the data is one of the most common statistical problems, commonly solved by tranditional Analysis of Variance (ANOVA), its multivariate generalization (MANOVA), or in general by linear models. The traditional methods are not applicable and very limited alternative methods exist to currently increasingly important problems in molecular biology where the dimensionality of the problem p is very large and the number of observations n is (relatively) small. The same "large p, small n" problem recurs also in other fields.

In biological experiments typical covariates are disease, drug treatment groups, gender or time-series, resulting in a multi-way experimental setup. The main challenge in biology is that the number of samples (for instance mice or human patients) is small due to economical and ethical cost, whereas the number of variables (such as genes or metabolites) is huge. Due to this, the traditional multivariate methods cannot be used, and on the other hand little research of multi-way analysis has been presented in the machine learning literature.

We have recently introduced a Bayesian method for solving this burning problem of multi-way analysis of small sample-size, high-dimensional datasets [1]. Moreover, the multi-way data-analysis problem becomes even more complicated when heterogeneous data with multiple covariates are integrated from multiple sources. Different data sources usually have distinct, unmatched variable-spaces with different dimensionalities. We have generalized ANOVA-type analysis to the case of multiple sources by considering the source ("view") as an additional covariate in the ANOVA-type analysis. The problem is impossible for traditional methods due to the different variable-spaces, but by utilizing dependencies between the sources the problem can be solved. We introduced a model (Figure 4.5; [2]) which is able to find the multi-way covariate-effects and to partion them into shared and source-specific effects. The method is applicable to any small sample-size, multi-source experiments, currently very popular in biological research.

References

[1] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Samuel Kaski, and Matej Orešič. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19(2):261–276, 2009.

[2] Ilkka Huopaniemi and Tommi Suvitaival and Janne Nikkilä and Matej Orešič and Samuel Kaski. Multi-Way, Multi-View Learning. In *NIPS 2009 workshop on Learning from Multiple Sources with Applications to Robotics*, 2009

4.5 Information visualization

Visualization of mutual similarities of entries in large high-dimensional data sets is a central subproblem in exploratory analysis and mining. It is makes sense to "look at the data" in all stages of data analysis, and reducing the dimensionality to two or three gives a scatterplot visualization.

It is generally not possible to show all the similarity relationships within a high-dimensional data set perfectly on a low-dimensional display; some properties are necessarily lost or misrepresented. All linear or nonlinear dimensionality reduction methods must make a compromise about which kinds of similarity relationships they aim to show, but which compromise is best for visualization? Many methods practically ignore this question because they are not designed to reduce the dimensionality of the data set lower than is possible without losing information; several such methods have difficulties when producing low-dimensional displays. Some methods choose the compromise implicitly in that they produce the lower-dimensional representation by minimizing a cost function, but the cost function has not been motivated from the point of view of visualization, that is, it is not obvious why a projection that minimizes the cost function should be a good visualization.

It has been difficult to assess the quality of visualizations since the task of visualization has not been well-defined. We have addressed this problem and introduced rigorously motivated measures for the quality of a visualization, as well as a nonlinear dimensionality reduction method that optimizes these measures and is therefore specifically designed for optimal visualization.

Visualization as information retrieval

We view visualization as an information retrieval task. An analyst looking at a scatterplot can choose any point (data item) and find its neighbors (similar other items) in the visualization. The visualization helps in this task of retrieving similar items, and quality of retrieval can be measured with standard information retrieval measures *precision* and *recall*. Any information retrieval method needs to make a compromise between these measures, parameterized by the relative cost of false positives and misses. Since a visualizer is an information retrieval device as well, it needs to make the same compromise.

We have adapted the information retrieval measures to visualization by smoothing them and representing them as differences between distributions of points being neighbors. It turns out that the traditional measures are limiting cases of these more general measures. Once the relative cost λ of false positives and misses has been fixed, we can directly optimize the visualization to minimize the retrieval cost. We call the resulting visualization method the Neighborhood Retrieval Visualizer (NeRV) [7, 8]. NeRV outperforms several recent nonlinear dimensionality reduction methods both by the new measures and by traditional measures.

We have extended NeRV to supervised visualization [4], to linear visualization [2], and to visualization with ontological annotation [3].

In addition to NeRV, we have introduced methods for the specific application of visualizing convergence of Markov chain Monte Carlo (MCMC) sampling methods commonly used in Bayesian inference [5].

One of the popular nonlinear dimensionality reduction methods, Stochastic Neighbor Embedding (SNE; [1]) is a special case of NeRV, corresponding to optimizing recall only. We have additionally introduced other generalizations of SNE and efficient algorithms for computing it, as described in the following.

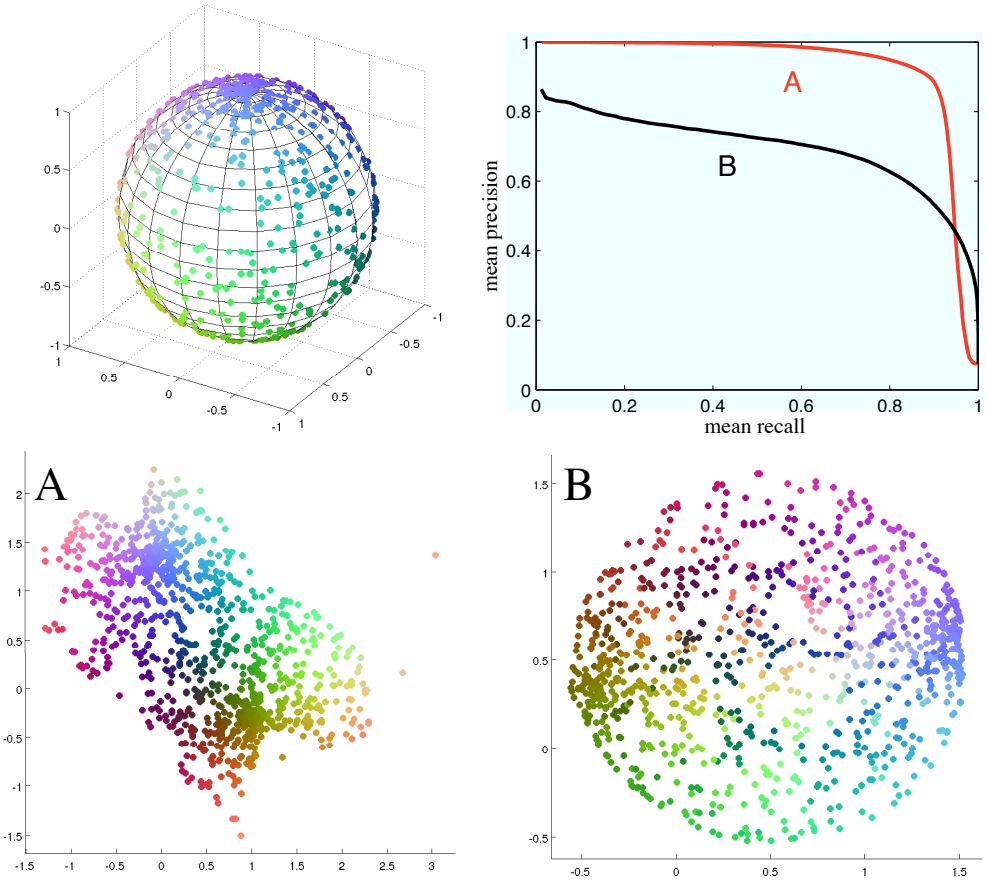


Figure 4.6: Demonstrating the precision-recall tradeoff in visualization. The task is to retrieve neighbors of points in the original space, based on their locations in the visualization. **Top left:** A three-dimensional dataset sampled from the surface of a sphere. **Bottom:** Two embeddings of the dataset. In **A**, the sphere has been cut open and folded out. This eliminates *false positives* (false neighbors), but there are some *misses* (missed neighbors) because points on different sides of the tear end up far away from each other. In contrast, **B** minimizes the number of misses by simply squashing the sphere flat; this yields many false positives because points on opposite sides of the sphere are mapped close to each other. **Top right:** mean precision–mean recall curves for the two projections. **A** has better precision (yielding higher values at the left end of the curve) **B** has better recall (yielding higher values at the right end of the curve).

Heavy-tailed Symmetric Stochastic Neighbor Embedding

Stochastic Neighbor Embedding (SNE) has been shown to be quite promising for data visualization. Currently, the most popular implementation, t-SNE [6], is restricted to a particular Student t-distribution as its embedding distribution. Moreover, it uses a gradient descent algorithm that may require users to tune parameters such as the learning step size, momentum, etc., in finding its optimum.

In [9], we have rigorously investigated the working mechanism of Heavy-tailed Symmetric Stochastic Neighbor Embedding (HSSNE). The several findings are: (1) we propose to use a negative score function to characterize and parameterize the heavy-tailed embedding similarity functions; (2) this finding has provided us with a power family of functions that

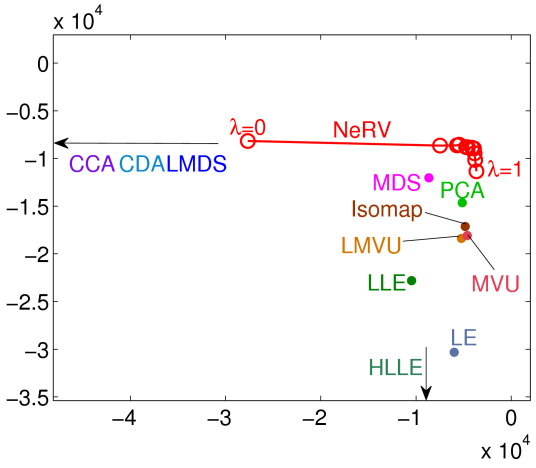


Figure 4.7: Comparison of visualization performance between several recent methods on a data set of mouse gene expression profiles, in terms of two novel measures: mean smoothed precision (vertical axis) and mean smoothed recall (horizontal axis). Our method NeRV performs best (best values near the upper right corner).

convert distances to embedding similarities; and (3) we have developed a fixed-point algorithm for optimizing SSNE, which greatly saves the effort in tuning program parameters and facilitates the extensions and applications of heavy-tailed SSNE. We have presented two empirical studies, one for unsupervised visualization showing that our optimization algorithm runs as fast and as good as the best known t-SNE implementation and the other for semi-supervised visualization showing quantitative superiority using the homogeneity measure as well as qualitative advantage in cluster separation over t-SNE. The latter results are shown in Figure 4.8.

References

[1] Geoffrey Hinton and Sam T. Roweis. Stochastic Neighbor Embedding. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 833–840. MIT Press, Cambridge, MA, 2002.

[2] Jaakko Peltonen. Visualization by Linear Projections as Information Retrieval. In José Príncipe and Risto Miikkulainen, editors, *Advances in Self-Organizing Maps (proceedings of WSOM 2009)*, pages 237-245. Springer, Berlin Heidelberg, 2009.

[3] Jaakko Peltonen, Helena Aidos, Nils Gehlenborg, Alvis Brazma, and Samuel Kaski. An information retrieval perspective on visualization of gene expression data with ontological annotation. In *Proceedings of ICASSP 2010*, to appear.

[4] Jaakko Peltonen, Helena Aidos, and Samuel Kaski. Supervised Nonlinear Dimensionality Reduction by Neighbor Retrieval. In *Proceedings of ICASSP 2009, the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1809–1812. IEEE, 2009.

[5] Jaakko Peltonen, Jarkko Venna, and Samuel Kaski. Visualizations for Assessing Convergence and Mixing of Markov Chain Monte Carlo Simulations. *Computational Statistics and Data Analysis*, 53:4453–4470, 2009.

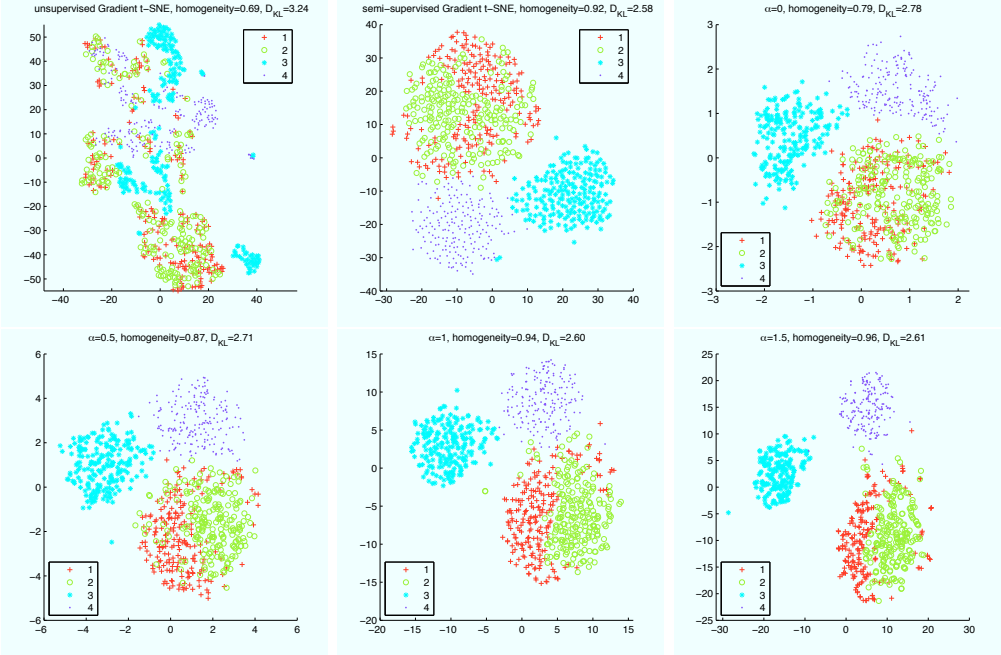


Figure 4.8: Semi-supervised visualization for the *vehicle* data set. The plots titled with α values are produced using the fixed-point algorithm of the power family of HSSNE.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[7] Jarkko Venna and Samuel Kaski. Nonlinear dimensionality reduction as information retrieval. In M. Meila and X. Shen, editors, *Proceedings of AISTATS*07, the 11th International Conference on Artificial Intelligence and Statistics (JMLR Workshop and Conference Proceedings Volume 2)*, pages 572–579, 2007.

[8] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[9] Zhirong Yang, Irwin King, Zenglin Xu, and Erkki Oja. Heavy-tailed symmetric stochastic neighbor embedding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2169–2177, 2009.

4.6 Networks

Machine Learning is in the midst of a "structural data revolution". After many decades of focusing on independent and identically-distributed examples, many researchers are now modelling inter-related entities that are linked together into complex graphs. A major driving force is the explosive growth of heterogeneous data collected in diverse sectors of the society. Example domains include bioinformatics, communication networks, and social network analysis.

Networks are a special case of structural data. Inferring properties of the network nodes, or vertices, from the links, or edges, has become a common data mining problem. Network data are typically not a complete description of reality but come with errors, omissions and uncertainties. Some links may be spurious, for instance due to measurement noise in biological networks, and some potential links may be missing, for instance friendship links of newcomers in social networks. Probabilistic generative models are a tool for modeling and inference under such uncertainty. They treat the links as random events, and give an explicit structure for the observed data and its uncertainty. Compared to non-stochastic methods, they are therefore likely to perform well as long as their assumptions are valid; they may reveal properties of networks that are difficult to observe with non-statistical techniques from the noisy and incomplete data, and they also offer a groundwork for new conceptual developments.

We have earlier introduced a family of Bayesian probablistic component models for analyzing interactions or graphs, called Interaction Component Model (ICM). We applied ICM to the task of detecting dense subnetworks from noisy protein-protein interaction networks, and additionally from multiple views; protein-protein interactions and gene expression data [2]. Such subnetworks are interpretable as functional gene modules or protein complexes. Our methods outperformed other state-of-the-art methods in this task of discovering functional subnetworks.

We further extended the ICM framework to handle multi-relational data [3], and to detect block structures [1]. For example, protein complexes consist of tightly interacting proteins, and the complexes in turn interact with other complexes.

References

[1] Juuso Parkkinen, Adam Gyenge, Janne Sinkkonen and Samuel Kaski. A block model suitable for sparse graphs. In *The 7th International Workshop on Mining and Learning with Graphs (MLG'09), Leuven, Belgium, July 2-4* (2009).

[2] Juuso Parkkinen and Samuel Kaski. Searching for functional gene modules with interaction component models. *BMC Systems Biology* 2010 4:4.

[3] Janne Sinkkonen, Janne Aukia and Samuel Kaski. Infinite mixtures for multi-relational categorical data. In *Proceedings of the 6th International Workshop on Mining and Learning with Graphs (MLG 2008)*, Helsinki, Finland, 2008.

Bioinformatics and Neuroinformatics

Chapter 5

Bioinformatics

Samuel Kaski, Jarkko Salojärvi, Gayle Leen, Arto Klami, Jaakko Peltonen, José Caldas, Andrey Ermolov, Ali Faisal, Ilkka Huopaniemi, Leo Lahti, Juuso Parkkinen, Abhishek Tripathi

5.1 Introduction

New so-called high-throughput measurement techniques have made possible genome-wide studies of gene function. Gene expression, gene regulation, protein content, protein interaction, and metabolic profiles can be measured and combined with the genetic sequence. The methods are used routinely in modern biology and medicine, and now the current challenge is to extract meaningful findings from the noisy and incomplete data masses, collected into both community resource and private data banks. The data needs to be analyzed, mined, understood, and taken into account in further experiments, which makes data analysis an integral part of biomedical research. Successful genome-wide analyses would allow a completely novel systems-level view into a biological organism.

We develop new modeling and data analysis principles needed for discovering the relevant signals and patterns from among the several measurement sources, and numerous earlier experiments collected into measurement databases. Our multi-source machine learning methods have proven to be very useful here, and new methods for retrieving and analyzing relevant experiments have promise for breakthroughs in making the data-driven sciences, biology and medicine, more cumulative. We have long-standing collaboration with European Bioinformatics Institute EBI (prof. A. Brazma), Laboratory of Cytomolecular Genetics (Prof. S. Knuutila), Department of Biological and Environmental Sciences, University of Helsinki (Prof. J. Kangasjärvi), VTT (Prof. M. Orešič), Finnish Institute for Molecular Medicine FIMM (prof. O. Kallioniemi), and smaller-scale or preliminary collaboration with several other groups.

5.2 Translational medicine on metabolic level

Translational medicine is a research field which attempts to more directly bring basic research findings to clinical practice. One of the necessary steps of this process is to translate inferences made on the molecular level, for example about metabolites, in model organisms into inferences about humans. Such translation is extremely challenging and the existing knowledge, if there is any, is currently largely tacit and only known to experts of the specific disease and model organism.

Metabolomics is the study of the set of all metabolites found in a sample tissue. Metabolite concentrations are affected strongly by diseases and drugs, and hence they complement the genomic, proteomic, and transcriptomic measurements in an excellent way, in studies of the biological state of an organism.

Our mission is to develop the computational methods needed for making molecular level translational medicine possible. We have developed new computational methods for mapping between the observed metabolomics data from model organisms and humans. In project TRANSCENDO we applied the methods to studies of the emergence of Type I diabetes, by computing mappings between non-obese diabetic (NOD) mice and children, and between the effects of a disease in several tissues. The project is collaboration within a consortium involving computational systems biology (M. Orešič, VTT), semantic modelling (Antti Pesonen, VTT), probabilistic modelling (us), and pharmacology and animal models of metabolic disease (Eriika Savontaus, University of Turku).

Metabolomic development in humans. Metabolic development in children developing into Type 1 diabetes is not well understood, and we develop computational methods in order to shed more light into it. We work on a unique data set of our collaborators [11], of metabolomic profiles derived from time series of blood samples of a large cohort of children.

We developed computational methods for studying dynamic differences between time-series measurements of two populations. In the first phase, differences between healthy boys and girls were studied [10], and at the moment we are moving forward to actual translational medicine.

The models operate under the assumption that the metabolic profiles are generated by a set of unobserved metabolic states, which can as the first approximation be modelled with Hidden Markov Models (HMM). HMM fits the assumption of latent states very well, is easy to compute and interpret, and can be extended into more flexible and expressive models. Moreover, HMM provides a way for probabilistic re-alignment of the time series, which takes into account the individual variation in the dynamics. The HMMs were able to separate the boys' and girls' metabolic states (Figure 5.1) more efficiently apart than traditional linear methods.

Disease-related dependencies between multiple tissues. A common setting in medical research is that a disease may be mainly located in a specific organ, for example in lungs, but it indirectly affects multiple tissues. Giving drugs to patients induces an analogous setup: the drugs may affect multiple other tissues in addition to the target tissue (and hence disease).

A typical analysis setup in any one-tissue (typically blood) biological experiments, looking for potential biomarkers for disease, is the diseased-healthy differential analysis. Biological experiments often contain additional covariates, such as drug treatment groups, gender or time-series, resulting in a multi-way experimental setup. Finding effects of multiple covariates from the data is a traditional statistical problem dealt with by Analysis

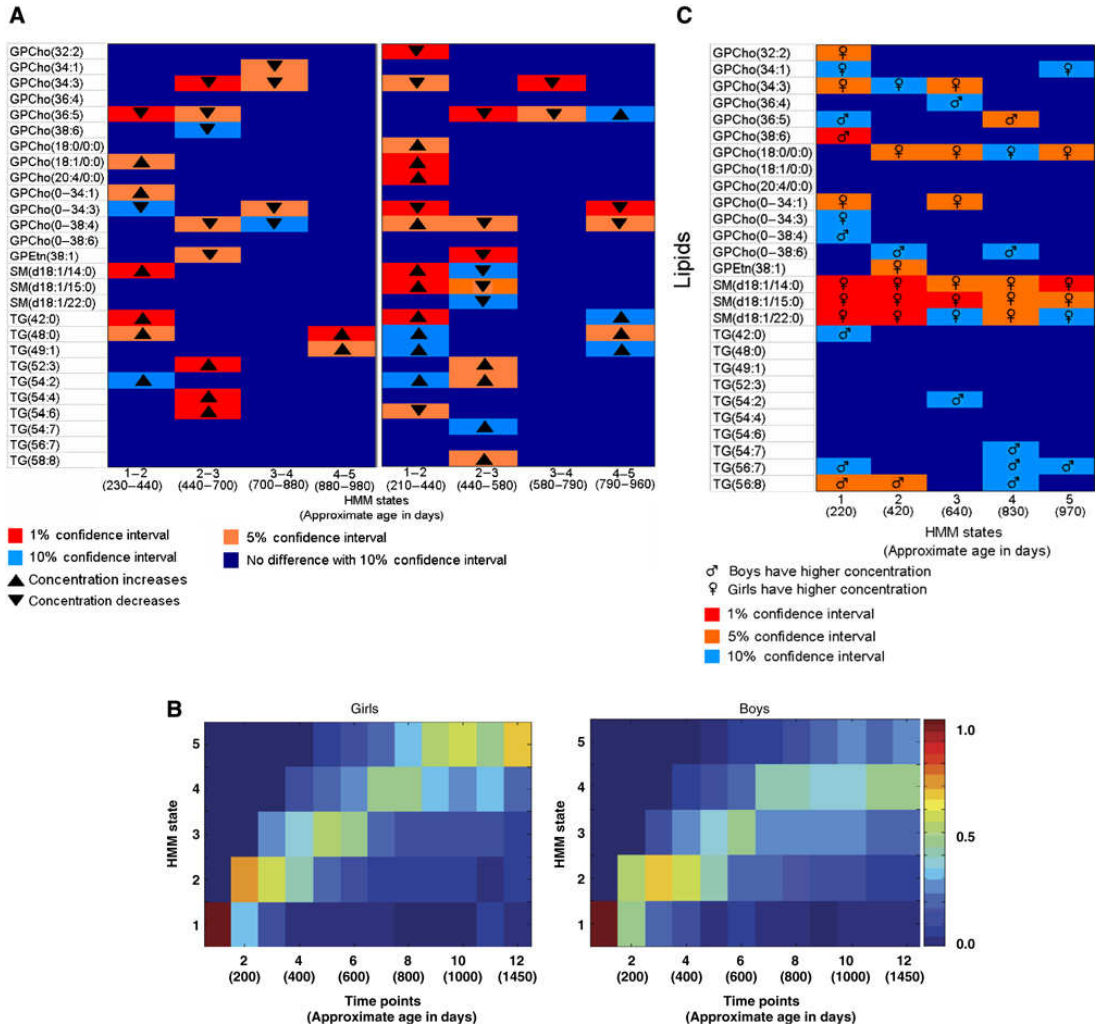


Figure 5.1: Boys and girls have different development of metabolic states.

of Variance (ANOVA) or in general by linear models. However, the main challenge in modern molecular biology is that the number of samples (such as mice or human patients) is small due to economical and ethical reasons, and the number of variables (genes or metabolites) is huge. Due to this, the traditional multivariate methods cannot be used, and few modern methods exist for this task. To address this broad and common set of problems, we developed a Bayesian model family for multi-way analysis of small sample-size, high-dimensional datasets [5].

The problem becomes even more interesting when the different data sources (here tissues) form different variable-spaces. Then stardard approaches are not applicable even in principle. Our model can be extended even to this case, by considering the different sources as different "views" in the sense of multi-view machine learning. The extended model is able to find the multi-way covariate-effects and to partion them into shared and source-specific effects. The method is applicable to any small sample-size, multi-source experiments, currently very popular in biological research. We call the general problem (Figure 5.2) Multi-Way, Multi-View Learning [6].

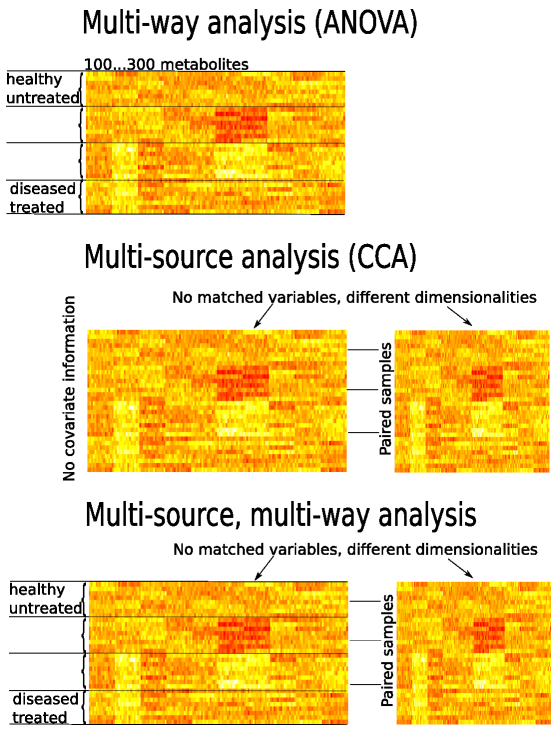


Figure 5.2: (a) Multi-way analysis studies datasets with two or more covariates for each sample. The task is to find the effects of the covariates in the data (b) Multi-source analysis studies dependencies between two or more datasets with paired samples without covariate information. (c) Multi-way, multi-source analysis combines both tasks. The task is to find shared and source-specific covariate effects. In the data matrices, rows represent samples, and columns represent variables.

5.3 Retrieval and visualization of relevant experiments

Repositories of genome-wide expression studies such as ArrayExpress [12] are becoming mature both in size and data annotation quality. This brings in the research question of how to systematically relate studies contained within those repositories. By allowing data to be re-used on a mass scale, researchers will be able to access a meaningful biological context to aid in the planning and analysis of new studies. This will in turn increase the statistical power of novel studies and put biological results in the context of previous studies. Most repositories contain basic text search functionalities that allow retrieving studies whose textual descriptions contain certain keywords (e.g. 'cancer'). This paradigm has several shortcomings. First, the textual description of an experiment or its results is not as information-rich as the actual data itself. Secondly, it does not provide any solution with respect to analyzing the retrieved study and rigorously comparing it to a novel study. In collaboration with the Brazma group at The European Bioinformatics Institute, which has created and maintains the ArrayExpress database, we are working towards developing machine learning methods that relate studies through their actual expression data, along with visualization tools that allow exploring and interpreting the results.

Content-based information retrieval for differential expression. Gene expression studies often involve differential expression analyses that allow assessing genes or pathways for consistent changes in expression in a phenotype of interest (e.g. cancerous tissue

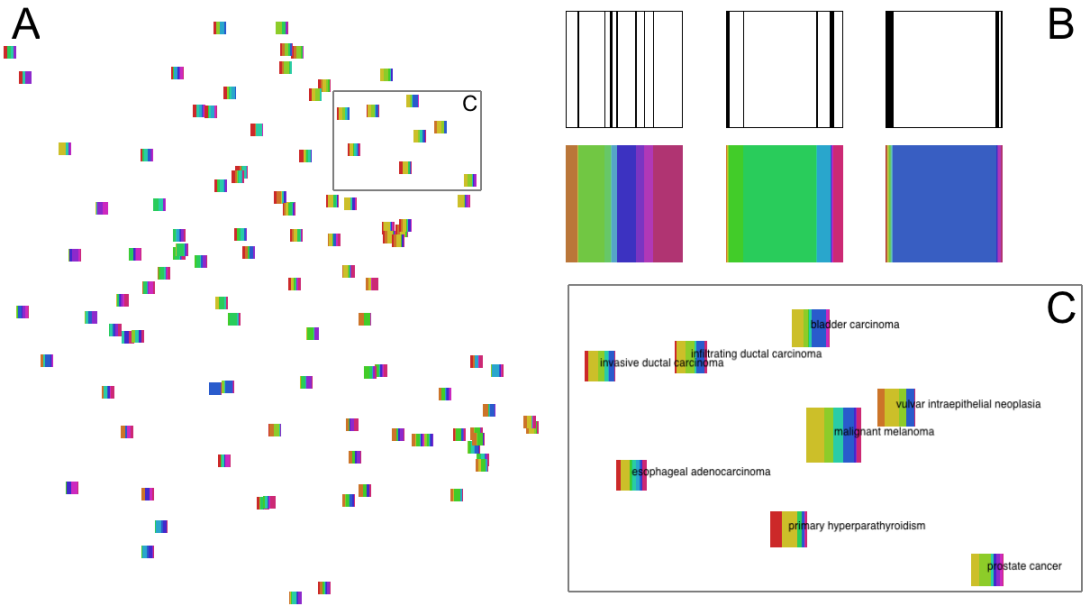


Figure 5.3: Figure taken from Caldas *et al.* [2]. (A) The experiment collection visualized as glyphs on a plane. (B) The method represents each experiment as a collection of so-called biological topics or components. Slice color and width represent the importance of each component in an experiment. (C) Enlarged region from (A) where glyphs have additionally been scaled according to their relevance to a query with a malignant melanoma experiment shown in the center. The surrounding experiments are either from cancer or from primary hyperparathyroidism, which is known to be associated with a higher incidence of cancer.

versus healthy tissue). Recently, it has been shown that differential expression analysis at the level of pathways or gene sets leads to improved and more robust results than differential expression at the gene level [15]. As the first prototype of our biological content-based information retrieval paradigm, we have developed a method that allows relating studies in a repository through shared patterns of gene set differential expression, using a combination of state-of-the-art nonparametric statistics and machine learning approaches [2]. We have also developed novel visualization tools that allow exploring both the data and the retrieval results. Our results show that, given a so-called query study, the method provides a set of other studies where most target the same biological question (e.g. cancer studies) (Fig.5.3). It is also able to find highly non-trivial relations between significantly different pathological entities which were confirmed in the literature. Finally, the retrieval results are interpretable, in the sense that the method provides the patterns of differential expression that are responsible for the inferred relevances (Fig.5.4). Although there has been previous work related to large-scale analysis of differential expression, ours is the first to highlight the potential of performing content-based, interpretable information retrieval in a rigorous and principled manner.

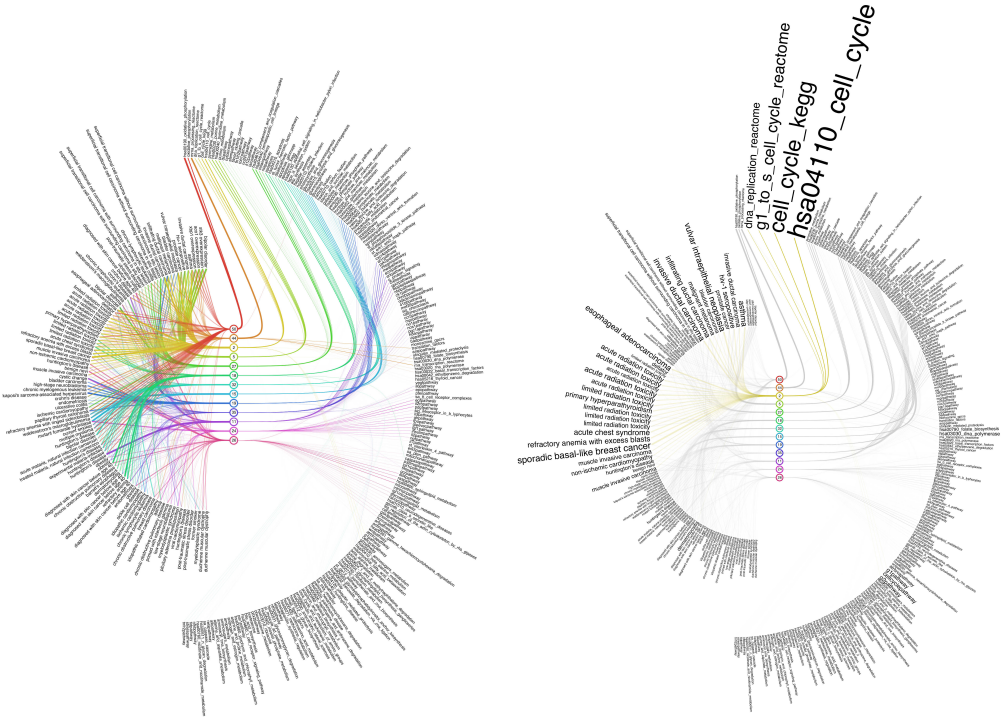


Figure 5.4: Figure taken from Caldas *et al.* [2]. A circular visualization that associates biological studies (labels on the left) to recurrent patterns (circles in the middle), and recurrent patterns to known biological pathways (labels on the right). The left figure shows the general visualization; the degree of association between studies, patterns, and pathways is encoded through edge opacity, where each color is specific to one pattern. In the right figure, the labels of both biological experiments and pathways are scaled according to the degree of association with recurrent pattern number 2. The figure shows an association between several cancer studies (e.g. sporadic basal-like breast cancer) and a collection of cell cycle-related biological pathways.

5.4 Fusion of heterogeneous biomedical data

A living cell is an extremely complex system, and hence integration of information from multiple sources is needed for revealing the true potential of the modern high-throughput measurement methods, such as gene expression or micro-RNA data, combined with relational information of the genes, environmental factors, and disease.

Much of the blooming data integration literature focuses either on well-targeted combinations of sources, such as using sequence-based regulators for explaining gene expression, or on well-focused prediction tasks such as predicting molecular interactions from several data sources. We have focused on knowledge discovery-types of problems where the goal is to discover what is relevant in massive data sets by searching for connections between data sources. This will become more concrete below. Additionally, we have worked on more specific but application-wise interesting problems, such as detection and analysis of deficiencies in the measurements [8].

Dependency modeling. We consider the data fusion problem of combining two or more data sources where each source consists of vector-valued measurements from the same object or entities but on different variables. The task is to include only those aspects which

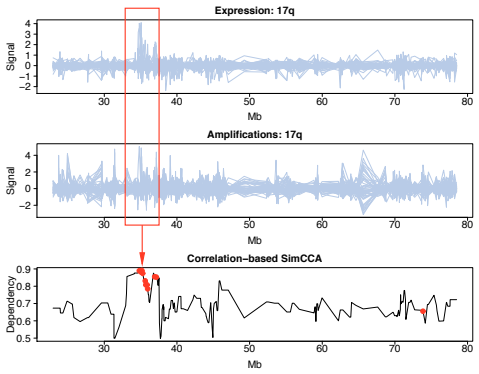


Figure 5.5: Gene expression, copy number signal, and the dependency between the two data sources along chromosome arm 17q in gastric cancer patients. The model detects known cancer associated genes (red dots) with high sensitivity. The Figure is from [9].

are *mutually* informative of each other. This task of including only shared aspects of data sources is motivated through two interrelated lines of thought. The first is noise reduction. If the data sources are measurements of the same entity corrupted by independent noise, discarding source-specific aspects will discard the noise and leave the shared properties that describe the shared entity. The second motivation is to analyze what is interesting in the data. We have investigated, for example, functional effects of DNA mutations by observing dependencies between gene expression and copy number levels across cancer genomes [9]. In these works, the shared variation of the data sources is of primary interest (Figure 5.5). The data set specific effects, often regarded as "noise", may have specific structure; its definition is simply that it is source-specific. The data set specific effects can also be of interest in certain applications. In [14], for example, the decomposition of gene and protein expression levels into shared and data set specific effects was used to distinguish between pre- and post-translational regulation.

We have developed novel ways to bring in prior information to dependency modeling tasks [9]. This helps to reduce overfitting and focus the modeling on the interesting parts of the data, which is critical in many biomedical applications with small sample sizes. We have also released an open-source software package for general fusion of biological data sets, using generalized canonical correlation analysis for both combining the data sets and finding a lower-dimensional representation for them [16].

Dependency models are potentially applicable for modeling other regulatory mechanisms such as transcription factors [7], or micro-RNAs that form a recently discovered and central class of cellular regulators. While causality and confounding factors are often unknown in these studies, detection of statistical dependencies provides a useful proxy for such effects. Future extensions of the dependency models will provide tools to detect multi-level relations between various regulatory mechanisms and gene activity.

Matching of entities. Most data fusion approaches assume co-occurring data sources, in the sense that all sources are different representations of the same entities. For example, in joint analysis of several mRNA experiments we assume the same set of genes has been measured in each experiment. Due to heterogenity of the biological data sources this assumption, however, does not always hold. If the experiments have been measured with different platforms the mapping between the probes might not be perfectly known, or in translational medicine the sources (tissues or species) might even have different entities altogether.

We have introduced a novel method that learns the matching of the entities in a data-driven way, using the actual measurements to find the co-occurrences [17]. The method is based on a very intuitive principle: The matching that gives maximal statistical dependency between the sources is most likely to be correct. The approach can either be used to complement a partial match found based on auxiliary data sources (such as sequence similarity when matching probes of two microarrays), or even to learn the matching from scratch.

Bayesian biclustering. Biclustering is the computational task of simultaneously clustering objects and inferring which features of the objects contribute to the grouping. It is a highly relevant area in gene expression bioinformatics, when one aims at finding restricted biological conditions where certain genes exhibit similar behavior, or alternatively at finding groups of genes with respect to which a set of biological conditions is similar. It is also deeply connected to the fields of content-based information retrieval and data fusion.

We have first adapted an existing promising model to the Bayesian framework, allowing the model to handle noise and endowing it with a rigorous inference engine [1]. More recently, we have developed a hierarchical nonparametric biclustering method [4]. Using recent advances in probabilistic machine learning, our method is able to generate a flexible tree structure of biclusters while keeping computations feasible. We showed that the model achieves state-of-the-art performance on a large data set, and that the model naturally lends itself to hierarchical content-based information retrieval. Finally, we highlighted how the information retrieval functionality can be used to mine for novel biological knowledge, via a case study that provides insight into the potential novel role of miR-224 in the association between melanoma and non-Hodgkin lymphoma.

Searching for functional modules. Functional gene modules and protein complexes have been sought from both protein-protein interaction and gene expression data with various clustering-type methods. We have devised a combined generative model for these data that directly models noise in both data types [13]. The model outperforms other state-of-the-art methods in the task of discovering functional modules. In addition, it is able to detect overlapping modules, in which proteins may have different roles.

References

[1] J. Caldas and S. Kaski Bayesian Biclustering with the Plaid Model. In J. Príncipe, D. Erdogmus and T. Adali, editors, *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing XVIII*, 2008.

[2] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25(12):i145–i153, 2009.

[3] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *BMC Bioinformatics*, 10(suppl.13):P1, 2009.

[4] J. Caldas and S. Kaski. Generative tree biclustering for information retrieval and microrna biomarker discovery. In *Proceedings of the 14th International Conference on Research in Computational Molecular Biology*, 2010. To appear.

[5] I. Huopaniemi, T. Suvitaival, J. Nikkilä, S. Kaski, and M. Orešič. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19(2):261–276, 2009.

[6] I. Huopaniemi and T. Suvitaival and J. Nikkilä and M. Orešič and S. Kaski. Multi-Way, Multi-View Learning. In *NIPS 2009 workshop on Learning from Multiple Sources with Applications to Robotics*, 2009

[7] P. Jaspers, T. Blomster, M. Brosché J. Salojärvi, R. Ahlfors, J. Vainonen, R. Reddy, R. Immink, G. Angenent, F. Turck, K. Overmyer, and J. Kangasjärvi. Unequally redundant rcd1 and sro1 mediate stress and developmental responses and interact with transcription factors. *The Plant Journal*, 60(2):268–279, 2009.

[8] L. Lahti, L.L. Elo, T. Aittokallio, and S. Kaski. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, to appear.

[9] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski. Dependency detection with similarity constraints. In *Proc. MLSP 2009 IEEE International Workshop on Machine Learning for Signal Processing*, 2009.

[10] J. Nikkilä, M. Sysi-Aho, A. Ermolov, T. Seppänen-Laakso, O. Simell, S. Kaski, and M. Orešič. Gender dependent progression of systemic metabolic states in early childhood. *Molecular Systems Biology*, 4:197, 2008.

[11] M. Orešič *et al.* Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. *Journal of Experimental Medicine*, 205(13):2975–2984, 2008.

[12] H. Parkinson *et al.* Arrayexpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–D872, 2009.

[13] J. Parkkinen and S. Kaski. Searching for functional gene modules with interaction component models. *BMC Systems Biology* 2010 4:4.

[14] S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski. Infinite Factorization of Multiple Non-parametric Views. *Machine Learning*, 79(1–2):201–226, 2010.

[15] A. Subramanian *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, page 0506580102, 2005.

[16] A. Tripathi, A. Klami, and S. Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

[17] A. Tripathi, A. Klami, and S. Kaski. Using dependencies to pair samples for multi-view learning. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1561–1564, 2009.

Chapter 6

Neuroinformatics

Ricardo Vigário, Miguel Almeida, Nicolau Gonçalves, Nima Reyhani, Jarkko Ylipaavalniemi, Elina Karp, Jayaprakash Rajasekharan, Jyri Soppela, Janne Nikkilä, Eerika Savia, Samuel Kaski, Erkki Oja

6.1 Introduction

Neuroinformatics has been defined as *the combination of neuroscience and information sciences to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain.* Aside from the development of new tools, the fields of application include often the analysis and modelling of neuronal behaviour, as well as the efficient handling and mining of scientific databases. The group aims at proposing algorithmic and methodological solutions for the analysis of elements and networks of functional brain activity, addressing several forms of communication mechanisms. Motivation and application areas include the understanding of ongoing brain activity and the neuronal responses to complex natural stimulation.

From a methodological viewpoint, the neuroinformatics group has studied properties of source separation methods, such as their reliability and extensions to subspaces. We have also assessed the suitability of such methods to the analysis of electrophysiological recordings (EEG and MEG), and functional magnetic resonance images (fMRI). We proposed also methods for the study of phase synchrony within the central nervous system, and between this and the peripheral nervous system. We have also developed methods for the analysis of neural responses of natural stimulation, based on a novel approach of capturing statistical dependencies between brain activity and the stimulus itself.

In addition to the analysis of fMRI recordings from natural stimulation, we have been also involved in the analysis of single trial event-related MEG data. Albeit its significantly higher temporal resolution, the signal-to-noise ratios are typically very poor, and averaging across hundreds of stimuli is often required. We currently search as well for efficient tissue segmentation of structural MRI.

Although not a natural research topic in neuroinformatics, we have also been involved in the study of phonocardiographic signals, to detect the sources and severities of cardiac murmurs in infants. The signal processing is a challenging one, and a successful application should have a great societal impact.

In addition to these ongoing but stable research topics, we have made a pilot in document mining. The goal is to extract, in a semi-automatic manner, functional information from neuroscience journals, hence reducing the dependence on curator intervention.

Research reported in this section has been carried out in collaboration with experts in neuroscience and cardiology. In the following, we highlight some of the results attained in the reported years.

References

[1] Ylipaavalniemi, J., E. Savia, R. Hari, R. Vigário, and S. Kaski. Towards True Brain Correlates: Dependencies Between Stimuli and Independent fMRI Sources. *Neuroimage 48*, 176–185, 2009.

[2] Almeida, M. and R. Vigário. Source separation of phase-locked subspaces. In *Proc. 8th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2009)*, Paraty, Brazil, pp. 203–210, 2009.

[3] Karp, E., L. Parkkonen, and R. Vigário. Denoising single trial event related magnetoencephalographic recordings. In *Proc. 8th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2009)*, Paraty, Brazil, pp. 203–210, 2009.

[4] Ylipaavalniemi, J., and J. Soppela. Arabica: Robust ICA in a Pipeline. In *Proc. 8th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2009)*, Paraty, Brazil, pp. 379–386, 2009.

[5] Gonçalves, N., Nikkilä, J., and R. Vigário. Partial clustering for tissue segmentation in MRI. To appear in *Proc. 15th Int. Conf. on Neuro- Information Processing (ICONIP'2008)*, Auckland, New Zealand, 2008.

[6] Gonçalves, N., and R. Vigário. Semi-automatic approach for brain tissue segmentation using MRI. In *1st INCF Congress of Neuroinformatics: Databasing and Modeling the Brain (Neuroinformatics 2008)*, Stockholm, Sweden, pp. 106, 2008.

[7] Vigário, R. and E. Oja. BSS and ICA in Neuroinformatics: From current practices to open challenges. *IEEE Rev. Biomed. Eng. 1*, 50–61, 2008.

[8] Ylipaavalniemi, J., and R. Vigário. Analyzing Consistency of Independent Components — An fMRI Illustration. *Neuroimage 39*, 169–180, 2008.

[9] Ylipaavalniemi, J., and R. Vigário. Matching complex activation patterns with features of natural stimuli. In *1st INCF Congress of Neuroinformatics: Databasing and Modeling the Brain (Neuroinformatics 2008)*, Stockholm, Sweden, pp. 71, 2008.

6.2 Complex neural responses to complex stimuli

Natural stimuli are increasingly used in fMRI studies to imitate real-life situations. Consequently, challenges are created for novel analysis methods, including new machine learning tools. With natural stimuli it is no longer feasible to assume single features of the experimental design alone to account for the brain activity. Instead, relevant combinations of rich-enough stimulus features could explain the more complex activation patterns.

We have proposed a novel two-step approach, where independent component analysis is first used to identify spatially independent brain processes, which we refer to as *functional patterns*. As the second step, temporal dependencies between stimuli and functional patterns are detected using either canonical correlation analysis (a journal article in NeuroImage) or its distribution-free variant Nonparametric Dependent Component Analysis (DeCA, a conference article in ICASSP'09). Our method looks for combinations of stimulus features and the corresponding combinations of functional patterns.

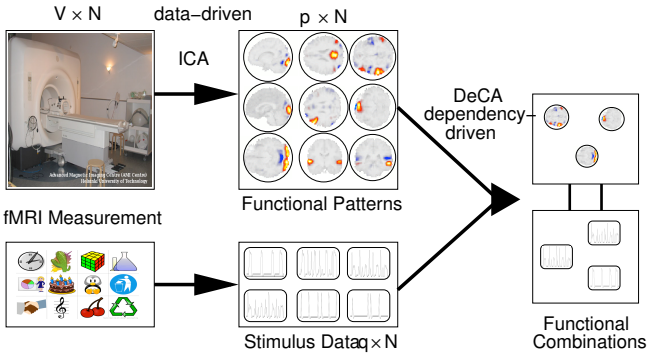


Figure 6.1: Sketch of the framework. (V = # of voxels, N = # of measurement time points, p = # of reliable ICA components, q = # of stimulus time courses).

This two-step approach has been used to analyze measurements from a fMRI study during multi-modal stimulation, in collaboration with Riitta Hari and co-workers. It seems promising to analyze data using natural stimulation.

6.3 Phase synchrony

Interest in phase synchronization phenomena has a long history, when studying the interaction of complex, natural or artificial, dynamic systems. Although not completely adopted, synchronization was attributed a role in the interplay between different parts of the central nervous system as well as across central and peripheral nervous systems. Such phenomena can be quantified by the phase locking factor, which requires knowledge of the instantaneous phase of an observed signal.

During the reported years, we extended the set of algorithmic tools for the identification of phase synchronous phenomena. Our earlier methods dealt with the extraction of sources phase-locked to a reference signal and the clustering of a population of oscillators into synchronous sub-populations. We proposed now a method for the extraction of phase-locked subspaces, following an approach akin to the underlying considerations in independent component analysis.

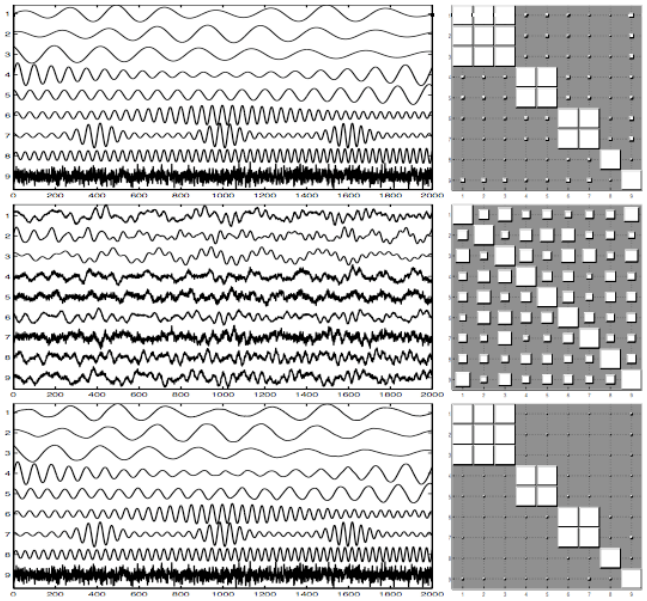


Figure 6.2: Results of one run of IPA: original sources (top left) and PLFs between them (top right); mixed signals (middle left) and PLFs between them (middle right); ex- tracted sources (bottom left) and PLFs between them (bottom right). Results obtained for ? = 0.2, after manually compensating for permutation, scaling and sign of the extracted sources. The Amari Performance Index was 0.06.

6.4 Single trial event related studies

Functional brain mapping is often performed by analysing neuronal responses evoked by external stimulation. Assuming constant brain responses to repeated identical stimuli, averaging across trials is usually applied to improve typically poor signal-to-noise ratios. However, since wave shape and latency vary from trial to trial, information is lost when averaging.

To mitigate this problem, and enable the identification of inter-trial signal variations, we proposed a method to correct the trial-to-trial jitters, in a visually evoked MEG study. The approach was based on a template-based denoising source separation framework. The results were physiologically plausible and presented a clear improvement compared to the classical averaging. We are currently searching as well for a competing approach to estimate, in addition to the jitters, variations in amplitude for each trial's evoked signal.

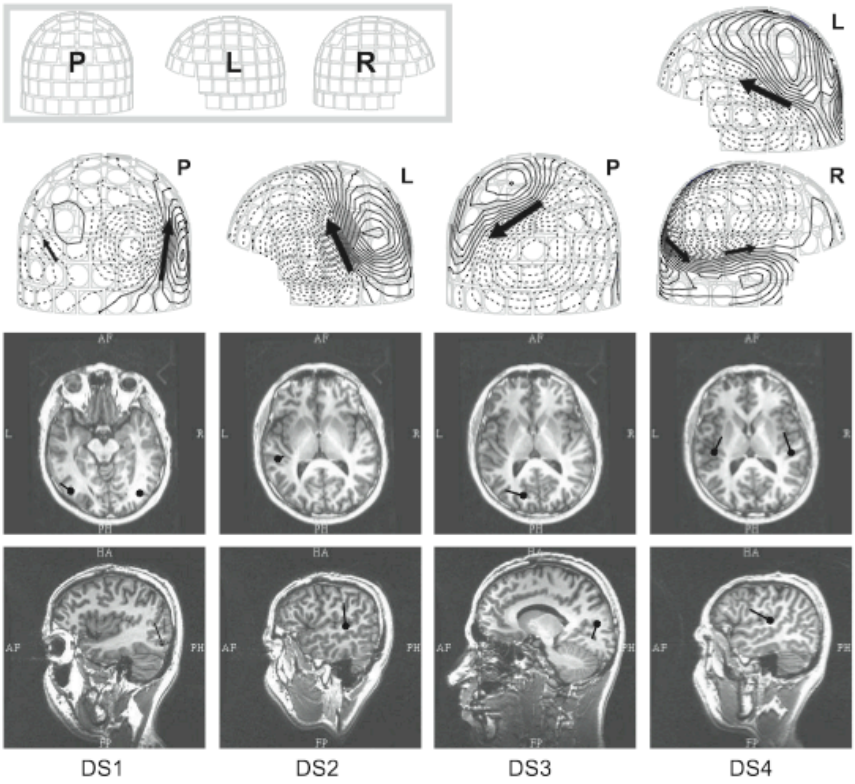


Figure 6.3: Fieldpatterns associated with the denoised sources (DS) and the locations of corresponding equivalent current dipoles. P stands for posterior, R forright and L for left view.

6.5 Tissue segmentation in MRI

MRI is a widely used clinical imaging technique, able to distinguish between soft tissues with an exceptional spatial resolution. In most clinical applications, several settings are used, depending on the targeted tissue, leading to a multi-spectral image set. Automatic follow-up of pathologies in the brain should make full use of the multi-spectra, and be capable of clear segmentation of each tissue.

Our approach is based on the discriminative clustering (DC) algorithm. Since DC is a supervised method, it requires labelled training data, which we produce through a boosting use of the self-organising maps. We achieved absolute classification results in par with the best methods currently in use. As a result of the clustering approach, partial volume information for each image voxel is available, and degenerative pathologies can be better assessed.

Figure 6.4: Classification result for each class of the brainweb data set. CSF, white and grey matter from left to right. The classification is shown overlaying a T2 sequence. Voxels in black correspond to the voxels that have most of their membership in that class.

Tissue	CSF	White	Gray	MS Lesion	Misclassification rate
normal-set	99.70%	97.05%	96.42%	–	3.11%
lesion-set	96.61%	98.90%	98.26%	34.38%	2.61%

Figure 6.5: Numerical results of the DC classification. The percentages shown correspond to the amount of voxels correctly classified.

6.6 Document mining

There is an ever increase in the number of scientific publications in many areas in general, and in neurosciences in particular. Hundreds of articles are published each month. When comparing the results one obtains with a given experimental setup and existing information in literature, one may validate, integrate or confront different opinions and theories. The compilation of such a vast amount of information is not only crucial, but currently also rather human-intensive.

With that in mind, we have conducted a pilot study on document mining of journal publications reporting results on fMRI experiments. We have focused on the image content of the articles. The rather positive preliminary results suggest that a more systematic use of the methodology, and its improvement may help as well reducing the amount of curating work required for the construction of functional databases.

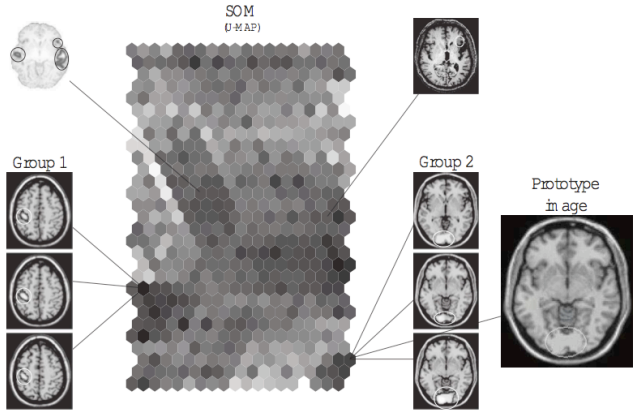


Figure 6.6: Self Organizing Map $\tilde{\text{N}}$ U-matrix trained with 16 dimensional feature vectors, from a set of 100 images extracted from 11 journal papers. Two distinct cluster regions are observed at the lower left and right sides of the map. The prototype image, depicted in the upper left corner fits the expected cluster.

Multimodal interfaces

Chapter 7

Content-based information retrieval and analysis

Erkki Oja, Jorma Laaksonen, Markus Koskela, Zhirong Yang, Ville Viitaniemi, Mats Sjöberg, He Zhang

7.1 Introduction

The Content-Based Information Retrieval Research Group studies and develops efficient methods for content-based information retrieval (CBIR) and analysis tasks and implements them in the PicSOM[1] CBIR system. During the years 2008 and 2009, the PicSOM search engine has been used in various old and new applications.

In the PicSOM CBIR system, parallel Self-Organizing Maps (SOMs) and Support Vector Machine (SVM) classifiers have been trained with separate data sets obtained from the multimodal object data with different feature extraction techniques. The different classifiers and their underlying feature extraction schemes impose different similarity functions on the images, videos, texts and other media objects. In the PicSOM approach, the system is able to discover those of the parallel classifiers that provide the most valuable information for retrieving relevant objects in each particular query.

7.2 Semantic concept detection from images and videos

Extracting semantic concepts from multimedia data has been studied intensively in recent years. The aim of the research on the multimedia retrieval research community has been to facilitate semantic indexing and concept-based retrieval of unannotated multimedia content. The modeling of mid-level semantic concepts is often essential in supporting high-level indexing and querying on multimedia data as such concept models can be trained off-line with considerably more positive and negative examples than what are available at query time.

In the course of previous years we have outlined and implemented our generic PicSOM system architecture for multimedia retrieval tasks. Detection of concepts from multimedia data—e.g. images and video shots—forms an important part of the architecture. In the PicSOM system, concept detection problem is formulated as a standard supervised learning problem. Our concept detection technology is fundamentally based on fusion of a large number of elementary detections, each based on a different low-level audiovisual feature extracted from the multimedia data. During the reporting period 2008–2009 we have continued to develop the technologies involved in the concept detection component of the PicSOM architecture. We have also successfully evaluated the concept detection performance of the PicSOM architecture by participating in international evaluation campaigns. These include PASCAL NoE Visual Object Classes (VOC) Challenge 2008 image analysis evaluation [1] and the annual TRECVID video analysis evaluations [2, 3]. Figure 7.1 illustrates the architecture for detecting concepts from video shots that was used in our TRECVID 2009 system.

We have recently enriched the set of audiovisual features that we use as the basis for for concept detection. As a part of that work, we have studied various aspects and extensions of the bag of visual words (BoV) model. In the BoV model images are represented with histograms of local image features. In our studies of BoV features we have addressed e.g. methods for quantisation of local image features [4, 5, 6], their distance measures [7] and spatial extensions of the BoV methodology [8]. In addition to feature extraction, we have continuosly developed the techniques for feature-wise elementary detection and detector fusion. We have also developed inter-shot temporal and cross-concept techniques for taking into account the dependecies that temporally adjacent video shots on one hand, and related concepts on the other hand typically exhibit [9].

[1] http://www.cis.hut.fi/picsom

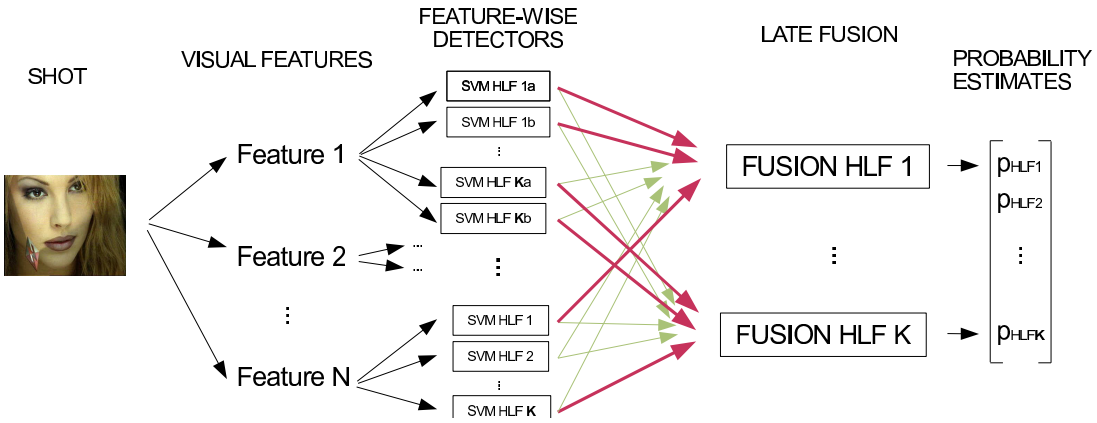


Figure 7.1: Fusion-based shot-wise concept detection module of the PicSOM system that was used for participating in the TRECVID 2009 evaluation.

7.3 Video search and retrieval

In recent studies it has been observed that, despite the fact that the accuracy of the concept detectors may be far from perfect, they can still be useful in supporting *high-level indexing and querying* on multimedia data [10]. We have found this to be true in particular for video search [11]. This is mainly because such semantic concept detectors can be trained off-line with computationally more demanding algorithms and considerably more training examples than what are typically available during interactive use.

Figure 7.2 gives an overview of the automatic video search process within PicSOM, with a detailed view of the concept-based submodule. In the top part of the figure a search query is presented, typically containing a *text query* and possibly also *visual examples*. The visual examples may consist of videos and/or images, demonstrating the visual properties of the desired retrieval response. Either or both of these two modalities of the search query are then used as input to the three parallel submodules of the search system: *text search*, *concept-based search* and *content-based search*. Based on its input, each module produces an estimate of the relevance of each database video to the given query. These scores are finally fused to produce the final search result which is a list of video shots ordered with decreasing estimated relevance to the query.

An important catalyst for research in video retrieval is provided by the annual TREC Video Retrieval Evaluation (TRECVID) workshop. The goal of the workshop series is to encourage research in multimedia retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested to compare their results. The search task in TRECVID models the task of an intelligence analyst who is looking for specific segments of video containing persons, objects, events, locations, etc. of current interest. The task is defined as follows: given a search test collection and a multimedia statement of information need, return a ranked list of shots which best satisfy the need.

We have successfully participated in TRECVID annually since 2005. In TRECVID 2008 we participated in the high-level feature extraction, automatic search, video summarization, and video copy detection tasks, using the PicSOM system framework [12]. In the high-level feature extraction experiments, we used SOM-based semantic concept modeling followed by a post-processing stage that utilizes the concepts' temporal and inter-concept co-occurrences. We also studied the effects of a more comprehensive feature selection scheme and the inclusion of audio features and face detection. The results show that more thorough feature selection can be useful, and that the temporal and inter-concept

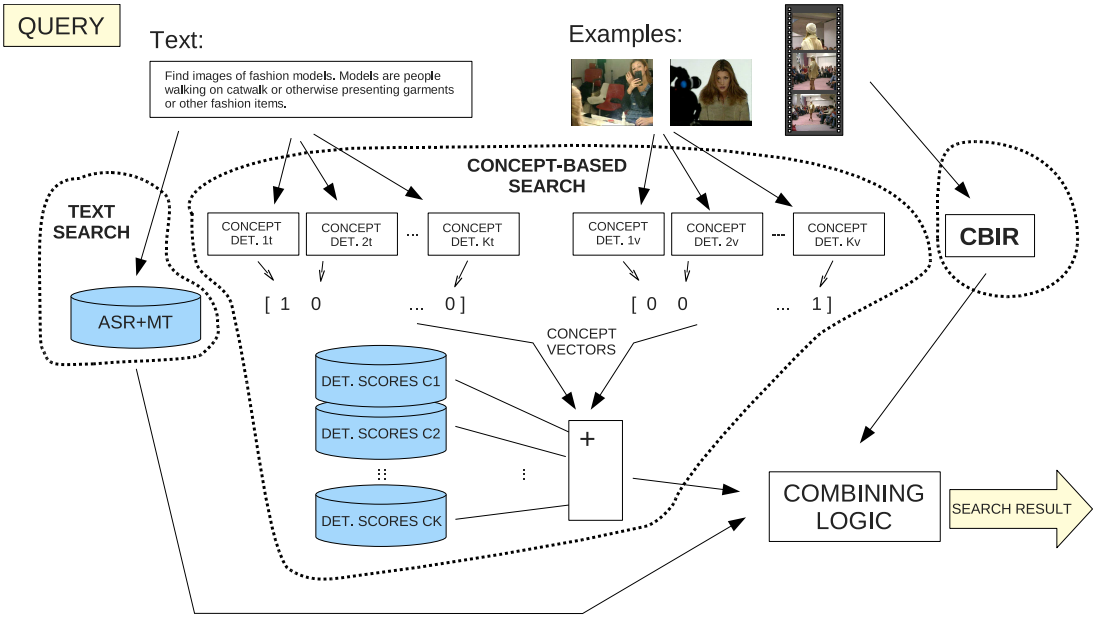


Figure 7.2: General architecture of the PicSOM search module.

co-occurrence analysis has the potential to improve the performance if good concept-wise post-processors can be chosen [13]. The use of audio features and face detection resulted in minor improvements. In TRECVID 2009 we used greatly improved Support Vector Machine (SVM) detectors for high-level feature extraction task, and our group had the sixth best result among the 20 participanting groups. These results could then be used in the search task, where we achieved the third best result.

7.4 Video analysis applications

In this section, two further applications of the content-based video analysis framework are described. These applications are *automatic video summarization* and *analysis of sign-language*.

Video summarization is a process where an original video file is converted to a considerably shorter form, which can then be used to facilitate efficient searching and browsing in large video collections. The aim of automatic summarization is to preserve as much as possible from the essential content and overall structure.

We have developed a technique for video summarization [14] using SOMs trained with standard visual features that have been applied in various multimedia analysis tasks. The produced summaries consist of collections of selected video clips from the original material. The method is based on initial shot segmentation, with the shots used in the following stages as basic units of processing. We then detect and remove unwanted "junk" shots, and apply face detection, speech detection, and motion activity estimation. Next, we compute the visual similarities between all pairs of shots and remove overly similar shots. We trace the trajectory of the frames within the shot in question and use the trajectory as a signature for the shot, which can then be compared to other shots' signatures to determine whether a shot is visually unique or similar to some other shots. Each remaining shot is then represented in the summary with a separately selected one-second clip.

We participated in the TRECVID 2008 rushes summarization task [2] and obtained

very promising results. Our summarization algorithm obtained average ground-truth inclusion performance with the shortest overall summaries over all the submissions.

We have also applied our methods for video content analysis in a multidisciplinary research project for the recognition and analysis of recorded Finnish Sign Language [15]. Automatic and semi-automatic computer vision techniques are used to recognize and analyze gestures and facial expressions in sign language videos (see Figure 7.3). The aim is to identify linguistic sign and gesture boundaries and to indicate which video sequences correspond to specific signs and gestures. This will facilitate indexing and the construction of an example-based open-access visual corpus of the Finnish Sign Language for which there already exists large amounts of non-indexed video material.

Figure 7.3: An example frame from the sign language video material. Left: Skin-color filtering. Right: Motion tracking.

References

[1] Ville Viitaniemi and Jorma Laaksonen. Techniques for image classification, object detection and object segmentation. In Monica Sebillo, Giuliana Vitiello, and Gerald Schaefer, editors, *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, volume 5188 of *Lecture Notes in Computer Science*, pages 231–234, Salerno, Italy, September 2008. Springer.

[2] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. PicSOM experiments in TRECVID 2008. In *Proceedings of the TRECVID 2008 Workshop*, Gaithersburg, MD, USA, November 2008.

[3] Mats Sjöberg, Ville Viitaniemi, Markus Koskela, and Jorma Laaksonen. PicSOM experiments in TRECVID 2009. In *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009.

[4] Ville Viitaniemi and Jorma Laaksonen. Experiments on selection of codebooks for local image feature histograms. In Monica Sebillo, Giuliana Vitiello, and Gerald Schaefer, editors, *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, volume 5188 of *Lecture Notes in Computer Science*, pages 126–137, Salerno, Italy, September 2008. Springer Verlag.

[5] Marcin Blachnik and Jorma Laaksonen. Image classification by histogram features created with learning vector quantization. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'08)*, pages 827–836, September 2008.

[6] Ville Viitaniemi and Jorma Laaksonen. Combining local feature histograms of different granularities. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 636–645, Oslo, Norway, 2009. Springer Verlag.

[7] Ville Viitaniemi and Jorma Laaksonen. Representing images with χ^2 distance based histograms of SIFT descriptors. In *Proceedings of 19th International Conference on Artificial Neural Networks (ICANN 2009)*, volume 5768 of *Lecture Notes in Computer Science*, pages 636–645, Limassol, Cyprus, 2009. Springer Verlag.

[8] Ville Viitaniemi and Jorma Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of ACM International Conference on Image and Video Retrieval (CIVR 2009)*, Fira, Greece, July 2009.

[9] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, Austria, May 2008.

[10] Alexander G. Hauptmann, Michael G. Christel, and Rong Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, April 2008.

[11] Markus Koskela, Mats Sjöberg, and Jorma Laaksonen. Improving automatic video retrieval with semantic concept detection. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 480–489, Oslo, Norway, 2009. Springer Verlag.

[12] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. PicSOM experiments in TRECVID 2008. In *Proceedings of the TRECVID 2008 Workshop*, pages 408–416, Gaithersburg, 2008.

[13] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, 2008.

[14] Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Ville Viitaniemi, and Hannes Muurinen. Rushes summarization with self-organizing maps. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 45–49, Augsburg, Germany, September 2007. ACM Press. Available online at `http://dx.doi.org/10.1145/1290031.1290039`.

[15] Markus Koskela, Jorma Laaksonen, Tommi Jantunen, Ritva Takkinen, Päivi Rainò, and Antti Raike. Content-based video analysis and access for finnish sign language – a multidisciplinary research project. In *Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages at 6th Language Resources and Evaluation Conference (LREC 2008)*, pages 101–104, Marrakech, Morocco, May-June 2008.

Chapter 8

Automatic speech recognition

Mikko Kurimo, Kalle Palomäki, Teemu Hirsimäki, Janne Pylkkönen, Ville Turunen, Sami Virpioja, Matti Varjokallio, Ulpu Remes, Heikki Kallasjoki, Reima Karhila, Teemu Ruokolainen, Tanel Alumäe, Sami Keronen, Andre Mansikkaniemi

8.1 Introduction

Automatic speech recognition (ASR) means an automated process that inputs human speech and tries to find out what was said. ASR is useful, for example, in speech-to-text applications (dictation, meeting transcription, etc.), speech-controlled interfaces, search engines for large speech or video archives, and speech-to-speech translation.

Figure 8.1 illustrates the major modules of an ASR system and their relation to applications. In *feature extraction*, signal processing techniques are applied to the speech signal in order to dig out the features that distinguish different phonemes from each other. Given the features extracted from the speech, *acoustic modeling* provides probabilities for different phonemes at different time instants. *Language modeling*, on the other hand, defines what kind of phoneme and word sequences are possible in the target language or application at hand, and what are their probabilities. The acoustic models and language models are used in *decoding* for searching the recognition hypothesis that fits best to the models. Recognition output can then be used in various applications.

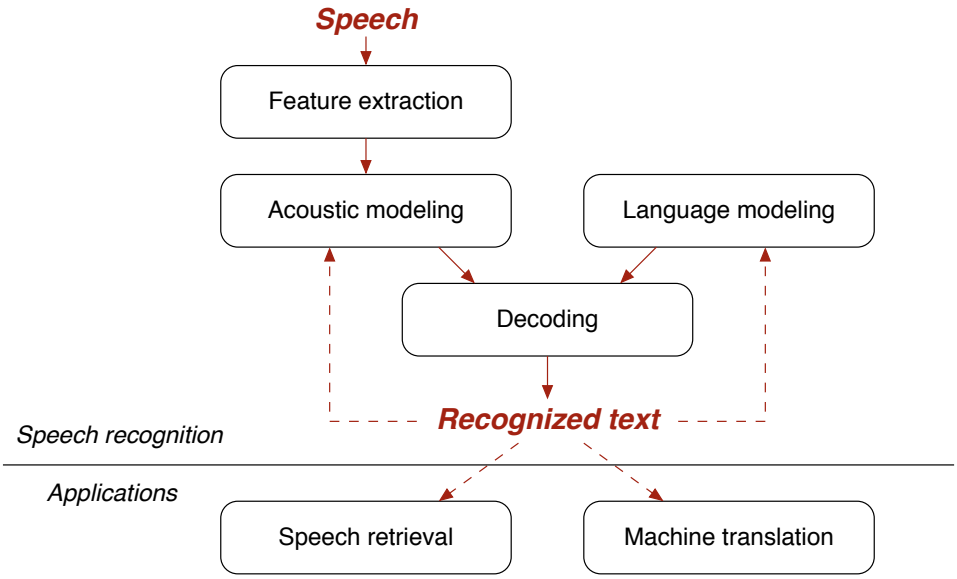


Figure 8.1: The main components of an automatic speech recognition system and their relation to speech retrieval and machine translation applications.

Our focus in ASR is large vocabulary continuous speech recognition (LVCSR). For several years, we have been developing new machine learning algorithms for each of the subfields and building a complete state-of-the-art recognizer to evaluate new methods and their impact. Originally, the recognizer was constructed for fluent and planned speech such as Finnish newsreading, where language models covering a very large vocabulary are required. Besides newsreading, other example tasks are political and academic speeches and other radio and television broadcasts where the language used is near the written style. Sofar, we have not seriously attempted to recognize Finnish spontaneous conversations, because enough Finnish training texts for learning the corresponding style do not exist. Our main training corpus for language modeling is the Finnish Language Bank at CSC. For acoustic modeling we use voice books, Finnish Broadcast Corpus at CSC and the SPEECON corpus.

In addition to the recognition of Finnish, we have performed experiments in English, Turkish and Estonian. To make this possible we have established research relations to

different top speech groups in Europe and U.S., e.g. University of Colorado, International Computer Science Institute ICSI, Stanford Research Institute SRI, IDIAP, University of Edinburgh, University of Sheffield, University of Cambridge, Bogazici University, and Tallinn University of Technology. The forms of collaboration have included researcher exchanges, special courses, workshops and joint research projects. We have also participated in several top international and national research projects funded by EU, Academy of Finland, Tekes, and our industrial partners. In the close collaboration with our Natural Language Processing group 10 we are also organizing an international competition called Morphochallenge to evaluate the best unsupervised segmentation algorithms for words into morphemes for information retrieval, statistical machine translation, LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network and described in Chapter 10.

In the EU FP7 project called EMIME, the aim is to develop new technologies for speech-to-speech systems. Although this has broadened the field of the group to include some aspects of speech synthesis, such as supervised and unsupervised adaptation in the same way as in ASR, text-to-speech (TTS) still plays a minor role compared to the strong ASR focus of the group.

8.2 Acoustic modeling

Acoustic modeling in automatic speech recognition means building statistical models for some meaningful speech units based on the feature vectors computed from speech. In most systems the speech signal is first chunked into overlapping 20-30 ms time windows at every 10 ms and the spectral representation is computed from each frame. A commonly used feature vector consists of mel-frequency cepstral coefficients (MFCC) which are the result of the discrete cosine transform (DCT) applied to the logarithmic mel-scaled filter bank energies. Local temporal dynamics can be captured by concatenating the first and second order delta features (time differences) to the basic feature vector.

The acoustic feature sequence in ASR is typically modeled using hidden Markov models (HMM). In basic form each phoneme is modeled by a separate HMM, where the emission distributions of the HMM states are Gaussian mixtures (GMMs). In practice, however, we need to take the phoneme context into account, so that for each phoneme there are separate HMMs for various phoneme contexts. This leads easily to very complex acoustic models where the number of parameters is in order of millions.

Estimating the parameters of complex HMM-GMM acoustic models is a very challenging task. Traditionally maximum likelihood (ML) estimation has been used, which offers simple and efficient re-estimation formulae for the parameters. However, ML estimation does not provide optimal parameter values for classification tasks such as ASR. Instead, discriminative training techniques are nowadays the state-of-the-art methods for estimating the parameters of acoustic models. They offer more detailed optimization criteria to match the estimation process with the actual recognition task. The drawback is increased computational complexity. Our implementation of the discriminative acoustic model training allows using several different training criteria such as maximum mutual information (MMI) and minimum phone error (MPE) [1]. It also enables alternative optimization methods such as gradient based optimization and constrained line search [2] in addition to the commonly used extended Baum-Welch method.

Our recent research has taken advantage of the flexibility of our system to use different discriminative training criteria by comparing different discriminative training methods in various configurations [3]. The research showed some guidelines in how to apply certain discriminative training methods in large scale acoustic model estimation.

The speech syntesis work related to the EMIME EU/FP7 project concentrates on the adaptation of HMM-based TTS models. The goal of the project is to personalize the output voice of a cross-lingual speech-to-speech system, to make it resemble the voice of the original speaker.

The features and models of TTS systems differ somewhat from those used in ASR. A shorter timestep, typically 5 ms is used, and the count of cepstral coefficients is twice or thrice that of typical ASR features. The acoustic models do not use GMMs - simple single-Gaussian models are used, but the amount of models is much higher. The TTS models are context-dependent on a more complicated level compared to the ASR models. A single phoneme has different models depending on its position within a word, syllable and sentence, as well as the surrounding phonemes.

Training acoustic models for high-quality voice for a TTS system therefore requires data of close to 1000 high-quality sentences from the target speaker. As this much data is not available in the target application of the project, the only feasible option is to train an average TTS voice and use adaptation techniques to change it to resemble the target speakers voice.

The adaptation of HMM-based TTS models is very similar to adaptation of ASR models. Maximum a posteriori (MAP) linear transformations are applied in similar fashion to

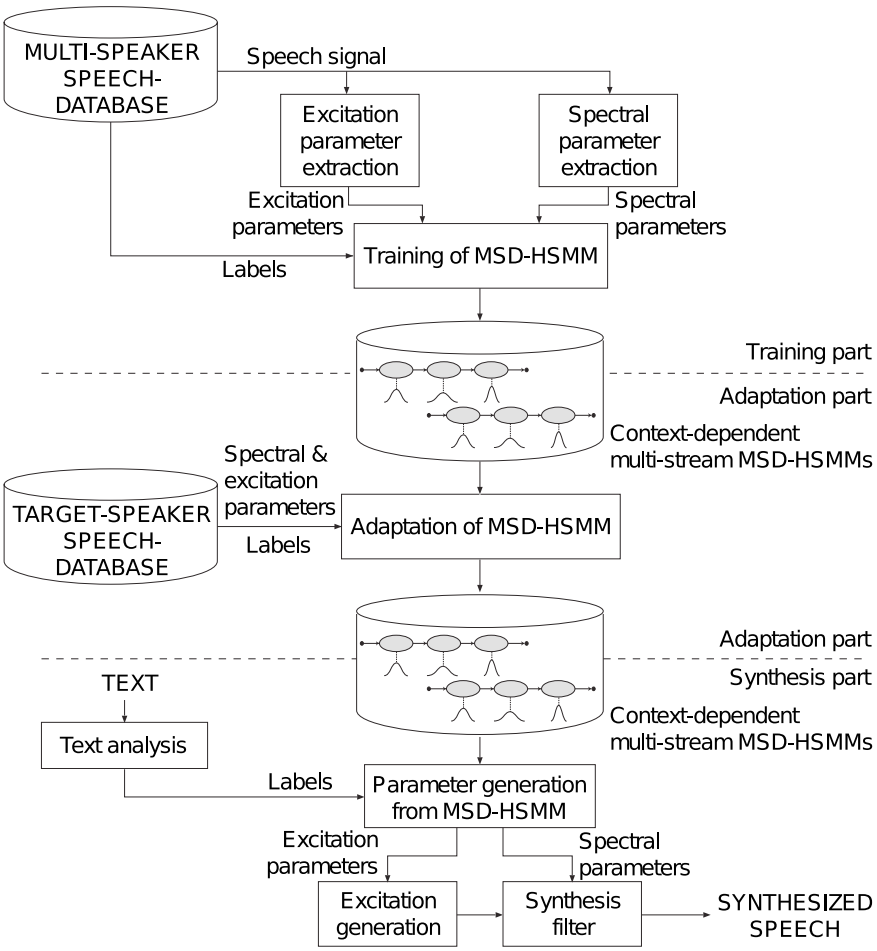


Figure 8.2: The HTS speech synthesis system for generating an average voice, adapting it to a target speaker and creating synthesized speech. From [4].

ASR adaptation. A collaborative investigation using data from several languages showed that adapting a general voice is a practical and effective way to mimic a target speaker's voice[4].

References

[1] D. Povey and P. C. Woodland, Minimum Phone Error and I-smoothing for Improved Discriminative Training. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, pages I-105–108, 2002.

[2] P. Liu, C. Liu, H. Jiang, F. Soong, and R.-H. Wang, A constrained line search optimization method for discriminative training of hmms. In *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 900–909, 2008.

[3] J. Pylkkönen, Investigations on Discriminative Training in Large Scale Acoustic Model Estimation. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, pp. 220–223, 2009.

[4] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian,R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, M. Kurimo Thousands of Voices for HMM-based Speech Synthesis. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, pp. 420–423, 2009.

8.3 Language modeling

In topic adaptation of language models, we take into account the underlying topic of speech by elevating the probabilites of the subvocabulary characteristic to its topic. Via topic adaptation, we aim at improving the recognition of topically important words. The potential benefit of topic adaptation relies on the success of retrieving the underlying topic correctly. In the master's thesis [1], we discuss the topic adaptation task in relation to multimodal interfaces. In the multimodal scenario, the contextual cues with which the topic is retrieved can not be assumed reliable nor large in size. The experiments with English large vocabulary speech recognition task showed that topic adaptation with these cue assumptions is feasible. The master's thesis was conducted as a part of projects PinView and UI-ART focusing on multimodal interfaces.

For Finnish, estimating the language model probabilities for words is difficult since there is a vast number of different word forms. For example, a single verb has theoretically thousands of inflected word forms. The natural way to attack the problem is to split words into smaller fragments and build the language models on the fragments instead of whole words. Since it is not obvious how the words should be split, we have studied what kind of word fragments are optimal for speech recognition systems. Experiments in Finnish, Turkish and Estonian recognition tasks indicate that an unsupervised data-driven splitting algorithm called Morfessor (see Section 10.2) improves recognition of rare words. [2]

In speech recognition systems solutions to the problem of vocabulary growth in morphologically rich languages proposed in the literature include increasing the size of the vocabulary and segmenting words into morphs. However, in many cases, the methods have only been experimented with low-order n-gram models or compared to word-based models that do not have very large vocabularies. In [3] we study the importance of using high-order variable-length n-gram models when the language models are trained over morphs instead of whole words. Language models trained on a very large vocabulary are compared with models based on different morph segmentations. Speech recognition experiments are carried out on two highly inecting and agglutinative languages, Finnish and Estonian. The results suggest that high-order models can be essential in morph-based speech recognition, even when lattices are generated for two-pass recognition. The analysis of recognition errors [4] reveal that the high-order morph language models improve especially the recognition of previously unseen words.

References

[1] T. Ruokolainen. Topic adaptation for speech recognition in multimodal environment. *Master's thesis, Helsinki University of Technology*, 2009.

[2] E. Arisoy, M. Kurimo, M. Saraclar, T. Hirsimäki, J. Pylkkönen, T. Alumäe and H. Sak. Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages, *Speech Recognition*, pages 193–204. I-Tech, Vienna, Austria, 2008.

[3] T. Hirsimäki, J. Pylkkönen and M. Kurimo. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 17(4):724–732, May 2009.

[4] T. Hirsimäki and M. Kurimo. Analysing recognition errors in unlimited-vocabulary speech recognition. In *Proceedings of the 2009 Annual Conference of the North ameri-*

can Chapter of the association for Computational Linguistics, NAACL 2009, Boulder, CO, May 31 – June 5 2009.

8.4 Applications and tasks

Speech retrieval and indexing

Large amounts of information is produced in spoken form. In addition to TV and radio broadcasts, more and more material is distributed on the Internet in the form of podcasts and video sharing web sites. There is an increasing need for content based retrieval of this material. Speech retrieval systems consist of two parts. First, an automatic speech recognition system is used to transcribe the speech into textual form. Second, an index is built based on this information.

The vocabulary of the speech recognizer limits the possible words that can be retrieved. Any word that is not in the vocabulary will not be recognized correctly and thus can not be used in retrieval. This is especially problematic since the rare words, such as proper names, that may not be in the vocabulary are often the most interesting from retrieval point of view. Our speech retrieval system addresses this problem by using morpheme-like units produced by the Morfessor algorithm. Any word in speech can now potentially be recognized by recognizing its component morphemes. The recognizer transcribes the text as a string of morpheme-like units and these units can also be used as index terms. We have shown that the morph-based approach for speech search suffers significantly less from OOV query words than a word based method [1].

Retrieval performance was further improved by utilizing alternative recognition candidates from the recognizer [1]. Speech recognizers typically produce only the most likely string of words, the 1-best hypothesis. Retrieval performance is decreased if a relevant term is misrecognized and is thus missing from the transcript. However, it is possible that the correct term was considered by the recognizer but was not the top choice. Thus, retrieval performance can be improved by extracting these alternative results from the recognizer and adding them to the index. A *confusion network* [2] provides a convenient representation of the competing terms along with a probability value for each term.

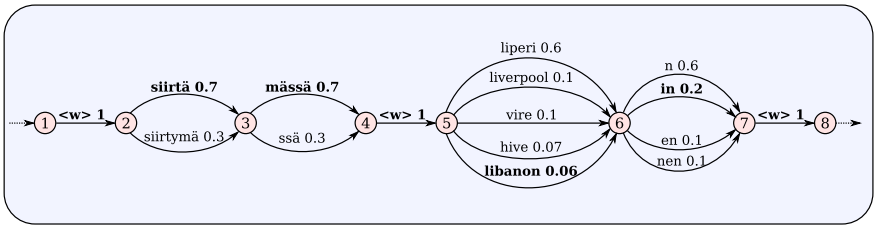


Figure 8.3: A confusion network of alternative recognition candidates for a segment of speech. `<w>` marks a word break boundary. The correct morphs are in bold.

Speech-to-speech translation

Speech-to-speech machine translation is in some ways the peak of natural language processing, in that it deals directly with our (humans') original, oral mode of communication (as opposed to derived written language). As such, it presents several important challenges:

1. Automatic speech recognition of the input using state-of-the-art acoustic and language modeling, adaptation and decoding

2. Statistical machine translation of either the recognized most likely speech transcript or the confusion network or the whole lattice including all the best hypothesis

3. Speech synthesis to turn the translation output into intelligible speech using the state-of-the-art synthesis models and adaptation

4. Intergration of all these components to aim at the best possible output and tolerate errors that may happen in each phase

A pilot study of Finnish-English speech-to-speech translation was carried out in the lab as a joint effort of the Speech Recognition, Natural Language Processing (Ch. 10) and Computational Cognitive Systems (Ch. 11) groups [3]. The domain selected for our experiments was heavily influenced by the available bilingual (Finnish and English) and bimodal (text and speech) data. Because none is readily yet available, we put one together using the Bible. As the first approach we utilized the existing components, and tried to weave them together in an optimal way. To recognize speech into word sequences we applied our morpheme-based unlimited vocabulary continuous speech recognizer [4]. As a Finnish acoustic model the system utilized multi-speaker hidden Markov models with Gaussian mixtures of mel-cepstral input features for state-tied cross-word triphones. The statistical language model was trained using our growing varigram model [5] with unsupervised morpheme-like units derived from Morfessor Baseline [6]. In addition to the Bible the training data included texts from various sources including newspapers, books and newswire stories totally about 150 million words. For translation, we trained the Moses system [7] on the same word and morpheme units as utilized in the language modeling units of our speech recognizer. For speech synthesis, we used Festival [8], including the built-in English voice and a Finnish voice developed at University of Helsinki. Further research on statistical machine translation is described in Section 13.

References

[1] V.T. Turunen. Reducing the effect of OOV query words by using morph-based spoken document retrieval. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, pages 2158–2161, September 2008.

[2] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech And Language*, 14:373–400, 2000.

[3] David Ellis, Mathias Creutz, Timo Honkela, and Mikko Kurimo. Speech to speech machine translation: Biblical chatter from Finnish to English. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 123-130, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing.

[4] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja and Janne Pylkkönen 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language* 20(4):515–541.

[5] Vesa Siivola Language models for automatic speech recognition: construction and complexity control. Doctoral thesis, Dissertations in Computer and Information Science, Report D21, Helsinki University of Technology, Espoo, Finland, 2006.

[6] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[7] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar, Alexandra Constantin, and Evan Herb. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

[8] The Festival Speech Synthesis System. University of Edinburgh. *http://festvox.org*

8.5 Noise robust speech recognition

Missing feature approaches

Using missing feature methods for noise compensation in automatic speech recognition is based on partitioning the compressed spectrographic representation of a noise corrupted speech signal to speech dominated i.e. realiable regions and noise dominated i.e. unreliable regions as illustrated in Figure 8.4. Information in the unreliable regions is assumed missing, so either the speech recognition system should ignore the unreliable components or the missing values be reconstructed using e.g. cluster-based imputation [1]. Experiments reported in [2] suggested that cluster-based imputation can significantly improve LVCSR performance under environmental noise but may not fully allow for simultaneous speaker or environment-based adaptation. We therefore modified the method to account for acoustic model adaptation estimated prior to reconstruction, which improved the speech recognition performance in certain noise environments as discussed in [3]. Additionally, we have been developing missing feature techniques that are particularly robust in the presence of reverberation noise [4, 5] and models that mimic certain principles of human speech recognition especially in the binaural system [6].

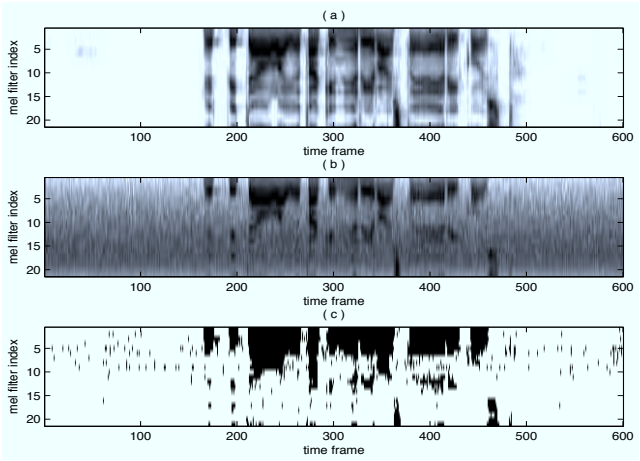


Figure 8.4: Logarithmic mel spectrogram of (a) an utterance recorded in quiet environment and (b) the same utterance corrupted with additive noise. The noise mask (c) constructed for the noisy speech signal indicates the speech dominated regions in black and the noise dominated regions in white.

Noise robust feature representations

One approach to noise robust speech recognition is to search for feature representations that are less affected by changes in environmental noise. In particular, common feature extraction schemes based on the short-time spectrum of the speech signal can be made more robust by using an estimate of the spectral envelope instead.

The stabilised weighted linear prediction (SWLP) signal modeling method [7], recently developed at the Department of Signal Processing and Acoustics at Helsinki University of Technology, was used to implement a more robust variant of the MFCC features currently used by our speech recognition system. The performance of the new features was evaluated in different noisy real-world environments using the SPEECON corpus. Improvements in

recognition rates were found in the case where acoustic models trained using clean speech only were used to recognize speech corrupted by noise [8, 9].

References

[1] B. Raj and R. M. Stern, Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, vol. 22, pages 101–116, 2005.

[2] U. Remes, K. J. Palomäki, and M. Kurimo, Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition. In *Proceedings of the 16th European Signal Processing Conference*, Lausanne, Switzerland, 2008.

[3] U. Remes, K. J. Palomäki, and M. Kurimo, Robust automatic speech recognition using acoustic model adaptation prior to missing feature reconstruction. In *Proceedings of the 17th European Signal Processing Conference*, Glasgow, Scotland, UK, pages 535–539, 2009.

[4] K. J. Palomäki, G. J. Brown and J. Barker, Techniques for handling convolutional distortion with "missing data" automatic speech recognition. *Speech Communication*, vol. 43, pages 123-142, 2004.

[5] G. J. Brown and K. J. Palomäki, A reverberation-robust automatic speech recognition system based on temporal masking (Abstract). *J. Acoust. Soc. Am.*, vol. 123, Acoustics 2008, Paris, France, page 2978, 2008.

[6] K. J. Palomäki and G. J. Brown, A computational model of binaural speech intelligibility level difference (Abstract). *J. Acoust. Soc. Am.*, vol 123, Acoustics 2008, Paris, France, page 3715, 2008.

[7] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, Stabilised weighted linear prediction. *Speech Communication*, volume 51, issue 5, pages 401–411, 2009.

[8] H. Kallasjoki, K. J. Palomäki, C. Magi, P. Alku, and M. Kurimo, Noise robust LVCSR feature extraction based on stabilized weighted linear prediction. In *Proceedings of the 13th International Conference Speech and Computer*, St. Petersburg, Russia, pages 221–225, 2009.

[9] J. Pohjalainen, H. Kallasjoki, P. Alku, K. J. Palomäki, and M. Kurimo, Weighted linear prediction for speech analysis in noisy conditions. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, pages 1315–1318, 2009.

Chapter 9

Proactive Interfaces

Samuel Kaski, Jorma Laaksonen, Mikko Kurimo, Arto Klami, Markus Koskela, Kai Puolamäki, Jarkko Salojärvi, Antti Ajanki, Mats Sjöberg, Ville Viitaniemi, He Zhang, Melih Kandemir, Laszlo Kozma, Lu Wei, Teemu Ruokolainen, Xi Chen, Erkki Oja

9.1 Introduction

The Proactive Interfaces research theme combines efforts of multiple research groups, including the Statistical Machine Learning and Bioinformatics group, lead by Professor Samuel Kaski, the Content-Based Information Retrieval group, lead by Docent Jorma Laaksonen, and the Speech Recognition group, lead by Docent Mikko Kurimo. In 2008, three major collaborative projects, PinView, UI-ART and Diem/MMR, have been launched which together form the AIRC flagship project *Proactive Interfaces*. In 2009, a fourth project, Image Based Linking, has been started.

9.2 Inferring interest from gaze patterns

Proactive systems anticipate the user's intentions and actions, and utilize the predictions to provide more natural and efficient user interfaces. One of the critical components in this loop is inferring the interests of the user, which is a challenging machine learning problem. Successful proactivity in varying contexts requires generalization from past experience. Generalization, on its part, requires suitable powerful (stochastic) models and a collection of data about relevant past history to learn the models.

We focus on inferring the interest and needs of the user from gaze patterns, measured with modern eye-tracking equipment. During complex tasks, such as reading, attention approximately lies on the location of the reader's gaze. Therefore eye movements should contain information, although very noisy, on the reader's interests. As a practical example of what can be inferred from eye movements, [1] uses discriminative Hidden Markov models to detect different processing states in the tasks of simple word search, question-answer, and finding the most interesting topic. The model detects, for example, switches between reading and scanning the text, which in turn helps in predicting the intention of the user.

Another line of work focuses on information retrieval tasks, where the tasks range from estimating relevance of specific text snippets to inferring implicit queries even the user cannot formulate accurately. The eye-movements collected while the user browses the retrieval results are informative of what the user was after, giving an estimated query that can be used for retrieving more relevant documents [2, 3]. The difficult learning problem, termed *learning to learn*, is in finding a regressor from word-level features to queries so that it generalizes to new queries and user interests. [2] solves this by incorporating both the inference of the implicit query and prediction of the relevance of unseen documents into a unified probabilistic model, while [3] utilizes SVM-classifiers in learning the relationships between how words are viewed and their importance for the task. The parameters of the models are optimized to maximize the average performance over a range of training queries, and the resulting query-independent predictors can be applied for topics with no training data.

Going beyond text retrieval tasks, [4] extends the information retrieval work for images, using eye movements for predicting relevance feedback to be used in content-based retrieval. The work is the first demonstration on gaze providing useful information also for media types that are less-structured than text, continued with improved inference and interface in [5].

9.3 Eye-movement enhanced image retrieval

PinView is an EU FP7 funded three-year Collaborative Project started on 1 January 2008 and coordinated by in AIRC. The goal of PinView is a proactive personal information navigator that allows retrieval of multimedia – such as still images, text and video – from

unannotated databases. During image browsing and searching with a task-dependent interface, the PinView system will infer the goals of the user from explicit and implicit feedback signals and interaction (eye movements, pointer traces and clicks, speech) complemented with social filtering. The collected rich multimodal responses from the user are processed with new advanced machine learning methods to infer the implicit topic of the user's interest as well as the sense in which it is interesting in the current context.

The PinView consortium combines pioneering application expertise with a solid machine learning background in content-based information retrieval. Besides AIRC, the project consortium includes University of Southampton (uk), University College London (uk), Montanuniversitaet Leoben (au), Xerox Research Centre Europe (fr), and celum gmbh (au). The publications of the PinView project's first two years are [4, 6, 7, 8, 9, 10, 11, 5].

As part of the project, we have developed novel gaze-based interfaces for image retrieval. The purpose is both to create interfaces that can be used without explicit control devices, which is useful for mobile environments and also for people with motor disabilities, but also to obtain more information from gaze. While gaze is informative of the user's interests in all settings, it is possible to create interfaces that provide more information compared to standard displays. We have developed GaZIR [5], a gaze-based zoomable interface for image retrieval, that can be operated with gaze alone. As explicit control with gaze is highly stressful, the GaZIR system uses gaze primarily for implicit information. Explicit control is used only for zooming in and out, while the actual retrieval feedback is learned implicitly from the gaze patterns and fed to the PinView content-based retrieval engine. Figure 9.1 shows a screenshot of the interface, showing the co-centric circles of images the user is currently browsing. The circle-shaped layout was chosen to break natural habits of browsing grid of images in a structured fashion, and hence to extract more information from the gaze trajectory.

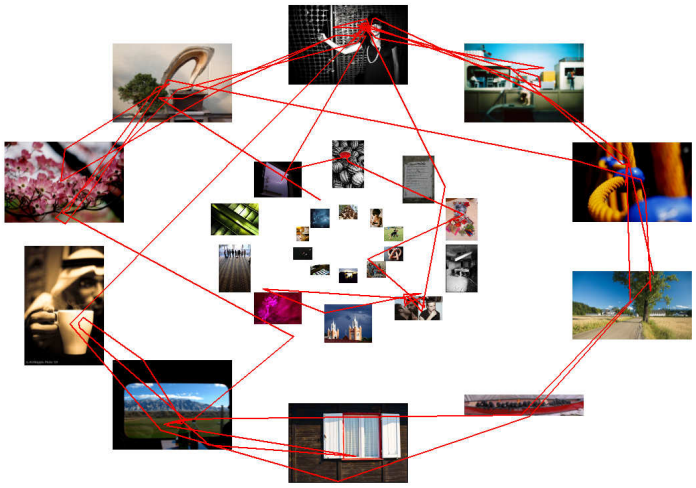


Figure 9.1: GaZIR, the gaze-based zoomable interface for image retrieval, in action. The user sees co-centric circles of images, and the system monitors the gaze of the user (red lines) while he is browsing the images. The subjective relevance of the images is inferred from the gaze trajectory and forwarded to the PicSOM content-based image retrieval system. When the user zooms in, PicSOM returns new sets of images closer to the implicitly defined intents of the user.

9.4 Contextual information interfaces

Contextual information interfaces provide access to information that is relevant in the current context. They use sensory signals, such as gaze patterns, to track the user's context and foci of interest, and to predict what kind of information the user would need at the present time. The information is retrieved from databases and presented in non-intrusive manner. Main challenges are extraction of context from visual and sensory data, construction of adaptive machine learning models that are able to utilize heterogeneous context cues to predict relevance, and undisturbing and easily understandable presentation of information. Novel statistical machine learning methods are used for multimodal information retrieval and for taking the context into account.

As a part of Urban Contextual Information Interfaces with Multimodal Augmented Reality (UI-ART) project, an interdisciplinary research project funded by TKK MIDE (Multidisciplinary Institute of Digitalisation and Energy) programme, we have build a pilot system that retrieves and displays abstract information about people and real world objects in augmented reality [12]. As a pilot application scenario, we have implemented a guide that displays relevant information to a visitor in a university department. The interface consists of either a head-worn display with an integrated gaze-tracker or a hand-held PC that can be pointed towards an interesting object. People and objects in the view are recognized from the video feed and information related to them is searched from a database. Retrieved textual annotations are augmented to the view and become part of the context the user can attend to. Evidence from gaze measurements and speech recognition is integrated to infer the user's current interests and annotations that match those are displayed. Figure 9.2 shows a snapshot of the UI-ART system's augmented reality display.

We studied one component of the pilot system, namely prediction of relevance from gaze patterns, in more detail in [13]. We trained a model to predict importance of objects in the scenes of a video, as reported by test subjects, based on gaze patterns recorded

Figure 9.2: The augmented reality of the UI-ART system in the Virtual Laboratory Guide pilot application.

while the subjects were watching the video. In this feasibility study we observed that gaze patterns provide useful information in inferring user interest.

If available, behavior of other people on the same or similar task can be an effective contextual cue. In [14] we introduced a collaborative filtering method that learns a latent structure both for users and documents. With this two-way generalization the model is able to make predictions when either new users or new documents are added to the dataset, unlike earlier state-of-the-art methods.

The Proactive Interfaces research group participates in the Device and Interoperability Ecosystem (DIEM) research programme of the TIVIT ICT SHOK. The project started in July 2008 and targets to enable new services and applications that are based on smart environments that comprise of digital devices containing relevant information for different purposes. The key is interoperability between devices from different domains. Our group is involved in the Mobile Mixed Reality (DIEM/MMR) work package together with the TKK Department of Media Technology, Nokia Research Center (NRC) and Technical Research Centre of Finland (VTT), among others.

The Image Based Linking project began in 2009. The project aims to provide new ways to get access to digital services for mobile phones with integrated digital cameras. This kind of methods can be used for various purposes linking digital information to the physical world. Possible application areas include outdoor advertising, magazine and newspaper advertising, tourist applications, and shopping. In the context of the Proactive Interfaces project, the researched technologies enable more sophisticated object and location recognition for the developed augmented reality applications.

References

[1] Jaana Simola, Jarkko Salojärvi, and Ilpo Kojo. Using hidden markov models to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9:237–251, 2008.

[2] Kai Puolamäki, Antti Ajanki, and Samuel Kaski. Learning to learn implicit queries from gaze patterns. In Andrew McCallum and Sam Roweis, editors, *Proceedings of ICML 2008, Twenty-Fifth International Conference on Machine Learning*, pages 760–767, Madison, 2008.

[3] Antti Ajanki, David R. Hardoon, Samuel Kaski, Kai Puolamäki, and John Shawe-Taylor. Can eyes reveal interest?—Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 19:307–339, 2009.

[4] Arto Klami, Craig Saunders, Teófilo de Campos, and Samuel Kaski. *Can relevance of images be inferred from eye movements?*, pages 134–140. ACM, New York, 2008.

[5] László Kozma, Arto Klami, and Samuel Kaski. GaZIR: Gaze-based zooming interface for image retrieval. In *Proc. ICMI-MLMI 2009, The Eleventh International Conference on Multimodal Interfaces and The Sixth Workshop on Machine Learning for Multimodal Interaction*, pages 305–312, New York, NY, USA, 2009. ACM.

[6] He Zhang, Markus Koskela, and Jorma Laaksonen. Report on forms of enriched relevance feedback. Technical Report TKK-ICS-R10, Helsinki University of Technology, Department of Information and Computer Science, Espoo, Finland, November 2008.

[7] Ville Viitaniemi and Jorma Laaksonen. Evaluation of pointer click relevance feedback in PicSOM. Technical Report TKK-ICS-R11, Helsinki University of Technology, Department of Information and Computer Science, Espoo, Finland, November 2008.

[8] Markus Koskela and Jorma Laaksonen. Specification of information interfaces in PinView. Technical Report TKK-ICS-R12, Helsinki University of Technology, Department of Information and Computer Science, Espoo, Finland, November 2008.

[9] Jorma Laaksonen. Definition of enriched relevance feedback in PicSOM. Technical Report TKK-ICS-R13, Helsinki University of Technology, Department of Information and Computer Science, Espoo, Finland, November 2008.

[10] Mats Sjöberg and Jorma Laaksonen. Optimal combination of SOM search in best-matching units and map neighborhood. In *Proceedings of 7th International Workshop on Self-Organizing Maps (WSOM 2009)*, volume 5629 of *Lecture Notes in Computer Science*, pages 281–289, St. Augustine, Florida, USA, 2009. Springer. Available online at: `http://dx.doi.org/10.1007/978-3-642-02397-2\_32`.

[11] Ville Viitaniemi and Jorma Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of ACM International Conference on Image and Video Retrieval (CIVR 2009)*, Fira, Greece, July 2009.

[12] Antti Ajanki, Mark Billinghurst, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, and Timo Tossavainen. Ubiquitous contextual information access with proactive retrieval and augmentation. Technical Report TKK-ICS-R27, Helsinki University of Technology, Department of Information and Computer Science, Espoo, Finland, December 2009.

[13] Melih Kandemir, Veli-Matti Saarinen, and Samuel Kaski. Inferring object relevance from gaze in dynamic scenes. In *Proc. ETRA 2010, Eye Tracking Research & Applications*, to appear.

[14] Eerika Savia, Kai Puolamäki, and Samuel Kaski. Latent grouping models for user preference prediction. *Machine Learning*, 74:75–109, 2009. Published online: 3 September 2008.

Chapter 10

Natural language processing

Krista Lagus, Mikko Kurimo, Timo Honkela, Sami Virpioja, Oskar Kohonen, Mari-Sanna Paukkeri, Mikaela Klami, Ville Turunen, Matti Varjokallio, Matti Pöllä, Ilari Nieminen, Tommi Vatanen

10.1 Introduction

Work in the field of natural language processing involves several research themes that have close connections to work carried out in other groups, especially speech recognition (Chapter 8) and Computational Cognitive Systems groups (Chapter 11). The objective of this research is to develop methods for learning general-purpose representations from text that can be applied to the recognition, understanding and generation of natural language. The results are evaluated in applications such as automatic speech recognition, information retrieval, and statistical machine translation.

During 2008–2009, our research has concentrated on finding suitable units of representations, such as morphemes, constructions, and keyphrases, in an unsupervised and language-independent manner. In addition, we have organized Morpho Challenges, international competitions funded by EU's PASCAL network, where multiple evaluations have been provided for algorithms for unsupervised morpheme analysis.

10.2 Unsupervised learning of morphology

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language, and they can be defined in a language-independent manner. It seems that even approximative automated morphological analysis is beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. These applications usually make use of *words* as vocabulary units. However, for highly-inflecting and agglutinative languages, this leads to very sparse data, as the number of possible word forms is very high.

Figure 10.1 shows the very different rates at which the vocabulary grows in various text corpora of the same size. The number of different unique word forms in the Finnish corpus is considerably higher than in the English ones, for example. In addition to the language, the size of the vocabulary is affected by the text type.

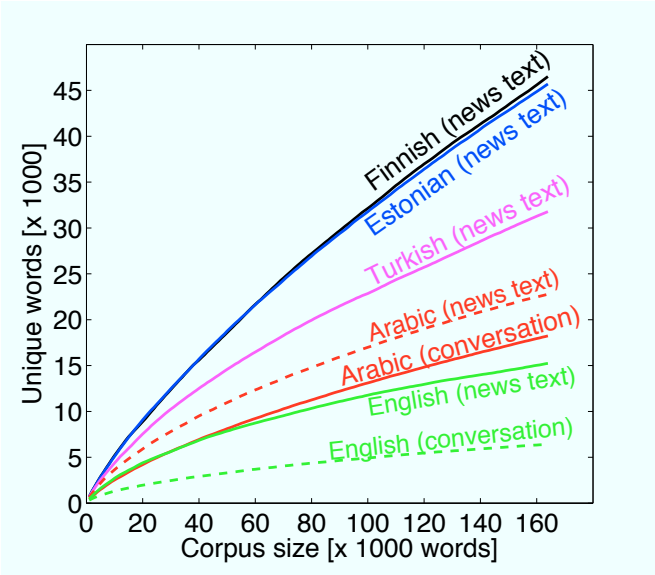


Figure 10.1: The number of different word forms (types) encountered in growing portions of running text (tokens) of various languages and text types.

We have developed language-independent, data-driven methods for the unsupervised discovery of morphemes. *Morfessor* [1] is a family of methods that perform segmentation of words into morpheme-like units. The different versions of Morfessor can be seen as instances of a general model. The model is strongly inspired by the Minimum Description Length (MDL) principle, although the later versions have been expressed in Maximum A Posteriori (MAP) estimation framework [2].

In *Allomorfessor*, Morfessor has been extended to account for the linguistic phenomenon of allomorphy. In allomorphy, an underlying morpheme-level unit has two or more surface realizations (e.g., "day" has an alternative surface form "dai" in "daily"). Recognizing the morpheme-level units should help with applications such as information retrieval and machine translation. In [3], the initial algorithm was evaluated in Morpho Challenge 2008 for English, Finnish, German, and Turkish languages, with moderate but promising results. An improved model performed significantly better in linguistic evaluation [4]. In Morpho Challenge 2009, the new Allomorfessor version performed very well in all languages and tasks, although the amount of allomorphs found by the algorithm was still limited [5].

References

[1] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[2] Krista Lagus, Mathias Creutz, Sami Virpioja, and Oskar Kohonen. Morpheme segmentation by optimizing two-part MDL codes. In *2009 Workshop on Information Theoretic Methods in Science and Engineering (WITMSE)*, Tampere, Finland, August 2009. Extended abstract.

[3] Oskar Kohonen, Sami Virpioja, and Mikaela Klami. Allomorfessor: Towards unsupervised morpheme analysis. In *Working Notes of the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[4] Oskar Kohonen, Sami Virpioja, and Mikaela Klami. Allomorfessor: Towards unsupervised morpheme analysis. In *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008 Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of Lecture Notes in Computer Science, pages 975-982. Springer, 2009.

[5] Sami Virpioja and Oskar Kohonen. Unsupervised morpheme analysis with Allomorfessor. In *Working Notes of the CLEF 2009 Workshop*, Corfu, Greece, 2009.

10.3 Unsupervised discovery of constructions

Construction grammar, originally developed by Charles Fillmore, is a grammatical theory that is beneficial for Natural Language Processing because it provides tools for modeling properties of language that traditional theory ignores (for an overview, see [1]). In particular, the different statistical properties of collocations, multi-word units and idioms are well known in Natural Language Processing.

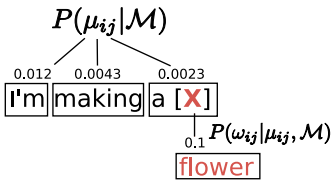


Figure 10.2: An illustration of the construction model used in [2].

In [2] we extended previous work in morphology learning into a method for learning multi-word constructions, as illustrated in figure 10.2. Since the construction grammar framework is a general one, in [3] we developed a framework for construction learning problems that includes both learning syntax and morphology.

References

[1] Adele E. Goldberg. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224, 2003.

[2] Krista Lagus, Oskar Kohonen, and Sami Virpioja. Towards unsupervised learning of constructions from text. In Magnus Sahlgren and Ola Knutsson, editors, *Proceedings of the Workshop on Extracting and Using Constructions in NLP of 17th Nordic Conference on Computational Linguistics, NODALIDA*, May 2009. SICS Technical Report T2009:10.

[3] Oskar Kohonen, Sami Virpioja, and Krista Lagus. A constructionist approach to grammar inference. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, Whistler, Canada, December 2009. Extended abstract.

10.4 Keyphrase extraction

A language-independent keyphrase extraction method, *Likey*, was developed as a follow-on to the earlier language-independent studies. The method utilises statistical analysis of language and comparison to a reference corpus, and it has a light-weight preprocessing phase. Most of the traditional methods for keyphrase extraction are highly dependent on the language used and the need for preprocessing is extensive. On the contrary, *Likey* enables independence from the language being analysed. It is possible to extract keyphrases from text in previously unknown language provided that a suitable reference corpus is available. *Likey* was tested with 11 European languages, including Germanic and Romance languages, Greek and Finnish. The evaluation method was based on Wikipedia articles and their intra-linking. The results were comparable to *tf.idf*, a statistical term weighting method. [1] The keyphrases produced by *Likey* were utilised as features in a web-based interface for collecting and analysis of information on authors and their publications [2].

A web-based demonstration of *Likey* is available at http://cog.hut.fi/likeydemo/. The system highlights keyphrases of a web document written in any of the eleven European languages. Keyphrases extracted from an article in a French online newspaper Le Monde are visualized by the demo in Figure 10.3. For example, the American swimmer "Michael Phelps" and word pair "médailles d'or" (English: *gold medals*) are extracted as keyphrases.

Figure 10.3: Keyphrases extracted by *Likey* from a French online news article.

References

[1] Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela. A language-independent approach to keyphrase extraction and evaluation. In *Coling 2008: Companion volume: Posters*, pages 83–86, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[2] Tommi Vatanen, Mari-Sanna Paukkeri, Ilari T. Nieminen, and Timo Honkela. Analyzing authors and articles using keyword extraction, self-organizing map and graph algorithms. In *Proceedings of the AKRR08*, pages 105-111, 2008.

10.5 Morpho Challenge

Morpho Challenge is a series of scientific competition annually organized by Adaptive Informatics Research Centre for the evaluation of new unsupervised morpheme analysis algorithms. The challenge is part of the EU Network of Excellence PASCAL Challenge Program and in 2008 and 2009 organized in collaboration with Cross-Language Evaluation Forum CLEF.

The objective of the challenge is to design statistical machine learning algorithms that discover which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, information retrieval, and statistical language modeling. The challenge has sofar been organized four times and the results have been published in PASCAL and CLEF workshops in Venice 2006 [1], in Budapest 2007 [2], in Aarhus 2008 [3, 4], and in Corfu 2009 [5].

In the 2009 Morpho Challenge, the evaluation of the submissions have performed by three complementary ways: *Competition 1*: The proposed morpheme analyses were compared to a linguistic morpheme analysis gold standard by matching the morpheme sharing word pairs [5]. *Competition 2*: Information retrieval (IR) experiments were performed, where the words in the documents and queries were replaced by their proposed morpheme representations and the search was based on morphemes instead of words. The IR evaluations were provided for Finnish, German, and English and participants were encouraged to apply their algorithm to all of them. The organizers performed the IR experiments using the queries, texts, and relevance judgments available in CLEF forum and morpheme analysis methods submitted by the challenge participants. The results show that the morpheme analysis has a significant effect in IR performance in all languages, and that the performance of the best unsupervised methods can be superior to the supervised reference methods. *Competition 3*: Statistical machine translation (SMT) experiments were performed, where the words in the source language sentences were replaced by their proposed morpheme representations and the alignment and translation was based on morphemes instead of words [5]. To make te results relevant to the state-of-the-art in SMT, the N-best translation hypotheses of the morpheme-based system were further combined with a conventional word-based system. The word-based system was trained with the same data, but keeping the words unsplit, and the combination was performed by using the minimum Bayes risk combination as in [6]. The experimented language-pairs were Finnish-English and German-English and the results showed that the best unsupervised methods improve the baseline word-based system.

References

[1] Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy and Murat Saraclar. Unsupervised segmentation of words into morphemes - Challenge 2005, An Introduction and Evaluation Report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*. Venice, Italy, April 12, 2006.

[2] Mikko Kurimo, Mathias Creutz, and Matti Varjokallio. Morpho Challenge evaluation using a linguistic Gold Standard. In *Advances in Multilingual and MultiModal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, Lecture Notes in Computer Science, Vol. 5152, pages 864-873. Springer, 2008.

[3] Mikko Kurimo, Ville Turunen, and Matti Varjokallio. Overview of Morpho Challenge 2008. In *Advances in Multilingual and MultiModal Information Retrieval, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Lecture Notes in Computer Science. Springer, 2009.

[4] Mikko Kurimo, Mathias Creutz, and Ville Turunen. Morpho Challenge evaluation by information retrieval experiments. In *Advances in Multilingual and MultiModal Information Retrieval, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Lecture Notes in Computer Science. Springer, 2009.

[5] Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. Overview and results of Morpho Challenge 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September 2009.

[6] Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73-76, Boulder, USA, June 2009. Association for Computational Linguistics.

Computational Cognitive Systems

Chapter 11

Cognitive Systems Research

Timo Honkela, Krista Lagus, Oskar Kohonen, Tiina Lindh-Knuutila, Mari-Sanna Paukkeri, Matti Pöllä, Juha Raitio, Sami Virpioja, Jaakko J. Väyrynen

11.1 Introduction

Computational Cognitive Systems group conducts research on artificial systems that combine *perception, action, reasoning, learning and communication*. This area of research draws upon biological, cognitive and social system approaches to understanding cognition. Cognitive systems research is *multidisciplinary and interdisciplinary*. It benefits from sharing and leveraging expertise and resources between disciplines. Methodologically, *statistical machine learning*, *pattern recognition* and *signal processing* are central tools within computational cognitive systems research. Our research focuses on modeling and applying methods of unsupervised and semisupervised learning for *conceptual modeling*, *machine translation and multilingual processing*, and *socio-cognitive modeling*. The general aim is to provide a methodological framework for theories of *conceptual development*, *symbol grounding and embodiment*, *communication among autonomous agents*, *situational activity analysis*, *social simulation*, and *constructive and expansive learning*.

11.2 Summary of collaboration

We have worked in close collaboration with other groups in Adaptive Informatics Research Centre, lead by Prof. Erkki Oja and Prof. Samuel Kaski, in particular natural language processing and multimodal interfaces (Dr. Mikko Kurimo and Dr. Jorma Laaksonen). We have also collaborated with the representatives of *Helsinki School of Economics* and *University of Art and Design Helsinki*. The collaboration with Helsinki School of Economics and *National Consumer Research Centre* has mainly taken place within Tekes-funded Kulta project that focuses on *modeling and simulation of changing consumer needs*. The project will be described in more detail in the subsequent sections. After 2009, these universities and Helsinki University of Technology, operate within the merged Aalto University. Helsinki University of Technology will operate as *Aalto University School of Science and Technology*. In Tekes-funded ContentFactory project, we have collaborated with *University of Helsinki* and in particular with Dr. Roman Yangarber and Prof. Lauri Carlson and their groups. Our specific topic related to multilingual terminology and ontology learning is described in some detail elsewhere in this report.

An important collaboration arena in the future will be the *EIT ICT Labs* that is built upon five co-location centres in Berlin, Eindhoven, Helsinki, Paris, and Stockholm. *European Institute of Innovation and Technology* (EIT) has nominated the EIT ICT Labs as one of its three first Knowledge and Innovation Communities (KIC). In the preparation of the EIT ICT Labs, Prof. Martti Mäntylä and Prof. Heikki Saikkonen from Helsinki University of Technology have had a central role. We foresee that this institution will be an important platform for research collaboration within our research area. With one of the partners, *Deutsche Forschungszentrum für Künstliche Intelligenz* (DFKI), we already share common interest in promoting efficient European communication through the joint EU-funded *Network of Excellence* in *Technologies for a Multilingual Europe* (T4ME). DFKI will serve as the coordinator of the NoE, lead by Prof. Hans Uszkoreit. The network will start its work in early 2010. The core consortium consists of twelve partners from Germany, France, Spain, Italy, Greece, Czech Republic, Finland, Ireland, the Netherlands and Slovenia, and is coordinated by DFKI, Germany. TKK (Aalto University) is the only Nordic partner in the consortium.

The group has been active in conducting and developing further international collaboration. In September 2008, the group was centrally responsible for organizing the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, AKRR'08 [1]. Timo Honkela has given keynote talks "From quantification

of information to quantification of meaning using socio-cognitive computing" at the 2008 IAPR Workshop on Cognitive Information Processing in Santorini, Greece, and "Conceptual Autonomy of Agents" at the International Conference on Agents and Artificial Intelligence, ICAART 2009 in Porto, Portugal. As a member of European Neural Network Society exectutive committee, Timo Honkela will serve as the programme co-chair of International Conference on Artificial Neural Networks 2011 (ICANN'11) and as the general chair of Workshop on Self-Organizing Maps 2011 (WSOM'11).

Year 2008 was the final year for the multi-national Project *MedIEQ* in which our group was actively involved. The project was co-funded by the European Commission under the Public Health programme. The project was set to pave the way towards the automation of quality labeling process of medical web sites. Matti Pöllä was our key representative in the project. Close collaboration with Prof. Eero Hyvönen's *Semantic Computing Research Group* at TKK was also conducted.

In 2009, two researchers and graduate students in our group, Tiina Lindh-Knuutila and Mari-Sanna Paukkeri, conducted a research visit abroad that lasted half a year. Tiina Lindh-Knuutila visited the International Computer Science Institute (ICS) at *University of California Berkeley*, USA. Mari-Sanna Paukkeri visited the School of Informatics at the *University of Edinburgh*, United Kingdom

The group has collaborated with Academician Teuvo Kohonen, for instance, to study the performance of the *Self-Organizing Map* (SOM) algorithm in vector quantization [2] Moreover, we have helped in updating the large SOM bibliography [3].

11.3 Summary of cognitive systems research areas

Our main research areas are *conceptual modeling*, *machine translation and multilingual processing*, and *socio-cognitive modeling*. We approach conceptual modeling as a dynamic phenomenon. Among humans, conceptual processing takes place as an individual and social process. We attempt to model this dynamic and constructive aspect of conceptual modeling by using statistical machine learning methods. We also wish to respect the overall complexity of the theme, for instance, not relying on explicit symbolic representations are the only means relevant in conceptual modeling. Our machine translation research builds on the conceptual modeling research as well as on the research on adaptive language technology.

Socio-cognitive modeling is our newest research area which builds on 1) the experience and expertise in modeling complex phenomena related to language learning and use at cognitive and social levels and 2) strong national and international collaboration especially with the representatives of social sciences and humanities. Socio-cognitive modeling mainly merges aspects of computer science, social sciences and cognitive science. The basic idea is to model interlinked social and cognitive phenomena.

References

[1] Timo Honkela, Mari-Sanna Paukkeri, Matti Pöllä, and Olli Simula (eds.). Proceedings of AKRR'08, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning. Helsinki University of Technology, Espoo, 2008.

[2] Teuvo Kohonen, Ilari T. Nieminen, and Timo Honkela. On the quantization error in SOM vs. VQ: A critical and systematic study. In Proceedings of WSOM'09, pages 133-144. Springer, 2009.

[3] Matti Pöllä, Timo Honkela, and Teuvo Kohonen. Bibliography of self-organizing map (SOM) papers: 2002-2005 addendum. Technical Report TKK-ICS-R23, Helsinki University of Technology, 2009.

Chapter 12

Conceptual modeling and learning

Krista Lagus, Timo Honkela, Tiina Lindh-Knuutila, Mari-Sanna Paukkeri, Juha Raitio, Oskar Kohonen, Paul Wagner

12.1 Introduction

Conceptual modeling is a task which has traditionally been conducted manually. In artificial intelligence, knowledge engineers have written descriptions of various domains using formalisms based on predicate logic and other symbolic representations such as semantic networks and rule-based systems. The development of expert systems in 1980s was a notable example of such efforts. As a modern related attempt, the Semantic Web can be mentioned. It seems that the complexity and changing nature of most of the domains makes such formalisms problematic in many real-world applications.

A problem often neglected in symbolic knowledge representation tradition is subjectivity. For us, it seems evident that major portions of individual conceptual systems are learned. Due to the individual and cultural differences, it is not believable that concepts could be modeled with static structures without making use of adaptive processes.

Another challenging topic related to conceptual modeling is contextuality. Contextuality is illustrated in Fig. 12.1. Human activity takes normally place in rich contexts in which the relationship between prototypical meanings of expressions and the situation may be complex. Phenomena like subjectivity and contextuality serve as motivation for the research that is described in the following.

Figure 12.1: An illustration of contextual effects in the interpretation of linguistic expressions. There is a prototypical red but the redness of a shirt typically differs considerably from the redness of skin or wine. In an image, white snow may not altogether be very white.

The theories of knowledge have traditionally been based on predicate logic and related methodologies and frameworks. The basic ontological assumption is that the world consists of objects, events and relationships. The language and the conceptual structures are then supposed to reflect rather straightforwardly this structure. Learning has been seen as a means to memorize the mapping from the epistemological domain (to put it simply: words) into the ontological domain (objects, events and relationships). This view has been dominant at least partly because of the consistent formalization of the theory through the use of symbolic logic. Moreover, the use of the von Neumann computer as the model or metaphor of human learning and memory has had similar effects and has strengthened the idea of the memory as a storage of separate compartments which are accessed and processed separately and which are used in storing and retrieving information more or less as such. [4]

Realistic simulations of the socio-economical and cultural levels are seemingly difficult to build due to the complexity of the overall system. The richness of human culture makes it difficult as a phenomenon to model. Moreover, already the world knowledge of a single human being is so vast that it is difficult to approach it successfully. However, useful development may be possible by taking into account the aspects presented, e.g., in [9, 10, 1, 8, 3]. For instance, Vygotsky [10] has stated that "... the world of experience must

be greatly simplified and generalized before it can be translated into symbols. Only in this way does communication become possible, for the individual's experience resides only in his own consciousness and is, strictly speaking, not communicable." Later, he continues: "The relation of thought to word is not a thing but a process, a continual movement back and forth from thought to word and from word to thought. In that process the relation of thought to word undergoes changes which themselves may be regarded as development in the functional sense." This means in practice that conceptualization is a complex process that takes place in a socio-cultural context, i.e., within a community of interacting individuals whose activities result into various kinds of cultural artifacts such as written texts.

The basic aim in our research group is to provide the means for a more or less automatic process of concept formation. This will facilitate both cost-effective development of knowledge-intensive systems as well as serve as a good basis for systems that can update themselves taking into account changes in the domain of interest.

Next we present three specific research areas within conceptual modeling with recent results. The *intersubjective communication model* aims at providing a general framework for explaining how communication between human or artificial agents that have different conceptual models can be successful and what kind of problems there also may be. We have also developed a *multiagent simulation model of conceptual development*. This model combines probabilistic modeling of concept naming with the self-organization of the underlying conceptual space in an agent population. In the third study, we have conducted an *analysis of philosophy students' conceptions*.

12.2 Intersubjective communication model

We have recently proposed a theoretical framework for modeling communication between two agents that have different conceptual models of their current context [5]. We have described how the emergence of subjective models of the world can be simulated and what the role of language and communication in that process is. We have considered the role of unsupervised learning in the formation of agents' conceptual models, the relative subjectivity of these models, and the communication and learning processes that lead into intersubjective sharing of concepts [5].

In this section, we introduce the basic definitions and notation used in our communication model for two agents. The key concept is the agent's internal view of its context, the *concept space*. The concept space is spanned by a number of features. We can use the terminology coined by Gärdenfors [2] calling each feature (f_i) a quality dimension. Dimensionalities of the concept spaces can be different for each agent . The concept space of agent 1 is N-dimensional metric space C^1, and for agent 2, C^2.

This work has several theoretical and practical implications including the possibility of approaching interoperability of information systems from a novel point of view. Some of these implications are discussed next (see the original article [5] for additional details and references).

The traditional notion of uncertainty in decision making does not cover the uncertainties caused by differences in conceptual systems of individual agents within a community. We claim that in all transactions including symbolic/linguistic communication the differences in the underlying conceptual systems play an important role. For instance, serious efforts have been made to harmonize or to standardize the classification systems used by business agents, e.g., using Semantic Web technologies. However, even if the standardization is conducted, there can not be any true guarantee that all the participating agents would share the meaning of all the expressions used in the business transactions in various contexts.

One implication is that in business transactions there should be means for checking what is a meant by some expressions by an access to a broader context (cf. symbol grounding). Moreover, rather than relying solely on a standardized conceptual system, one could introduce mechanisms of meaning negotiation. Before two business agents get into negotiation about, for instance, the price of some commodity, they should first check if they agree on what they refer to by the expressions that are used in the negotiation. This concern is valid both for human and computerized systems, even though humans are usually capable to conduct meaning negotiations even when they are not aware of it [5].

The harmonization of conceptual systems, such as the creation of ontologies for business transactions, has obvious benefits when, for instance, the interoperability of related information systems is considered. It appears ideal that all systems within some domain would use similar terminologies and shared ontologies. However, this approach can be claimed to be idealistic because the continuous change through innovations and other activities and the underlying learning processes within the human community lead into the situation carefully considered in the earlier chapters of this paper. All the agents have a conceptual system of their own, at least to some degree. Therefore, the harmonization of the conceptual systems should be considered only as a relative goal. One may aim for a larger degree of sharing of the conceptual system as before. A central theme is then to assess the associated benefits and costs. Here we do not try to provide any means to estimate the benefit of well working harmonized conceptual system implemented, e.g., as an ontology within the Semantic Web framework [5].

The costs stem from two main sources: the development of a shared conceptual systems

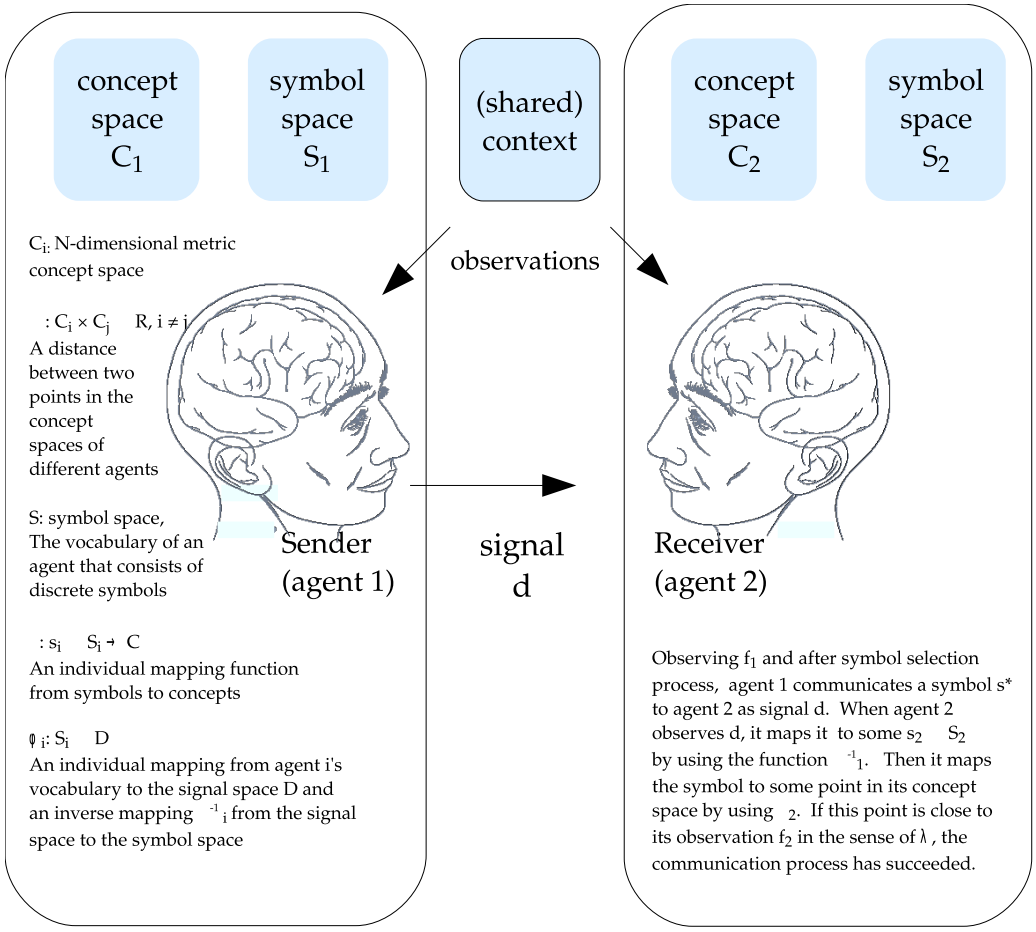


Figure 12.2: In the intersubjective communication framework, there are two distance measures ω and λ. ω gives a distance between two points inside the concept space of the agent, i.e. $\omega : C^i \times C^i \rightarrow \mathbb{R}, i = 1, 2,$. λ gives a distance between two points in the concept spaces of the different agents, i.e. $\lambda : C^i \times C^j \rightarrow \mathbb{R}, i \neq j$. The symbol space S^1 of the agent 1 is its vocabulary that consists of discrete symbols. Similarly, the vocabulary of agent 2 consists of symbols S^2. An agent i has an individual mapping function ξ^i that maps the symbol $s^i \in S^i$ to C^i. An agent i expresses each symbol $s^i \in S^i$ as a signal d in the signal space D. The signal space D is multidimensional, continuous and shared between the agents. Each agent i has an individual mapping function ϕ^i from its vocabulary to the signal space, i.e. $\phi^i : S^i \rightarrow D$ and an inverse mapping ϕ^{-i} from the signal space to the symbol space.

and the use of it. The development of an ontology typically consists of defining the concepts and the relationships between the concepts. The typical stages of an ontology building process are the following: (1) domain analysis resulting into the requirements specification, (2) conceptualization resulting into the conceptual model, (3) implementation that leads into the specification of the conceptual model in the selected representation language, and (4) the ontology population i.e. the generation of instances and their alignment to the model that results into the instantiated ontology [5].

The estimation of costs related to the use of ontologies is rather difficult. There are many kinds of uses of ontologies that require higher or lower degree of familiarity of

conceptual structures of the domain. A widely cited claims from expertise research is the 10-year rule, first proposed in relation to expertise development among chess players, and later generalized to other domains. The essential content of the rule is that anyone seeking to perform at world-class level in any significant domain must engage in sustained, deliberate practice in the activity for a period of at least ten years. This figure serves only as a upper bound of an estimate for a person to learn to master the conceptual content of a complex domain [5].

12.3 Multiagent simulation model of conceptual development

In [6], we present a model that combines probabilistic modeling of concept naming with the self-organization of the underlying conceptual space in an agent population. In this multi-agent simulation framework, we study emergence of a common vocabulary. The self-organizing map is used for the purpose of transferring sensory perceps into a conceptual level representations.

In the community of agents, we assume that each agent has its own representation. While the representations are alike due to similar training data, the representations are not exactly the same. On top of the concept emergence, we studied shared vocabulary emergence using a naming game paradigm , in which two agents share a common perceived context, and they attempt to find a name to match their observation. Each agent matches the observation to their concept map by finding the best-matching unit for that data point in the self-organizing map. For that given map unit, each agent then selects the term to denote that observation based on the maximum likelihood, $\max(P(C|T))$, which is estimated as the number of successful uses of the term for a given map node, proportional to all of the successful uses of all the terms in that node. The likelihood is estimated for all the terms associated with the BMU and for those nodes adjacent to it, and the term with the highest likelihood is selected and uttered. If no term is found to be associated with the color or its neighborhood in the self-organizing map, a new term is invented. The hearer estimates the likelihood $P(C|T)$ in similar fashion. When a number of games is played, a common vocabulary emerges in the population.

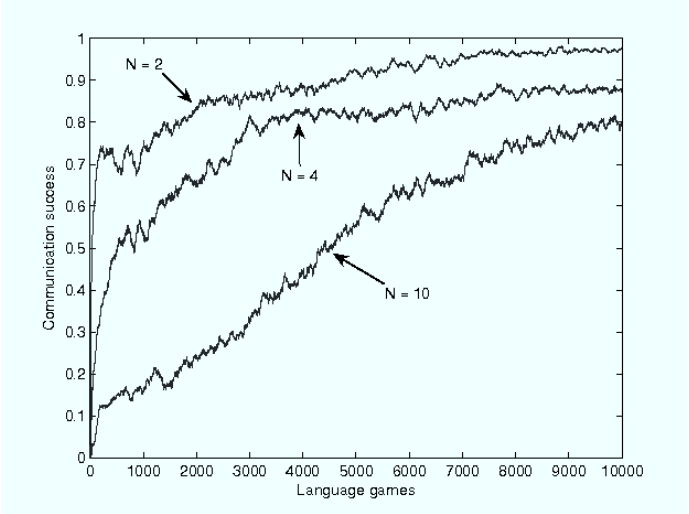


Figure 12.3: Communication success for $N = 2$, $N = 4$ and $N = 10$ agents in the population.

Figure 12.3 shows the communication success for two, four and ten agents, each averaged over 10 simulation runs. In the two-agent case, the communication success, the fraction of successful games of the previous hundred games played, rises rapidly to $CS = 0.8$ and then steadily up to $CS = 0.95$ during the $10,000$ simulated games. The communication success for four agents grows slower than in the previous experiment, but still increases up to $CS = 0.86$, where it seems to settle. The bigger population size, in the ten-agent case yields into considerably slower convergence, reaching approximately $CS = 0.8$

in 10,000 games.

All the language games are played pair-wise, i.e. only two agents of the whole population participate in each game, and other agents have access to the words only through subsequent language games with the same topic. This means that when the population size grows, the convergence to common vocabulary is considerably slower. More competing words for a given topic emerge, and it simply takes longer for each agent to see a representative subset of the topics.

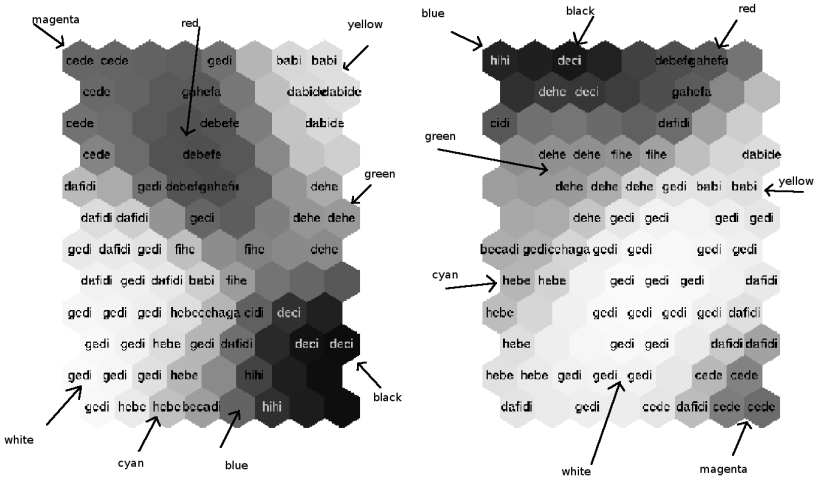


Figure 12.4: The conceptual memories of the agents in the two-agent simulation. Only the most probable label for each node is shown.

Figure 12.4 shows the conceptual maps of the two agents in the first experiment. The colors denote the converged RGB values of the prototype vectors of the map. The map has organized well and transformations from one color to other are smooth. The eight prototypical colors used are more prominent, since they are represented more in the data than the intermediate colors that have resulted from added noise.

When comparing the figures, it is evident that for most prototypical colors there are one or two words that are preferred: *deci* is preferred for black or dark, *hihi* for blue, *fehe* for green, *hebe* for cyan, *defebe* and *gahefa* for red, *cede* for magenta, and *babi* and *dabide* for yellow. For white, the most common word used is *gedi*, but there are also competing labels for bluish white, pinkish white and so on because white covers a larger area in the space. The conceptual memories support the conclusion already visible in the communication success ratio – that a common vocabulary for the agents has emerged.

12.4 Analysis of philosophy students' conceptions

In collaboration with researchers from Helsinki University, Anna-Mari Rusanen, Otto Lappi and Mikael Nederström, we have used the self-organizing map algorithm to analyze and visualize the initial conceptions of philosophy students [7].

The general theoretical approach of this study was based on the conceptual change paradigm. There is a large body of research which shows that novices conceptions do differ from those of experts, but researchers still remain divided not only about the nature of those differences, and also the status of novices' belief systems. Some researchers claim that novices belief systems are weakly organized systems that are internally inconsistent, piecemeal and incoherent. Other researchers argue that novice belief systems are not only internally quite coherent but they may also share the essential properties of scientific theories.[7]

To obtain information on the students' conceptions, we used a multiple choice-questionnaire. The questionnaire included 63 thematically selected items. Three thematic sets of questions probed (1) the subjects' ontological commitments with regard to the mind and the body, (2) hypothetical questions that relate to the possible spatial and temporal attributes of bodyless minds (3) hypothetical questions that relate to the possible perceptual and cognitive attributes of bodyless minds. Each of these conceptual subdomains was probed with multiple questions, and the students' responses were examined, coded in the binary format and used to train a self-organizing map for visualization.[7]

To summarize the results, the overall structure of the map suggests that the students do not share a clear and coherent set of beliefs on the spatiotemporal attributes of an immaterial mind, whereas in the case of sensory and cognitive capacities they are quite consistent. The question of internal coherence of this recurring set of belief remains an open question, however. The SOM map cannot address this question directly. However, it does show that if the students are incoherent, they are consistently incoherent in the same way.[7]

References

[1] M. Cole and Y. Engeström. *Distributed Cognition*, chapter A Cultural-Historical Approach to Distributed Cognition, pages 1–47. Cambrigde University Press, 1991.

[2] P. Gärdenfors. *Conceptual Spaces*. MIT Press, 2000.

[3] K. Hakkarainen, T. Palonen, S. Paavola, and E. Lehtinen. *Communities of networked expertise: Professional and educational perspectives*. Elsevier, 2004.

[4] Timo Honkela, Teemu Leinonen, Kirsti Lonka, and Antti Raike. Self-organizing maps and constructive learning. In *Proceedings of ICEUT'2000, International Conference on Educational Uses of Communication and Information Technologies*, pages 339–343. IFIP, 2000.

[5] Timo Honkela, Ville Könönen, Tiina Lindh-Knuutila, and Mari-Sanna Paukkeri. Simulating processes of concept formation and communication. *Journal of Economic Methodology*, 15(3):245-259, 2008.

[6] Tiina Lindh-Knuutila, Juha Raitio, and Timo Honkela. Combining self-organized and Bayesian models of concept formation. In J. Mayor, N. Ruh, and K. Plunkett, editors,*Connectionist Models of Behaviour and Cognition II Proceedings of the Eleventh*

Neural Computation and Psychology Workshop, number 18 in Progress in Neural Processing, pages 193-204. World Scientific, April 2009.

[7] Anna-Mari Rusanen, Otto Lappi, Timo Honkela, and Mikael Nederström. Conceptual coherence in philosophy education - visualizing initial conceptions of philosophy students with self-organizing maps. In B. C. Love, K. McRae, and V. M. (Eds.) Sloutsky, editors, Proceedings of the 30th Annual Conference of the Cognitive Science Society, pages 64-70, Austin, Texas, 2008.

[8] J. Lindblom and T. Ziemke. Social situatedness of natural and artificial intelligence: Vygotsky and beyond. *Adaptive Behavior*, 11(2):79–96, 2003.

[9] L. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, Massachusetts, 1978. (Originally published in 1934).

[10] L. Vygotsky. *Thought and Language*. MIT Press, Cambridge, Massachusetts, 1986. (Originally published in 1934).

Chapter 13

Learning to translate

Jaakko J. Väyrynen, Sami Virpioja, Timo Honkela, Mikko Kurimo, Marcus Dobrinkat, Tero Tapiovaara, Tommi Vatanen

13.1 Introduction

Our research on multilinguality and machine translation (MT) uses novel methods that are based on adaptivity. An MT system is *learning to translate* rather than needs to be programmed to do so. The advances in statistical machine translation have shown that the adaptive paradigm can help in reducing the system development costs dramatically. However, these systems rely on representations that do not capture many relevant linguistic aspects, neither take into account the wealth of knowledge that is known about human cognitive processes related to natural language understanding, translation and interpretation.

13.2 Analysis of complexity of European languages

We have studied differences between the European Union languages using statistical and unsupervised methods [8]. The analysis has been conducted at different levels of language including lexical, morphological and syntactic levels. Our premise is that the difficulty of the translation could be perceived as differences or similarities in different levels of language. Two approaches were selected for the analysis. A Kolmogorov complexity based approach was used to compare the language structure in syntactical and morphological levels. A morpheme-level comparison was conducted based on an automated segmentation of the languages into morpheme-like units.

13.3 Learning interlingual mappings

We have also developed an approach for finding interlingual mappings using the Self-Organizing Map (SOM) algorithm [4]. The semantic or conceptual space is explicitly modeled in the SOM-based approach. This can be constrasted with the commonly used Bayesian approach. This approach resembles, to some degree, the idea of using a knowledge-based interlingua in machine translation. The underlying philosophical assumptions about knowledge are, however, quite different. In a knowledge-based interlingua, the semantics of natural language expressions are typically represented as propositions and relations in symbolic hierarchical structures. The SOM can be used to span a continuous and multidimensional conceptual space in a data-driven manner. Moreover, the approach provides a natural means to deal with multimodal data (cf. [9]).

13.4 Applying morphology learning to statistical machine translation

Languages of rich morphology pose a problem for statistical machine translation methods, which usually apply words as the smallest units of translation. We have studied how unsupervised learning of morphology (see Section 10.2) can be used to help in the task. In a joint work with University of Cambridge [3], automatic morphological segmentations by Morfessor [1] were shown to improve the translations provided by the well-known Moses system [6]. The approach combines individual translation models that use alternative morphological decompositions using Minimum Bayes Risk decoding. Statistically significant improvents were obtained for two tasks: Arabic to English task, where two different morphological analyses were applied for Arabic, and Finnish to English, where word-based model was combined with one where Morfessor was applied for Finnish. The method was applied also in the machine translation tasks of Morpho Challenge 2009 (see Section 10.5).

13.5 Experiments in speech-to-speech machine translation

In a join effort with Speech Recognition (Ch. 8) and Natural Language Processing (Ch. 10) groups, we conducted experiments with speech-to-speech machine translation from Finnish to English [2]. The experiment is described in detail in Section 8.4.

13.6 Automatic machine translation evaluation

The feasibility of normalized compression distance as an automatic machine translation evaluation measure has been investigated [10]. The examined distance metric is based on an approximation of the Kolmogorov complexity between translated text and a reference translation. Compared to many state-of-the-art automatic measures, normalized compression distance is theoretically justified while providing competitive correlation to human judgments,.

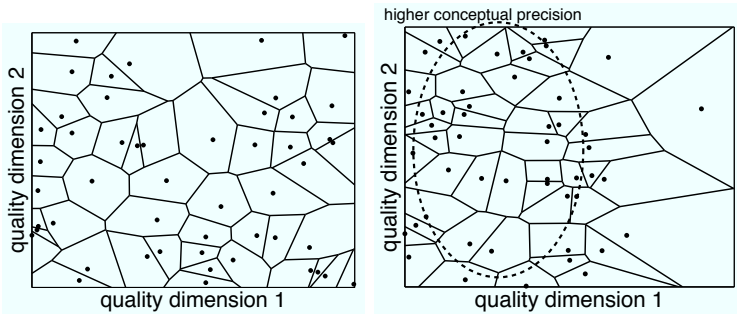


Figure 13.1: Illustration of differing conceptual densities of two agents having a 2-dimensional quality domain. Points mark the locations of the prototypes of concepts. Lines divide the concepts according to Voronoi tessellation. Both agents can discriminate an equal number of concepts, but abilities of the agent B are more focused on the left half of the quality dimension 1, whereas agent A represents the whole space with rather equal precision.

13.7 Within-language translation

The research related to machine translation includes also within-language translation activities. The basic idea is to conduct translation or paraphrasing between two different ways using the same language. In a preparatory study towards this direction, automated classification into layperson and expert use of medical language was conducted using the SVM (support vector machine) method [7].

In general, to provide motivation for this line of research, two persons may often have very different conceptual density related to a topic under consideration. For instance, in Fig.13.1 person A has a rather evenly distributed conceptual division of the space, whereas person B has a more fine-grained conceptual division on the left side of the conceptual space, but has lower precision on the right side of the space [5].

If some agents speak the *same language*, many of the symbols and the associated concepts in their vocabularies are the same. A subjective conceptual space emerges through an individual self-organization process. The input for the agents consists of perceptions of the environment, and expressions communicated by other agents. The subjectivity of the conceptual space of an individual is a matter of degree. The conceptual spaces of two individual agents may be more or less different. The convergence of conceptual spaces stem from two sources: similarities between the individual experiences (as direct perceptions of the environment) and communication situations (mutual communication or exposure to the same linguistic/cultural influences such as upbringing and education, and artifacts such as newspapers, books, etc.) [5]. In a similar manner, the divergence among conceptual spaces of agents is caused by differences in the personal experiences/perceptions and differences in the exposure to linguistic/cultural influences and artifacts. These aspects are handled in more detail in the section on socio-cognitive modeling 14.

References

[1] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[2] David Ellis, Mathias Creutz, Timo Honkela, and Mikko Kurimo. Speech to speech machine translation: Biblical chatter from Finnish to English. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 123-130, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing.

[3] Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73-76, Boulder, USA, June 2009. Association for Computational Linguistics.

[4] Timo Honkela, Sami Virpioja, and Jaakko Väyrynen. Adaptive translation: Finding interlingual mappings using self-organizing maps. In *Vera KurkovÃ¡, Roman Neruda, and Jan Koutnik, editors, Proceedings of ICANN'08*, volume 5163 of *Lecture Notes in Computer Science*, pages 603-612. Springer, 2008.

[5] Timo Honkela, Nina Janasik, Krista Lagus, Tiina Lindh-Knuutila, Mika Pantzar, and Juha Raitio. Modeling communities of experts - conceptual grounding of expertise. Technical Report TKK-ICS-R24, Helsinki University of Technology, 2009.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar, Alexandra Constantin, and Evan Herb. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

[7] Marja Ollikainen. *Matching medical documents to users; Lääketieteellisten dokumenttien sovitus käyttäjille.* Master's Thesis. Helsinki University of Technology, Department of Information and Computer Science, Espoo, 2008.

[8] Markus Sadeniemi, Kimmo Kettunen, Tiina Lindh-Knuutila, and Timo Honkela. Complexity of European Union languages: A comparative approach. *Journal of Quantitative Linguistics*, 15(2):185-211, 2008.

[9] Mats Sjöberg, Jorma Laaksonen, Timo Honkela, and Matti Pöllä. Inferring semantics from textual information in multimedia retrieval. *Neurocomputing*, 71(13–15):2576–2586, 2008.

[10] Jaakko J. Väyrynen, Tero Tapiovaara, Kimmo Kettunen, and Marcus Dobrinkat. Normalized compression distance as an automatic MT evaluation metric. In *Machine Translation 25 Years on*, to appear.

Chapter 14

Socio-cognitive modeling

Timo Honkela, Krista Lagus, Tiina Lindh-Knuutila, Mari-Sanna Paukkeri, Juha Raitio, Eric Malmi

14.1 Introduction

Socio-cognitive modeling is a new research area that merges aspects of computer science, social sciences and cognitive science. The basic idea is to model interlinked social and cognitive phenomena. Our focus has traditionally been in modeling individual cognition that learns and uses language, or in building models of language using statistical machine learning methods. Already for a long time, we have been interested in language and its use as a dynamic phenomenon rather than as a static structural object. Thereafter, we have widened our interest to language as a socio-cultural phenomenon that encodes human knowing and further to other socio-cognitive phenomena, however often related to language. In other words, cognition and intelligent activity are not only individual processes but ones which rely on socio-culturally developed cognitive tools. These include physical and conceptual artifacts as well as socially distributed and shared processes of intelligent activity embedded in complex social and cultural environments [5].

At the socio-cultural level, humans create and share conceptual artifacts such as symbols, words and texts. These are used as mediators between different minds. In communicating and sharing knowledge, individuals have to make a transformation between their internal representation into an explicit representation to be communicated and vice versa, as Vygotsky pointed out already in the 1930s. The internalization and externalization processes take place as a continuous activity. In externalization, the internal view is externalized as explicit and shared representations. Vygotsky also investigated child development and how this was guided by the role of culture and interpersonal communication [19]. He observed how higher mental functions develop historically in cultural groups and individually through social interactions. The specific knowledge gained by children represents the shared knowledge of a culture including the social norms, e.g., related to language use. In our research, we are interested how norms emerge, evolve, and disintegrate at a sociocultural level, how the norms are internalized and externalized by individuals, how they are followed or occasionally deliberately not followed, and how they are implicitly represented in linguistic expressions and explicitly represented as externalized rules.

One approach in socio-cognitive modeling is social simulation. It aims at exploring and understanding of social processes by means of computer simulation. Social simulation methods can be used to to support the objective of building a bridge between the qualitative and descriptive approaches used in the social sciences and the quantitative and formal approaches used in the natural sciences. Collections of agents and their interactions are simulated as complex non-linear systems, which are difficult to study in closed form with classical mathematical equation-based models. Social simulation research builds on the distributed AI and multi-agent system research with a specific interest of linking the two areas. The research area of simulating social phenomena is growing steadily (see, e.g., [18]).

In Kulta project, we have been modeling and simulating the changing needs of consumers in collaboration with Helsinki School of Economics (including Prof. Mika Pantzar, Prof. Raimo Lovio and his group, and Aleksi Neuvonen) National Consumer Research Center (Dr. Tanja Kotro, and Mikael Johnson). The project, funded by Tekes, ends during spring 2010 but many results are already available including [10, 11, 15, 13]. Also some other results have been partially based or connected to the Kulta project including [7]. As an interdisciplinary effort, a wide range of methodologies has been developed, refined and/or applied. These have summarized in Fig. 14.1. In addition to the research partners, the network has included collaborators from various sectors: information and communications technology (Nokia), energy (Helsingin Energia), gaming (Finland's Slot Machine Association), and consultancy (Pöyry).

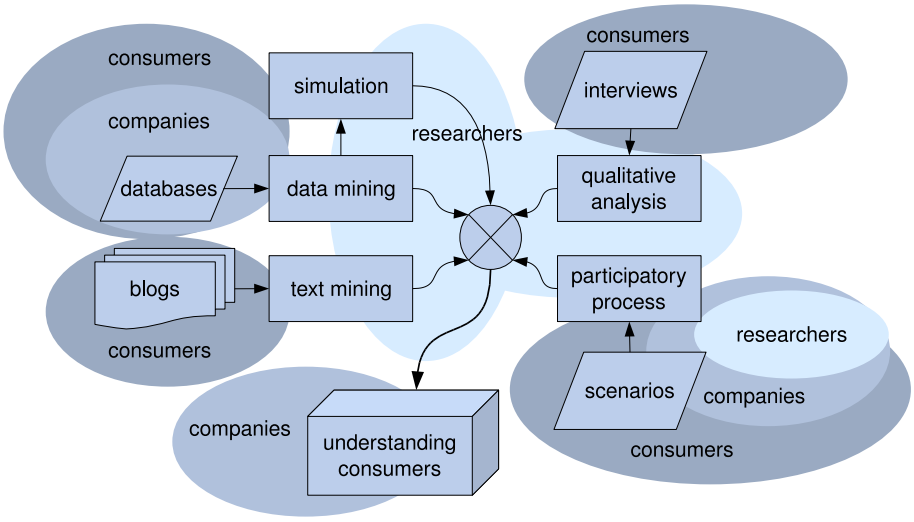


Figure 14.1: A schematic diagrams of the means used in Kulta project for obtaining understanding on changing consumer needs.

A central theoretical starting point in Kulta project has been practice theory as formulated by Prof. Mika Pantzar and Prof. Elisabeth Shove. In their theory, it is assumed that practices consist of three basic elements: material (materials, technologies and tangible, physical entities), image (domain of symbols and meanings), and skill (competence, know-how and techniques) [16, 17]. Practices come into existence, persist and disappear when links between these foundational elements are made, sustained or broken: material, image and skill co-evolve. The disintegration of the links leads into fossilization [16, 17].

In the following, we describe several areas of our research related to socio-cognitive modeling. Modeling expertise is considered both at individual and social level as well as in its implicit and explicit forms. Based on the results of a collaboration effort with Prof. Kai Hakkarainen's group from University of Helsinki, we present how development of knowledge structures in the web can be grounded on evolving knowledge practices and tools supporting them. This area of research is nowadays coined with the term "pragmatic web". The development of a social simulation model based on practice theory is also described. We continue by reporting four results on analyzing complex socio-cognitive phenomena and data. First, we discuss in some detail how text mining based on the self-organizing map can be used to support qualitative research. Second and third, based on the collaboration in Kulta project especially with Dr. Tanja Kotro, we describe means for analyzing consumer data and supporting democratic innovation in organizations by collecting and analyzing observations, ideas and questions. Fourth, we present an analysis of the relationship between the popularity of political parties in parliamentary elections and the socio-economic situation in Finland between 1954 and 2003.

14.2 Modeling expertise at individual and social level

Finding ways in which communities of experts can benefit from each other is a question shared by the machine learning community and social sciences alike. Considerable research in machine learning methods has shown that communities of experts can provide consistently better classifications and decisions than single experts in various tasks and domains.

In our research, we have extended the perspective on communities of experts to cover the wider context of socio-cognitive research. In particular, we consider how the formation and use of expertise relates to the modeling of concept formation, integration and use in human and artificial agents. We have presented a methodological framework for the computational modeling of these phenomena with a specific emphasis on unsupervised statistical machine learning of heterogeneous conceptual spaces in multi-agent systems [7].

We consider different computational models that have been used to represent individual expertise. In particular, we make a distinction between explicit representations (such as rule systems) and implicit representations (such as artificial neural networks [7].

It seems that an individual's rationality is an adaptive tool that does not follow (only) the principles of symbolic logic or probability theory as such, but includes various "cognitive survival strategies", such as a collection of heuristics as pointed out, for instance, by Gerd Gigerenzer and his colleagues [3]. The difference between explicit and implicit knowledge is usually defined by referring to language. If knowledge is represented as interpretable linguistic expressions, it is considered to be explicit, otherwise implicit. Computational intelligence methods such as neural networks and statistical machine learning have provided models of implicit (unconscious, intuitive) understanding [7]. The nature of knowing also depends on the source of experience on the concept or topic (direct versus indirect), illustrated in Fig. 14.2.

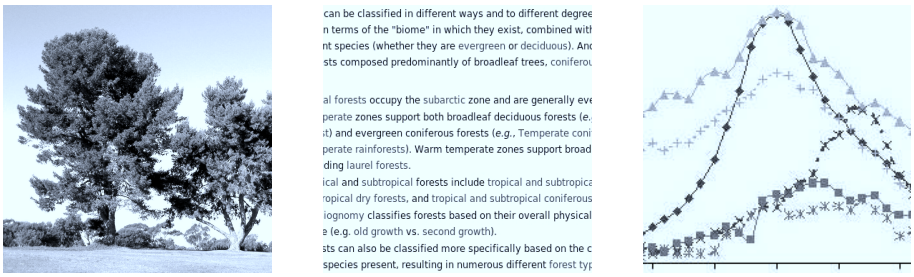


Figure 14.2: An illustration of different sources of knowing i.e., direct experience, written information, and numerical information.

The social level of expertise refers to competencies that arise from social interaction, knowledge sharing, and collective problem solving ([5]. Cognition and intelligent activity rely on socio-culturally developed cognitive tools. These include physical and conceptual artifacts as well as socially distributed and shared processes of intelligent activity embedded in complex social and cultural environments [5]. Expertise at the social level is constituted in interaction between individuals, communities, and larger networks supported by cognitive artifacts.

When a community of conceptually heterogeneous human experts collaborate in order to solve challenging problems, for instance, in the environmental, health or consumer domains, they are likely to encounter a number of knowledge-related challenges [1]. Some of these challenges stem from differences in the conceptual systems of the individual experts. These kinds of situations call for means of highlighting the conceptual differences and

resolving the resulting communication blocks. We present three strategies for this. These three strategies are, in order of increasing complexity, a) clarifying naming conventions, b) visualizing differences in conceptual density, and c) providing augmenting data that mediates between the different conceptual systems [7].

Communication across borders of expertise in collaborative problem solving efforts can, in principle, be achieved in two ways: (1) by bringing forth a combination of the opinions of the experts by, e.g., voting, or (2) by a more involved sharing or integration of expertise and experience at the conceptual level [7]. A particular form of sharing expertise is sharing prototypes. This refers to a process in which an expert communicates prototypical cases to the other expert. In the methodological context of the self-organizing map and other prototype-based conceptual models, prototype sharing means transmitting a collection of model vectors [7].

14.3 Knowledge practices and pragmatic web

We have collaborated with University of Helsinki and the Knowledge-Practices Laboratory project in which 22 organizations from 15 European countries take part. This integrating project (IP), coordinated by Prof. Kai Hakkarainen's research group. The semantic web has been the general foundation of KP-Lab project, but it also addresses practice-based issues extending mere semantic considerations and highlighting the importance of examining the boundaries of the semantic and pragmatic webs. This has been the central area of common interest for Hakkarainen's research group and the computational cognitive systems group [4].

The Pragmatic Web consists of the tools, practices and theories describing why and how people use information. In contrast to the Syntactic Web and Semantic Web the Pragmatic Web is not only about form or meaning of information, but about social interaction which brings about e.g. understanding or commitments.

The transformation of existing information into information relevant to a group of users or an individual user includes the support of how users locate, filter, access, process, synthesize and share information. Social bookmarking is an example of a group tool, end-user programmable agents are examples of individual tools

In the context of the pragmatic web, creating, using, and developing knowledge rather than mere transmission of information or social exchange becomes the central concern. So far, discourses concerning the pragmatic web have, however, mainly addressed contextual aspects of using information by diverse communities of practice. Also highlighted have been various processes of negotiation of meaning which take place in the context of knowledge usage [8].

The pragmatic web may elicit knowledge creation by 1) providing a technological infrastructure for augmenting the functioning of more or less distributed epistemic communities, 2) facilitating automated analysis and interpretation of large bodies of data generated by the users, and 3) adapting to and coevolving with human knowledge practices. Knowledge practice refers to personal and social practices related to working with knowledge. Current theories of social practices highlight both the inseparability of knowing and doing and the creative and improvisational aspect of practice. Here the term "knowledge" is used in the broadest sense, to include explicit official discourses, implicit habits of expert working; and further yet to that which underlies the competencies of experts, for example, so called "procedural knowledge".

An essential factor in pragmatics is context. When the semantic level is dealt with in a context-free manner, investigators tend to focus on prototypical meanings. Resulting models consist of a set of entities and relations connecting those entities. In their actual use at the pragmatic level, meanings are imprecise and changing, biased at any moment by the particular social and external context. The contextual process of meaning attribution is simultaneously both socio-cultural and cognitively subjective [8]. While the semantic web has been preoccupied with standardization of knowledge and systems of knowledge based on ontologies determined mainly by experts beforehand, the envisioned pragmatic web is oriented toward adapting to the special needs of customers and user communities. Rather than simply assimilating to already existing knowledge ontologies, the vision is to engage user communities in active negotiation and interpretation of meaning and to the development of knowledge structures grounded on their evolving practices and epistemic pursuits.

14.4 Social simulation and ensemble models

Agent-based social simulation has provided the scientists a promising tool for analyzing social phenomena without costly real-life experiments. Recently, massive sources of social data have also started to become available. In [13] we explore the possibilities of social simulation in demonstrating the practice theory of social sciences. We present a stochastic multiagent simulation framework for modeling the diffusion of practices among a group of agents. The evolution of the system state in the high-dimensional space of practices is visualized and analyzed using the SOM.

In general, the agents implement stochastic behavior based on their individual representations of beliefs, utilities and context. Ensembles of such agents are simulated for stochastic forecasting of state distributions of the modeled phenomenon. Our goal has been to build a framework that can handle incomplete data and probabilistic interaction models. Based on the framework, we propose a recommender system that combines the traditional collaborative filtering and content-based methods. To evaluate the framework and the recommender system, we applied it to music listening data from the Last.fm service.[13]

In collaboration with Dr. Amaury Lendasse in AIRC and colleagues, we have investigated the application of adaptive ensemble models of Extreme Learning Machines to the problem of one-step ahead prediction in (non)stationary time series [6]. This research is described more in detail elsewhere in this report.

14.5 Text mining in qualitative research

Text mining using the SOM presents an interesting methodological opportunity for qualitative research. Qualitative researchers aim to gather rich understanding of human behavior and the reasons for the behavior. In qualitative research, small but focused samples are therefore more often used, rather than large samples. We have argued that the SOM is particularly efficient in improving inference quality within qualitative research, with regard to both confirmatory and exploratory research [9]. Within the theory-driven or deductive mode of qualitative research, the SOM can be used to test the adequacy of conceptual frameworks created before the analysis of the data. In the data-driven or inductive mode, the SOM can be applied in creating emerging category systems describing and explaining the data.

The SOM (and related methods) can be considered as a quantitative method or research tool that is particularly well suited to the aim of respecting complexity rather than trying to do away with it. The SOM can produce not only one but a multitude of perspectives on some data. In relation to very large data sets of the kind, some of these multiple perspectives might be such that no human would, even in principle, be able to produce them. This follows from the fact that the computational method can be used to process writings or sayings of thousands or even millions of persons, something that is beyond the scope of any individual researcher. Yet applying the SOM allows us access to potentially highly relevant and novel categories and patterns that "really are there," even if we do not as yet know it. This would appear to be particularly true when it comes to various nonconscious categorizations. Thus, it would appear that applying the quantitative method of the SOM could take us even beyond situational analysis in that it is capable of revealing subconscious operations of the human mind, which the consciously operating human mind of the situational analyst will never be able to discover [9]. In general, a cartographer of social life can greatly benefit from taking the text mining results into account.

14.6 Analysis of consumer data

In [10] we present the concept of 'open data': a kind of consumer data produced by the consumers themselves from their perspective and for their own purposes that is not intended to be used primarily as consumer data. It is shared publicly in such a way that it can be used as basis for the business and nonprofit organizations in their quest for novelty and understanding of changing consumer trends, also for the benefit of the consumers themselves. We discuss and analyze three cases of opportunities brought by open data: web enhanced brand communities, the weak signals approach and conceptual mapping, which is in its early phase of development.

14.7 Supporting democratic innovation in organizations

When talking about innovations in organizations, our aim is to put the power to innovate in the hands of the people in organizations. We have developed a concept of a tool to support innovative open practices within organizations and to avoid problems often noticed in organizational practice, such as problems in sharing understanding about consumers and markets, and lack of creating organizational memory. [11] The tool is called Note and it helps with the problems of vanishing organizational memory, disappearing or badly accessible notes, and wide range of ideas that are hard to be organized. Note is a shared electronic noteboard where the employees of a company write down their observations, ideas and questions. The underlying data processing system processes the notes and links similar or related ideas together. The text processing is carried out by statistical text mining methods. The notes are short, about 1–20 words each, and they may be written in any language for which there is also textual background material available. The background material is used as a sample of general language for the clustering methods. [15] A demo version of the tool and its data processing methods has been implemented.

14.8 Analysis of political popularity patterns

The complex phenomena of political science are typically studied using qualitative approach, potentially supported by hypothesis- driven statistical analysis of some numerical data. We have examined the use of the self-organizing map in this area and explored the relationship between parliamentary election results and socio-economic situation in Finland between 1954 and 2003 [14]. In the following, we discuss some of the specific results and findings. The variable maps (or component planes, as they are traditionally called) show the distribution of each separate variable on the map. These are presented in Fig. 14.3.

It is commonly believed that being in the government will cause a popularity reduction in the next election. According to our analysis, this is true for the four largest parties: Centre Party (KESK), Social Democratic Party (SDP), National Coalition Party (KOK) and Left Alliance (LEFT). This observation is not however valid for the other parties. There is a strong negative correlation between the Centre Party (KESK) and inflation (COLI(T), COLI(T-1) and COLI(T-2)). During high unemployment the popularity of the Centre Party has been decreasing and during low unemployment it has been increasing. The Centre Party's position as the largest party that has been many times in the government could cause these findings. Voters have punished it because of unfavorable economic situations or developments. The popularity of the National Coalition Party (KOK) has the same feature as the popularity of the Centre Party. During high unemployment it has been decreasing and during low unemployment it has been increasing. During the existence of the Green League (GREENS) the approval ratings of the Social Democratic Party and the Greens have had negative correlation. The popularity of the Left Alliance (LEFT) has been decreasing within the whole period of the study.[14]

A change that took place in the late 1970s is clearly discernable. Many dependences between variables changed their features. Correlations turned from negative to positive and vice versa. For example, turnout has a positive correlation with the Change of Gross Domestic Product per Capita (CGDP(T), CGDP(T-1) and CGDP(T-2)) in the 1950s and 1960s. In the 1990s and 2000s, there is, on the contrary, a negative correlation. Earlier economic growth has potentially provided possibilities to be politically active and later it has made people negligent.[14]

References

[1] H. Bruun, R. Langlais & N. Janasik (2005) Knowledge networking: A conceptual framework and typology. *VEST*, 18(3-4): 73-104.

[2] B. Castellani & F.W. Hafferty (2009) *Sociology and Complexity Science: A New Field of Inquiry*. Springer.

[3] S. Gigerenzer, P. Todd, and the ABC Research Group (1999) *Simple heuristics that make us smart*. New York: Oxford University Press.

[4] Kai Hakkarainen, Ritva Engeström, Sami Paavola, Pasi Pohjola, and Timo Honkela. Knowledge practices, epistemic technologies, and pragmatic web. In *Proceedings of I-KNOW'09 and I-SEMANTICS'09: the 4th AIS SigPrag International Pragmatic Web Conference Track (ICPW 2009)*, pages 683-694. Verlag der Technischen Universität Graz, 2009.

[5] K. Hakkarainen, T. Palonen, S. Paavola & E. Lehtinen (2004) *Communities of networked expertise: Professional and educational perspectives*. Elsevier, Amsterdam.

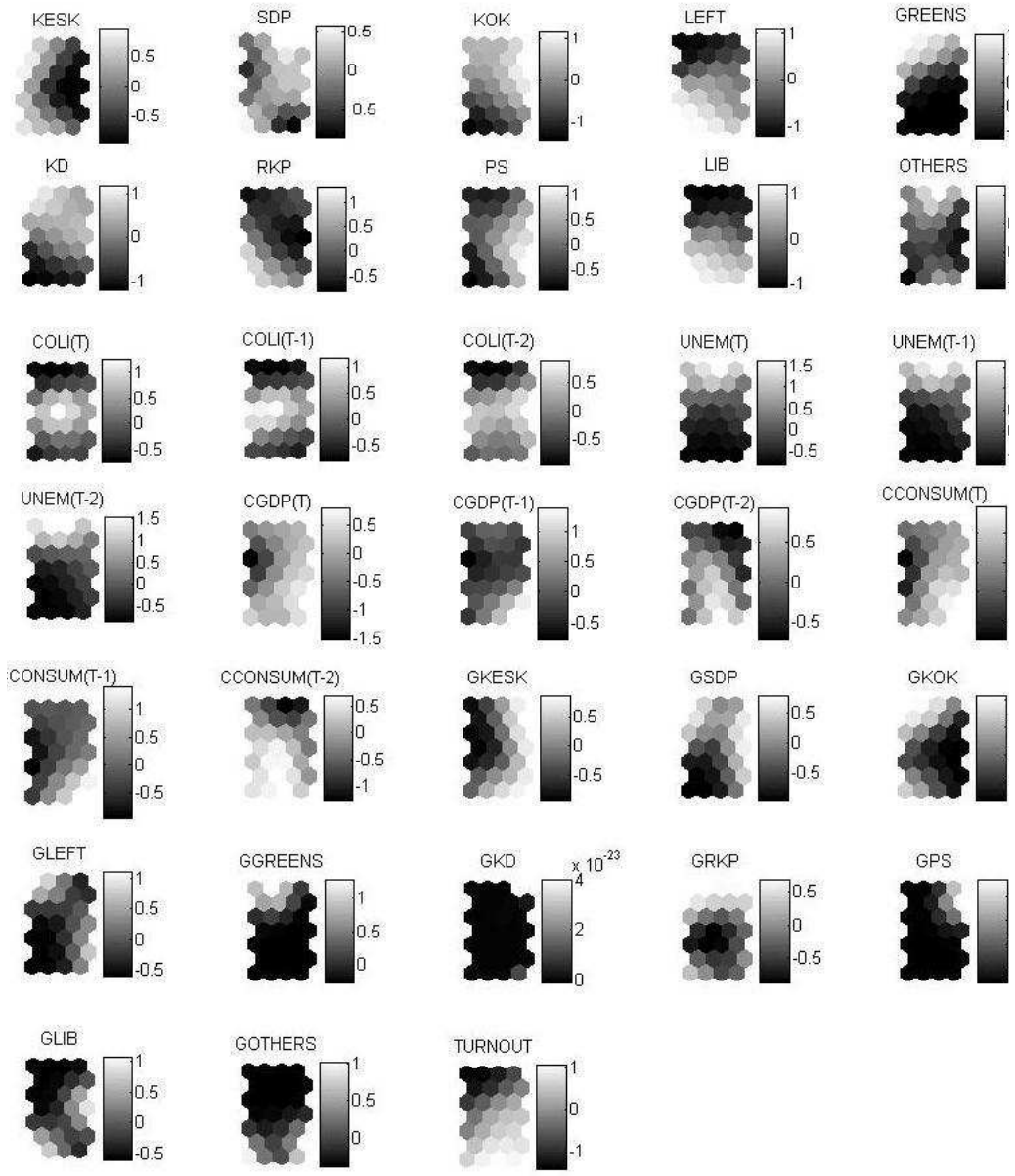


Figure 14.3: The variable maps of all variables used in the study. The acronyms for the political parties are as follows: KESK: Centre Party of Finland, SDP: Social Democratic Party of Finland, KOK: National Coalition Party, VAS: Left Alliance, GREENS: Green League, KD: Christian Democrats in Finland, RKP: Swedish People's Party, PS: True Finns, and LIB: Liberals. National economic conditions are analyzed using four measurements: Change of Cost of Living Index (COLI), Unemployment Rate (UNEM), Change of Gross Domestic Product per Capita (CGDP), and Change of Total Consumption per Capita (CCONSUM). These four monetary values are transformed into constant prices of the year 2000. For each measurement, there are three variables included in the data: the first at elections year (marked with COLI(T), UNEM(T), CGDP(T), and CCONSUM(T)), the second at a year before elections (marked with COLI(T-1), etc.) and the third at two years before elections (marked with COLI(T-2), etc.)

[6] Mark van Heeswijk, Yoan Miche, Tiina Lindh-Knuutila, Peter A. J. Hilbers, Timo Honkela, Erkki Oja, and Amaury Lendasse. Adaptive ensemble models of extreme learning machines for time series prediction. In *Proceedings of ICANN*, Volume 2, pages 305-314, 2009.

[7] Timo Honkela, Nina Janasik, Krista Lagus, Tiina Lindh-Knuutila, Mika Pantzar, and Juha Raitio. Modeling communities of experts - conceptual grounding of expertise. Technical Report TKK-ICS-R24, Helsinki University of Technology, 2009.

[8] Timo Honkela, Ville Könönen, Tiina Lindh-Knuutila, and Mari-Sanna Paukkeri. Simulating processes of concept formation and communication. *Journal of Economic Methodology*, 15(3):245-259, 2008.

[9] Nina Janasik, Timo Honkela, and Henrik Bruun. Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods*, 12(3):436-460, 2009

[10] Tanja Kotro, Tiina Lindh-Knuutila, and Elina Hiltunen. How to analyze various consumer data in the future? In Marileena Koskela and Markus Vinnari, editors, *Future of the consumer society: Proceedings fo the conference Future of the Consumer Sciety, number 7 in FFRC eBook*, pages 135-143. Finland Future Research Centre, May 2009.

[11] Tanja Kotro and Mari-Sanna Paukkeri (2009) Micro democracy - Enhancing Openness in Innovation in an Organizational Context. *In Proceedings of OPEN 2009*. Helsinki, Finland, October 2009.

[12] Lasse Lindqvist, Timo Honkela, and Mika Pantzar, Visualizing practice theory through a simulation model, Technical Repprt E9, Helsinki University of Technology, Laboratory of Computer and Information Science, 2007.

[13] Eric Malmi, Juha Raitio, and Timo Honkela. Modeling practice diffusion with an agent-based social simulation framework. In *Proceedings of the 6th European Social Simulation Association Conference, ESSA 2009*, page 53, Guildford, U.K., September 2009. Extended abstract.

[14] Pyry Niemelä and Timo Honkela. Analysis of parliamentary election results and socio-economic situation using self-organizing map. In *Proceedings of WSOM'09*, pages 209-218, 2009.

[15] Mari-Sanna Paukkeri and Tanja Kotro (2009) Framework for Analyzing and Clustering Short Message Database of Ideas. In *Proceedings of I-KNOW'09,the 9th International Conference on Knowledge Management and Knowledge Technologies*. Graz, Austria, September 2009.

[16] M. Pantzar, The choreography of everyday life: A missing brick in the general evolution theory, *World Futures, The Journal of General Evolution*, vol. 27, no. 3, pp. 207-226, 1989.

[17] E. Shove and M. Pantzar Consumers, producers and practices: understanding the invention and reinvention of nordic walking, *Journal of Consumer Culture*, vol. 1, pp. 43-64, 2005.

[18] R. Sun (2006) *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press.

[19] L. Vygotski (1962) *Thought and Language*. MIT Press. (Originally published in Russian in 1934.)

Chapter 15

Immune system inspired computing

Matti Pöllä and Timo Honkela

15.1 Introduction

Artificial immune systems, motivated by the natural immune principles, are emerging to overcome the limitations of conventional model-based techniques in combating with the uncertainties, vagueness, and high-dimensions of real-world problems.

In a collaboration project, lead by Professor Seppo Ovaska from the Institute of Intelligent Power Electronics of TKK, we have been investigating the fusion of artificial immune systems and soft computing [4] with an application in text data mining [5].

15.2 Anomaly detection

The task of detecting anomalies in a collection of data is one of the primary research topics of immunology-inspired engineering. Biological immune systems have evolved into various successfull mechanisms for detecting bacteria and viruses while having no prior information on them. Correspondingly, immunology-inspired anomaly detection attempts to mimic these mechanism to develop classification algorithms for anomaly detection.

In [3] we have introduced a method for applying a negative selection algorithm to anomaly detection in textual data. The sparsity of discrete sequential data, such as written language, is decreased by analyzing individual character frequencies in a subset of the corpus. We use a collection of Wikipedia articles to show how little information on the original article is needed to detect and locate the changed parts.

The use of generative statistical models has also been the topic of recent research on text anomaly detection. In [1] we have augmented the mixture model scheme by Stibor [2] for arbitrary strings into a mixture-of-multinomials model. Aside from text mining, the presented method for anomaly detection is applicable for other types of symbolic sequential data such as gene and protein sequence analysis. We also compare the use of such generative models with the one-class support vector machine.

References

[1] Matti Pöllä. A Generative Model for Self/Non-Self Discrimination in Strings. *Proceedings of ICANNGA'09: International Conference on Adaptive and Natural Computing Algorithms*, pages 293-302. Springer-Verlag, April 2009.

[2] Thomas Stibor. Discriminating Self from Non-Self with Finite Mixtures of Multivariate Bernoulli Distributions. *Proceedings of Genetic and Evolutionary Computation Conference*, pages 127-134. ACM Press, 2008.

[3] Matti Pöllä and Timo Honkela. Change detection of text documents using negative first-order statistics, *Poceedings of AKRR'08, The Second International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 48-55. Porvoo, Finland, September 2008.

[4] Matti Pöllä, Timo Honkela, and Xiao-Zhi Gao. Biologically inspired clustering: Comparing the neural and immune paradigms. In *Proceedings of NICSO 2007 Workshop on Nature Inspired Cooperative Strategies for Optimization*, pages 179–188, Acireale, 2008. Springer-Verlag. Other paper.

[5] Matti Pöllä and Timo Honkela. Change detection of text documents using negative first-order statistics. In *Proceedings of AKRR'08*, pages 48–55, Porvoo, 2008. Other paper.

Adaptive Informatics Applications

Chapter 16

Intelligent data engineering

Miki Sirola, Kimmo Raivio, Pasi Lehtimäki, Jukka Parviainen, Jaakko Talonen, Golan Lampi, Teemu Poikela, Eimontas Augilius, Olli Simula

16.1 Data analysis in industrial operator support

Miki Sirola, Jukka Parviainen, Jaakko Talonen, Golan Lampi, Teemu Poikela, Eimontas Augilius

Early fault detection with data-analysis tools in nuclear power plants was one of the main goals in NoTeS2-project (test case 4) in TEKES technology program MASI. The industrial partner in this project was Teollisuuden Voima Oy, Olkiluoto nuclear power plant. Data analysis was carried out with real failure data, training simulator data and design based data, such as data from isolation valve experiments. A control room tool, visualization tools and various visualizations were developed.

Fault dynamics and dependencies of power plant elements and variables was inspected to open the way for modelling and creating useful statistics to detect process faults. In our research we succeeded to use data mining to learn from industrial processes and find out dependencies between variables by Principal Component Analysis (PCA) and Self-Organizing Map (SOM). Also a segmentation method was developed to detect automatically different process states of stored datasets.

An adaptive model was developed to primary circulation system to detect leakage in steam lines. A fault was defined as an unpermitted deviation of the variable. Also K-means clustering in time was used for monitoring and detecting pre-stage of process fault [1].

When fault or its pre-stage is detected, current process state should be diagnosed and operators should be informed efficiently. Process monitoring was improved by concepts of generated control limits and alarm balance. All these fault detection and diagnosis methods were programmed with Matlab. Data Management Tool (DMT) is an interface for off-line analysis of stored Olkiluoto datasets including preprocessing, variable selection and other developed methods, see Figure 16.1.

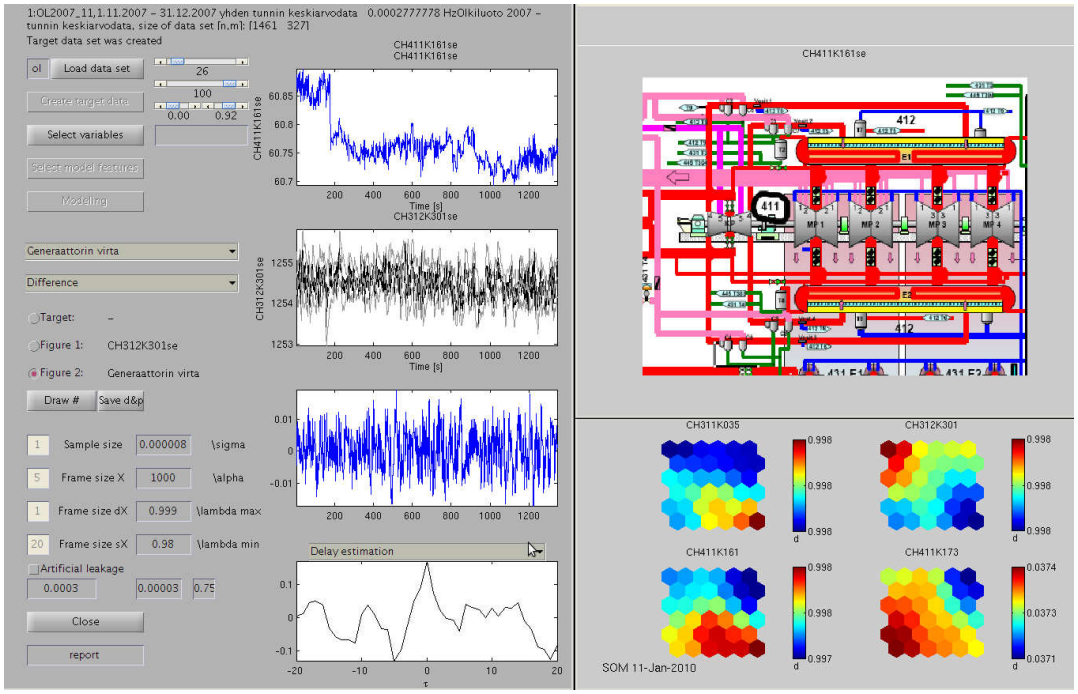


Figure 16.1: DMT User Interface.

Decision support prototype DERSI for failure management in nuclear power plants was developed [2]. It is a control room tool for operator or analysis tool for expert user. It combines neural methods and knowledge-based methods. DERSI utilizes Self-Organizing Map (SOM) method and gives advice by rule-based reasoning. The operator is provided by various informative decision support visualizations, such as SOM maps for normal data and failure data, state U-matrix, quantization error for both component level and state U-matrix, time-series curves and progress visualizations. DERSI tool has been tested in fault detection and separation of simulated data.

All visualizations developed in the project are collected for making a first proposal for wide monitoring screens in [3]. Industrial applications by using SOM are presented in [4]. Also a seminar course on this topic was held in the autumn term 2009.

References

[1] J. Talonen and M. Sirola. Abnormal Process State Detection by Cluster Center Point Monitoring in BWR Nuclear Power Plant. In *Proceedings of the International Conference on Data Mining (DMIN)*, volume I, II, pages 247–252, July 2009.

[2] G. Lampi. Self-organizing maps in decision support: a decision support system prototype. Master's thesis, Helsinki University of Technology, 2009.

[3] M. Sirola, J. Parviainen, J. Talonen, G. Lampi, T. Alhonnoro, and R. Hakala. Early fault detection with SOM based methods and visualizations - new contents for wide monitoring screens. *EHPG-Meeting of OECD Halden Reactor Project*, May 2008. Loen, Norway. 11p.

[4] M. Sirola, J. Talonen, and G. Lampi. SOM based methods in early fault detection of nuclear industry. In *Proceedings of the 17th European Symposium On Artificial Neural Networks ESANN'09*, April 2009.

16.2 Cellular network optimization

Kimmo Raivio, Pasi Lehtimäki

Structure of mobile networks gets more and more complicated when new network technologies are added to the current ones. Thus, advanced analysis and tuning methods are needed to optimize the performance of the network. Adaptive methods can be utilized, for example, to detect anomalous behavior of network elements [3] and to adjust configuration parameters of the network [1].

In order to automate the configuration parameter optimization, a computational method to evaluate the performance of alternative configurations must be available. In data-rich environments like cellular networks, such predictive models are most efficiently obtained with the use of past data records.

In blocking prediction, the interest is to compute the number of blocked requests at different conditions. This can be based on the use of well known Erlang-B formula. The expected value for the number of blocked requests is obtained by multiplying the number of arriving requests with the blocking probability, leading to $B = \lambda p(N_c|\lambda, \mu, N_c)$. The expected value for the congestion time is $C = p(N_c|\lambda, \mu, N_c)$ and the expected value for the number of channels in use is $M = \sum_{n=0}^{N_c} np(n|\lambda, \mu, N_c)$.

In [1], it was shown that the Erlang-B formula does not provide accurate predictions for blocking in GSM networks if low sampling rate measurements of arrival process are used in the model. More traditional regression methods can be used for the same purpose with the assist of knowledge engineering approach in which Erlang-B formula and regression methods are combined. With the use of Erlang-B formula, the dependencies between B, C and M that remain the same in each base station system need not be estimated from data alone. The data can be used to estimate other relevant and additional parameters that are required in prediction. In this research, a method to use Erlang-B formula and measurement data to predict blocking has been developed. The regression techniques are used to estimate the arrival rate distribution describing the arrival process during short time periods. The Erlang-B formula is used to compute the amount of blocking during the short time periods.

Suppose that the time period is divided into N_s segments of equal length. Also, assume that we have a vector $\boldsymbol{\lambda} = [0 \;\; 1\Delta_\lambda \;\; 2\Delta_\lambda \;\; \ldots \;\; (N_\lambda - 1)\Delta_\lambda]$ of N_λ possible arrival rates per segment with discretization step Δ_λ. Let us denote the number of blocked requests during a segment with arrival rate λ_i with $B_i = \lambda_i p(N_c|\lambda_i, \mu, N_c)$, where $p(N_c|\lambda_i, \mu, N_c)$ is the blocking probability given by the Erlang distribution. Also, the congestion time and the average number of busy channels during a segment with arrival rate λ_i are denoted with $C_i = p(N_c|\lambda_i, \mu, N_c)$ and $M_i = \sum_{n=0}^{N_c} np(n|\lambda_i, \mu, N_c)$. In other words, the segment-wise values for blocked requests, congestion time and average number of busy channels are based on the Erlang-B formula.

Now, assume that the number of segments with arrival rate λ_i is θ_i and $\sum_i \theta_i = N_s$. Then, the cumulative values over one hour for the number of requests T, blocked requests B, congestion time C and average number of busy channels M can be computed with

$$\begin{bmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_{N_\lambda} \\ B_1 & B_2 & \ldots & B_{N_\lambda} \\ \frac{C_1}{N_s} & \frac{C_2}{N_s} & \ldots & \frac{C_{N_\lambda}}{N_s} \\ \frac{M_1}{N_s} & \frac{M_2}{N_s} & \ldots & \frac{M_{N_\lambda}}{N_s} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{N_\lambda} \end{bmatrix} = \begin{bmatrix} T \\ B \\ C \\ M \end{bmatrix} \tag{16.1}$$

or in matrix notation $\mathbf{X}\boldsymbol{\theta} = \mathbf{Y}$.

Now, the problem is that the vector $\boldsymbol{\theta}$ is unknown and it must be estimated from the data using the observations of $\mathbf{Y}$ and matrix $\mathbf{X}$ which are known a priori. Since the output vector $\mathbf{Y}$ includes variables that are measured in different scales, it is necessary to include weighting of variables into the cost function. By selecting variable weights according to their variances estimated from the data, the quadratic programming problem

$$\min_{\boldsymbol{\theta}} \quad \left\{ \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} + \mathbf{f}^T \boldsymbol{\theta} \right\} \tag{16.2}$$

$$w.r.t \quad 0 \leq \theta_i \leq N_s, \ i = 1, 2, ..., N_\lambda, \tag{16.3}$$

$$\sum_{i=1}^{N_\lambda} \theta_i = N_s \tag{16.4}$$

is obtained where $\mathbf{f} = -\mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{Y}$ and $\mathbf{H} = \mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{X}$ include the weighting matrix $\mathbf{W}$. In other words, the goal is to find the vector $\boldsymbol{\theta}$ that provides the smallest prediction errors for variables T, B, C and M.

The optimization problem could be solved for each of the N_d observation vectors separately, leading to N_d solution vectors $\boldsymbol{\theta}$ for hour h. Since we are interested in long-term prediction of blocking, we should somehow combine the solution vectors so that behavior common to all solution vectors are retained and non-regular properties of the demand are given less attention.

Using probabilistic models solutions of different arrival rates can be combined. At first the total number of arrived requests is estimated from probabilities of observing a segment with certain arrival rate. The same model can be used to map segment-wise blocking candidates to the total number of occurrences of blocked requests during one period. Similarly, the cumulative values for the average number of busy channels and the congestion time can be computed [2].

References

[1] P. Lehtimäki and K. Raivio. Combining measurement data and Erlang-B formula for blocking prediction in GSM networks. In *Proceedings of The 10th Scandinavian Conference on Artificial Intelligence (SCAI)*, Stockholm, Sweden, May 26 - 28 2008.

[2] Pasi Lehtimäki. *Data Analysis Methods for Cellular Network Performance Optimization.* Doctoral dissertation, TKK Dissertations in Information and Computer Science TKK-ICS-D1, Helsinki University of Technology, Faculty of Information and Natural Sciences, Department of Information and Computer Science, Espoo, Finland, April 2008.

[3] M. Multanen, K. Raivio, and P. Lehtimäki. Outlier detection in cellular network data exploration. In *Proceedings of the 3rd International Workshop on Performance Analysis and Enhancement of Wireless Networks (PAEWN)*, Okinawa, Japan, March 25 - 28 2008.

Chapter 17

Time series prediction

Amaury Lendasse, Francesco Corona, Federico Montesino-Pouzols, Patrick Bas, Antti Sorjamaa, Mark van Heeswijk, Laura Kainulainen, Eric Severin, Elia Liitiäinen, Tuomas Kärnä, Yu Qi, Yoan Miche, Emil Eirola, Dušan Sovilj, Olli Simula

17.1 Introduction

Amaury Lendasse

What is Time series prediction? Time series prediction (TSP) is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. The common point to their problems is the following: how can one analyze and use the past to predict the future? Many techniques exist: linear methods such as ARX, ARMA, etc., and nonlinear ones such as artificial neural networks. In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict the future values. The common difficulty to all the methods is the determination of sufficient and necessary information for an accurate prediction.

A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information.

Our contributions in TSP research. The TSP group is a new research group. It has been created in 2004. A notable achievement has been the organization of the first European Symposium on Time Series Prediction (ESTSP'08) on September 2008 in Porvoo [1]. For this symposium, a time series competition has been organized and a benchmark has been created.

In the reporting period 2006 - 2007, TSP research has been established as a new project in the laboratory. Nevertheless, TSP research has already been extended to a new direction: Chemoinformatics.

This Chapter starts by introducing some theoretical advances undertaken during the reporting period, including the presentation of the ESTSP'08 competition. Also the problem of input selection for TSP is reported. The applications range includes Chemoinformatics.

In 2010, as we believe that i'The Times They Are A-Changin", the TSP group will evolve and will become the "Environmental and Industrial Machine Learning Group".

17.2 European Symposium on Time Series Prediction

Amaury Lendasse, Olli Simula and Timo Honkela

Introduction

Time series forecasting is a challenge in many fields. In finance, one forecasts stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. The common point to their problems is the following: how can one analyze and use the past to predict the future? Many techniques exist including linear methods such as ARX or ARMA, and nonlinear ones such as the ones used in the area of machine learning. In general, these methods try to build a model of the process that is to be predicted. The model is then used on the last values of the series to predict future ones. The common difficulty to all methods is the determination of sufficient and necessary information for a good prediction. If the information is insufficient, the forecasting will be poor. On the contrary, if information is useless or redundant, modeling will be difficult or even skewed. In parallel with this determination, a suitable prediction model has to be selected. In order to compare different prediction methods several competitions have been organized, for example, the Santa Fe Competition, the CATS Benchmark Competition and the ESTSP'07 Competition.

After the competitions, their results have been published and the time series have become widely used benchmarks. The goal of these competitions is the prediction of the subsequent values of a given time series (3 to 100 values to predict). Unfortunately, the long-term prediction of time series is a very difficult task. Furthermore, after the publication of results, the real values that had to be predicted are also published. Thereafter, it becomes more difficult to trust in new results that are published: knowing the results of a challenge may lead, even unconsciously, to bias the selection of model; some speak about "data snooping". It becomes therefore more difficult to assess newly developed methods, and new competitions have to be organized.

This text is based on papers presented at the joined ESTSP'08 (European Symposium on Time Series Prediction) [1] and AKRR'08 (Adaptive Knowledge Representation and Reasoning) conferences. This shared event took place in Porvoo, Finland, from 17th to 19th of September, 2008. The goal of joining these conferences was to create an interdisciplinary forum for researchers who may widen their scope of attention beyond the usual scope of research. The crossfertilization took place, for instance, by offering the attendees shared keynote talks. Prof. Marie Cottrell (Paris University 1) gave a talk on data analysis using Self-Organizing Maps. Prof. José Príncipe (University of Florida) described information theoretic learning and kernel methods. Dr. Harri Valpola (Helsinki University of Technology) explained how to extract abstract concepts from raw data using statistical machine learning methods. One specific shared theme of interest was anticipation, i.e., how an agent makes decisions based on predictions, expectations, or beliefs about the future. Anticipation is an important concept when complex natural cognitive systems are considered.

ESTSP'08 Competition

The goal of the ESTSP'08 competition was to predict the future of three very different Time Series[1]. Firstly, the length and the sampling period of the time series are very different. Secondly, the origin of each time series varies. The data and the origins, i.e., environment, electric load, and internet traffic, are described below in more detail. In order to provide the participants an equal opportunity for success, the origins of the three time series were kept secret until the end of the competition.

Data sets

Chemical descriptors of environmental condition

This series is part of a multidimensional time series of monthly averages of different chemical descriptors of a certain area of the Baltic Sea. The series is made of 354 samples and spans for 29.5 years. This competition data set is shown in Figure 17.1. For this time series, the goal was to predict the next 18 values of the third time series, using the two other one as exogenous variables.

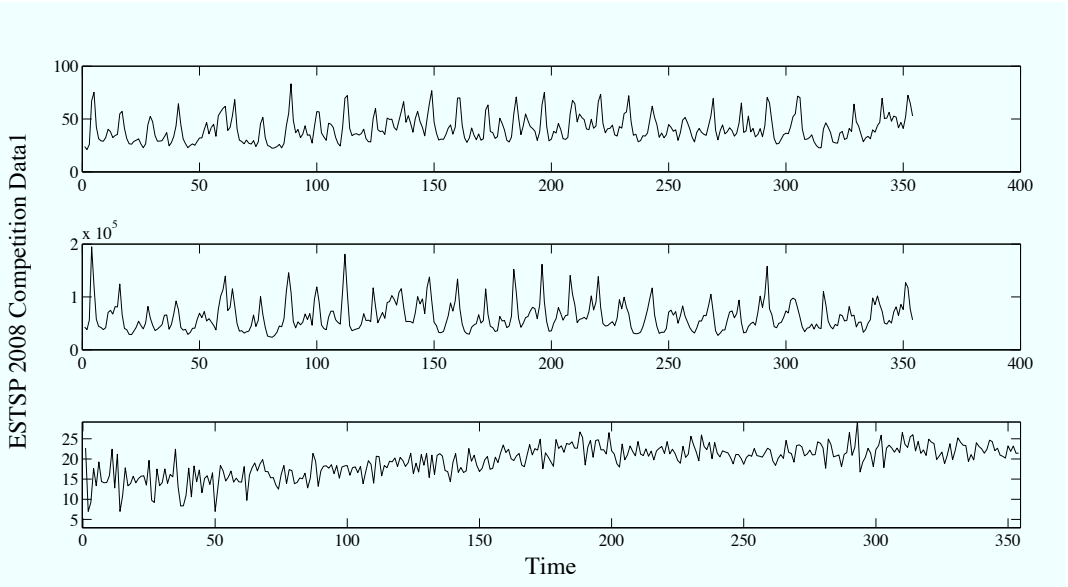


Figure 17.1: ESTSP 2008 competition data 1.

Traffic in a data network

The second dataset from the ESTSP 2008 competition is a univariate time series consisting of 1300 samples that describe the daily average amount of traffic in a data network. The competition data set 2 is shown in Figure 17.2. For this time series, the goal was to predict the next 100 values of the time series.

Electric load

The third dataset was a univariate time series consisting of 31614 samples that describe the daily average amount of eletric load. The competition data set 2 is shown in Figure 17.3. For this time series, the goal is the prediction of the next 200 values of the time series.

[1]The data sets can be downloaded from
http://www.cis.hut.fi/projects/tsp/index.php?page=timeseries.

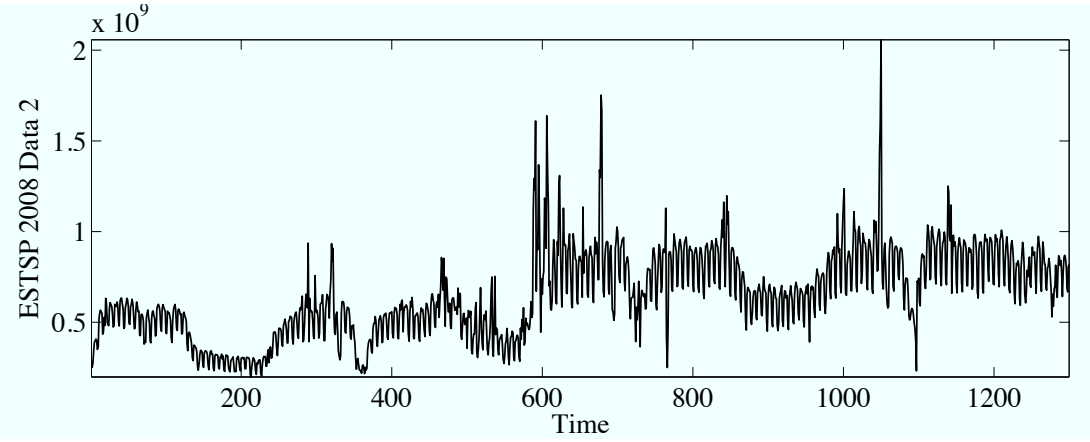


Figure 17.2: ESTSP 2008 competition data 2.

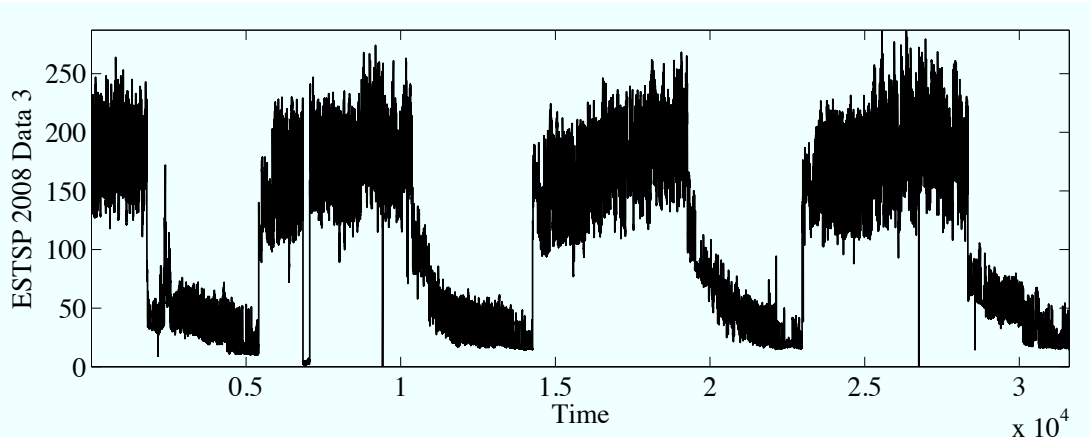


Figure 17.3: ESTSP 2008 competition data 3.

Results

Twenty sets of predictions have been submitted to the competition. The results in Table 17.1 present the Normalized Test Mean Squared Error for the 3 predictions respectively. We present only the results of the participants that agreed to have their results published. The winners of the competition were Rubio, Herrera, Pomares, Rojas and Guillen.

Summary of the Best Papers

The papers can be classified in 3 distinct categories:

1. The authors that participated to the competition.

	Kourentzes	Bontempi	Olteanu	Wyffels	Espinoza	Adeodato	Rubio	Montesino
Data 1	0.07	0.12	0.19	0.157	0.112	0.151	0.079	0.16
Data 2	0.212	0.431	0.359	0.529	0.266	0.49	0.208	0.4
Data 3	0.25	1.802	1.655	1.582	0.464	1.611	0.036	1.344
Total	0.178	0.785	0.735	0.756	0.281	0.751	0.107	0.635

Table 17.1: Competition Results: Test NMSE for Each Data Set.

2. The papers that presents new methods for the analysis and/or prediction of Time Series but did not participate in the competition.

3. The papers that participated in the ESTSP08-AKRR'08 Special Session on Prediction for Finance organized by Prof. Eric Séverin.

Competition Papers

Crone and Kourentzes propose a data driven, fully automated methodology to specify multilayer perceptrons for time series prediction using a combination of iterative (neural network) filters and wrappers. Their approach is capable of identifying unknown time series frequencies, multiple overlying seasonality, and additional relevant features without human expert intervention. The approach has shown promising performance in forecasting by ranking second in the ESTSP competition.

Pouzols and Barriga deal with an automatic methodology for clustering-based fuzzy inference models. A number of clustering methods are compared and an extension of Improved Clustering for Function Approximation is proposed. The appraoch yields compact models and its accuracy and speed compare favorably against MLP, LS-SVM and ELM models for a diverse set of time series benchmarks.

Ben Taieb, Sorjamaa and Bontempi present a new multiple-output approaches for Multi-Step-Ahead Time Series Forecasting and compares it to state-of-the-art approaches. The extensive validation made with the series of the NN3 competition shows that the multiple-output paradigm is very promising and able to outperform conventional techniques.

Reservoir Computing has been shown to perform well in chaotic time series prediction. **Wyffels and Schrauwen** extend these results by a comparison of multiple Reservoir Computing strategies for time series prediction (including research on regularization, influence of reservoir size and decomposition) in the domain of noisy, seasonal time series prediction for industrial purposes. They compare their approach to standard approaches such as ARIMA modeling and NAR modeling using LS-SVMs.

Rubio, Herrera, Pomares, Rojas and Guillen present a kernelized version of the weighted k-nearest neighbours method (KWKNN) for regression problems and address the creation of specific-to-problem kernels for time series data. This unified framework for kernel and k-nearest neighbours methods allows for a comparison of KWKNN with LSSVM using time series prediction examples with interesting results. Additionally, a parallel implementation of KWKNN, developed in order to speed up the method and make it practical for large datasets, is proposed and applied to a large scale problem.

General Papers

Sovilj, Sorjamaa, Yu, Miche and Séverin present a methodology for long-term time series prediction that can also be applied to standard regression tasks. The methodology consists of two main steps: (1) input variable scaling or projection with Delta Test, optimized with Genetic Algorithm, and (2) prediction on the projected data using two models, Optimally-Pruned Extreme Learning Machine and Optimally-Pruned k-Nearest Neighbors. The methodology is tested on two time series prediction tasks and one financial regression problem.

Nybo provides an applied perspective from the petroleum industry. Normally conservative, this industry nonetheless shows an increasing interest in machine learning and data mining. The paper gives a taste of the new opportunities in this industry and goes

on to show how a successful choice of machine learning algorithms becomes governed by the industry's work processes and the user's behavioural mode.

Souza and Barreto provide a comprehensive performance evaluation of the use of vector quantization (VQ) algorithms to building local models for inverse system identification. Statistical hypothesis testing is carried out through the Kolmogorov-Smirnov test in order to study the influence of the VQ algorithms on the performances of the local models. Tests on four benchmarking input-output time series reveal that the resulting local models achieve performances superior to standard global MLP-based model.

Lemke and Gabrys describe how the performance of the time series forecasting algorithms differ depending on the data set used. However, for a limited data set of similar time series, it can be possible to determine one particular method or combination of methods that performs best. Following this idea, the article presents an empirical study extracting characteristics of time series in order to generate domain knowledge. This knowledge is then used to dynamically select or combine different forecasting algorithms.

Mateo, Sovilj and Gadea present a method that uses genetic algorithms to select an optimum set of input variables that minimizes the Delta Test on a dataset. The nearest neighbor computation has been speeded up by using an approximate method. The scaling and projection of variables has been addressed to improve the interpretability.

Guillen, Herrera, Rubio, Pomares, Lendasse and Rojas present a totally new approach for the problem of filtering the outliers, reducing the noise and defining a good subset of samples. The approach is based in the concept of Mutual Information with the advantage of just having one parameter to be tuned. The simple idea is efficient and easy to implement, providing satisfactory results within a wide range of problems.

Korpela, Mäkinen, Nöjd, Hollmén and Sulkava present a Markov-switching autoregressive model. Its performance is compared with other statistical and machine learning methods in a new kind of real-world change detection problem with environmental time-series.

Financial Prediction Papers

du Jardin presents two main results. It is shown that a neural-network-based model for predicting bankruptcy performs better when designed with appropriate variable selection techniques than when designed with methods commonly used in the financial literature. Furthermore, it has been found that there is a relationship between the structure of a prediction model and its ability to reduce Type I errors.

Séverin deals with the advantages of the self-organizing map algorithm in the field of corporate finance. Not only the SOM method is able to improve the classical method for bankruptcy prediction but it also questions the scoring models.

17.3 Tools for long-term prediction of time series

Amaury Lendasse, Yu Qi, Yoan Miche, Emil Eirola, Dusan Sovilj, Olli Simula and Antti Sorjamaa

The time series prediction problem is the prediction of future values based on the previous values and the current value of the time series (see Equation 17.1).

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{17.1}$$

The previous values and the current value of the time series are used as inputs for the prediction model. One-step ahead prediction is needed in general and is referred as Short-Term Prediction. But when multi-step ahead predictions are needed, it is called Long-Term Prediction problem.

Unlike the Short-Term time series prediction, the Long-Term Prediction is typically faced with growing uncertainties arising from various sources. For instance, the accumulation of errors and the lack of information make the prediction more difficult. In Long-Term Prediction, performing multiple steps ahead prediction, there are several alternatives to build models. Two variants of prediction strategies are studied and compared:the Direct (see Equation 17.2) and the Recursive Prediction Strategies (see Equation 17.1).

$$\hat{y}_{t+k} = f_k(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{17.2}$$

In order to perform Long-Term Prediction, several tools have been studied and developed:

- Nonparametric noise estimation
- Imputation of Missing Data
- OP-ELM and Ensembles of ELM

17.4 Nonparametric noise estimation

Elia Liitiäinen, Francesco Corona, Emil Eirola, Olli Simula and Amaury Lendasse

The residual variance estimation problem (or Nonparametric noise Estimation) is well-known in machine learning and statistics under various contexts. Residual variance estimation can be viewed as the problem of estimating the variance of the part of the output that cannot be modeled with the given set of input variables. This type of information is valuable and gives elegant methods to do model selection. While there exist numerous applications of residual variance estimators to supervised learning, time series analysis and machine learning, it seems that a rigorous and general framework for analysis is still missing. For example, in some publications the theoretical model assumes additive noise and independent identically distributed (iid) variables. The principal objective of our work is to define such a general framework for residual variance estimation by extending its formulation to the non-iid case. The model is chosen to be realistic from the point of view of supervised learning. Secondly, we view two well-known residual variance estimators, the Delta test and the Gamma test in the general setting and we discuss their convergence properties.

Contributions:

Minimizing the Delta test for variable selection in regression problems

There exists a wide variety of models that are able to approximate any function such as Radial Basis Function Neural Networks , Multilayer Perceptrons, Fuzzy Systems , Gaussian Process , Support Vector Machines (SVM) and Least Square SVM, etc. however, they all suffer from the Curse of Dimensionality. As the number of dimensions d grows, the number of input values required to sample the solution space increases exponentially, this means that the models will not be able to set their parameters correctly if there are not enough input vectors in the training set. Many real life problems present this drawback since they have a considerable amount of variables to be selected in comparison to the few number of observations. Thus, efficient and effective algorithms to reduce the dimensionality of the data sets are required. Another aspect that is improved by selecting a subset of variables is the interpretability of the designed systems.

The literature presents a wide number of methodologies for feature or variable selection although they have been focused on classification problems. Therefore, specific algorithms for regression must be designed. Recently, it has been demonstrated in how the Delta Test (DT) is a quite powerful tool to determine the quality of a subset of variables. The latest work related to feature selection using the DT consisted in the employment of a local search technique such as Forward-Backward. However, there are other alternatives that allow to perform a global optimization of the variable selection like Genetic Algorithms (GA) and Tabu Search (TS). One of the main drawbacks of using global optimization techniques is their computational cost. Nevertheless, the latest advances in computer architecture provide powerful clusters without requiring a large budget, so an adequate parallelization of these techniques might ameliorate this problem. This is quite important in real life applications where the response time of the algorithm must be acceptable from the perspective of a human operator. Our reserach proposes several new approaches to perform variable selection using the DT as criterion to decide if a subset of variables is adequate or not. The new approaches are based in local search methodologies, global optimization techniques and the hybridization of both [2].

Residual variance estimation in machine learning

The problem of residual variance estimation consists of estimating the best possible generalization error obtainable by any model based on a finite sample of data [3, 4, 5]. Even though it is a natural generalization of linear correlation, residual variance estimation in its general form has attracted relatively little attention in machine learning. In our research, we examine four different residual variance estimators and analyzed their properties both theoretically and experimentally to understand better their applicability in machine learning problems. The theoretical treatment differs from previous work by being based on a general formulation of the problem covering also heteroscedastic noise in contrary to previous work, which concentrates on homoscedastic and additive noise. Secondly, we demonstrate practical applications in input and model structure selection. The experimental results show that using residual variance estimators in these tasks gives good results often with a reduced computational complexity, while the nearest neighbor estimators are simple and easy to implement.

17.5 Imputation of missing data in climatology and finance

Antti Sorjamaa, Olli Simula and Amaury Lendasse

Meteorology and climate research are two rapidly growing fields with an increasing need for accurate and large measurement datasets. The African continent represents a clear example of the current challenges in these fields. The drought and humidity imbalance create extreme conditions for both the people on the continent and the very necessary research. Lake Tanganyika is located in the African Rift in the center of the African continent. It is an important source of proteins for the people around it and the fish industry provides not only the food for the people, but also gives thousands of workers a job.
The importance to the people and the extraordinary size and shape of the lake make it really valuable for the climate research, but the size brings also difficulties. The size and the shape of the lake make it hard to adequately measure the bio-geo-hysical parameters, such as surface temperature. But due to the current political and economical situation in Africa, the satellite is the only valid option. The data measured by satellite includes a vast number of missing values, due to clouds, technical difficulties and even heavy smoke from forest fires. The missing values make a posteriori modeling a difficult problem and the filling procedure a mandatory preprocessing step before climate modeling.
A great number of methods have been already developed for solving the problem by filling the missing values, for example, Kriging and several other Optimal Interpolation methods, such as Objective Analysis. One of the emerging approaches for filling the missing values is the Empirical Orthogonal Functions (EOF) methodology. The EOF is a deterministic methodology, enabling a linear projection to a high-dimensional space. Moreover, the EOF models allow continuous interpolation of missing values even when a high percentage of the data is missing. In our research, an improvement to the standard EOF method is presented, called EOF Pruning. It enhances the accuracy of the EOF methodology and even speeds up the calculation process [6].
Academics as well as practitioners often face the problem of missing data in financial time series. Non-quotation date, too recent inception date, intention not to report a bad performance or mistake of data provider are some of the reasons why missing values occur recurrently in financial databases. Moreover, in order to achieve good performance, most financial models need complete and cylindrical samples. Thus, most of the time, imputation methods have to be applied before running the model. A number of methods have been developed to solve the problem and fill the missing values, both commercial and academical. The methods in both sectors can be classified into two distinct categories : deterministic methods and stochastic methods. Self-Organizing Maps (SOM) aim to ideally group homogeneous individuals, highlighting a neighborhood structure between classes in a chosen lattice. The SOM algorithm is based on unsupervised learning principle where the training is entirely stochastic, data-driven. No information about the input data is required. Recent approaches propose to take advantage of the homogeneity of the underlying classes for data completion purposes. Furthermore, the SOM algorithm allows projection of high-dimensional data to a low-dimensional grid. Through this projection and focusing on its property of topology preservation, SOM allows nonlinear interpolation for missing values.
Empirical Orthogonal Functions (EOF) are deterministic models, enabling linear projection to high-dimensional space. They have also been used to develop models for finding missing data. Moreover, EOF models allow continuous interpolation of missing values, but are sensitive to the initialization. Our research proposes a new method, which combines the advantages of both the SOM and the EOF. The nonlinearity property of the SOM

is used as a de-noising tool and then continuity property of the EOF method is used to recover missing data efficiently [7].

17.6 OP-ELM and ensembles of ELM

Yoan Miche, Antti Sorjamaa, Mark van Heeswijk, Tiina Lindh-Knuutila, Timo Honkela, Erkki Oja, Olli Simula and Amaury Lendasse

,

The amount of information is increasing rapidly in many fields of science. It creates new challenges for storing the massive amounts of data as well as to the methods, which are used in the data mining process. In many cases, when the amount of data grows, the computational complexity of the used methodology also increases.
Feed-forward neural networks are often found to be rather slow to build, especially on important datasets related to the data mining problems of the industry. For this reason, the nonlinear models tend not to be used as widely as they could, even considering their overall good performances. The slow building of the networks comes from a few simple reasons; many parameters have to be tuned, by slow algorithms, and the training phase has to be repeated many times to make sure the model is proper and to be able to perform model structure selection (number of hidden neurons in the network, regularization parameters tuning. . .).
Guang-Bin Huang et al. propose an original algorithm for the determination of the weights of the hidden neurons called Extreme Learning Machine (ELM). This algorithm decreases the computational time required for training and model structure selection of the network by hundreds. Furthermore, the algorithm is rather simplistic, which makes the implementation easy.
In our research, a methodology called Optimally-Pruned ELM (OP-ELM), based on the original ELM, is proposed. The OP-ELM methodology is compared using several experiments and two well-known methods, the Least-Squares Support Vector Machine (LS-SVM) and the Multilayer Perceptron (MLP). Finally, a toolbox for performing the OP-ELM has been developed [8].

Ensembles of ELM have also been used for Time Series Prediction. A large number of application areas of time series prediction involve nonstationary phenomena. Therefore, contrary to the stationary case, one cannot assume that one can use what has been learned from past data and one has to keep learning and adapting the model as new samples arrive. Possible ways of doing this include: 1) retraining the model repeatedly on a finite window of past values and 2) using a combination of different models, each of which is specialized on part of the state space.
Besides the need to deal with nonstationarity, another motivation for such an approach is that one can drop stationarity requirements on the time series. This is very useful, since often we cannot assume anything about whether or not a time series is stationary.
Ensemble methods have been applied in various forms (and under various names) to time series prediction, regression and classification. A non-exhaustive list of literature that discusses the combination of different models into a single model includes bagging, boosting, committees, mixture of experts, multi-agent systems for prediction, classifier ensembles, among others.
In order to construct the ensemble model, a number of Extreme Learning Machines (ELMs) of varying complexity are generated, each of which is individually trained on the data. After training, these individual models are combined in an ensemble model. The output of the ensemble model is a weighted linear combination of the outputs of the individual models. During the test phase, the ensemble model adapts this linear combination over time with the goal of minimizing the prediction error: whenever a particular model has

bad prediction performance (relative to the other models) its weight in the ensemble is decreased, and vice versa. In our first experiments, we tested the performance of this adaptive ensemble model in repeated one-step ahead prediction on a time series that is known to be stationary (the Santa Fe A Laser series). The main goal of this experiment is to test the robustness of the model and to investigate the different parameters influencing the performance of the model. In the second experiments, the model is applied to another time series (Quebec Births) which is nonstationary and more noisy than the Santa Fe time series [9].

17.7 Chemoinformatics

Francesco Corona, Elia Liitiäinen, Tuomas Kärnä, Olli Simula and Amaury Lendasse

Many analytical problems related to spectrometry require predicting a quantitative variable through a set of measured spectral data. For example, one can try to predict a chemical component concentration in a product through its measured infrared spectrum. In recent years, the importance of such problems in various fields including the pharmaceutical, food and textile industries have grown dramatically. The chemical analysis by spectrophotometry rests on the fast acquisition of a great number of spectral data (several hundred, even several thousands).

In spectrometric problems, one is often faced with databases having more variables (spectra components) than samples; and almost all models use at least as many parameters as the number of input variables. These two problems, colinearity and risk of overfitting, already exist in linear models. However, their effect may be even more dramatic when nonlinear models are used (there are usually more parameters than in linear models, and the risk of overfitting is higher). In such high-dimensional problems, it is thus necessary to use a smaller set of variables than the initial one.

We have proposed methods to select spectral variables by using concepts from information theory [10, 11, 12, 13, 14]:

- the measure of mutual information
- the measure of topological relevance on the Self-Organizing Map
- the Functional Data Analysis (FDA)
- Nonparametric Noise Estimation

One particular application has been studied in the field of Oil Production.

In this industrial application, there has been applied process data. The aim has been to get new empirical modeling tools, which are based on information technology. The outcome has been emphasized on tools, which are suitable in fast data mining from large data sets. The test cases have included:

- Analysis of instrumental data, on-line monitoring data and quality data
- Non-linear processes
- Identification of delays between stages in industrial processes
- Robust variable selection methods

17.8 Steganography and steganalysis

Yoan Miche, Amaury Lendasse, Patrick Bas and Olli Simula

Steganography has been known and used for a very long time, as a way to exchange information in an unnoticeable manner between parties, by embedding it in another, apparently innocuous, document. For example, during the 80's, Margaret Thatcher decided to have each word processor of the government's administration members changed with an unique word spacing for each, giving a sort of invisible signature to documents. This was done to prevent the continuation of sensitive government information leaks.

Nowadays steganographic techniques are mostly used on digital contents. The online newspaper, Wired News, reported in one of its articles on steganography that several steganographic contents have been found on web-sites with very large image database such as eBay.

Most of the time research about steganography is not as much to hide information, but more to detect that there is hidden information. This reverse part of the steganography is called steganalysis and is specifically aimed at making the difference between genuine documents, and steganographied – called stego – ones. Consequently, steganalysis can be seen as a classification problem where the goal is to build a classifier able to distinguish these two sorts of documents.

During the steganographic process, a message is embedded in an image so that it is as undetectable as possible. Basically, it uses several heuristics in order to guarantee that the statistics of the stego content (the modified image) are as close as possible to the statistics of the original one. Afterwards, steganalysis techniques classically use features extracted from the analyzed image and an appropriately trained classifier to decide whether the image is genuine or not.

In our work, a widely used and known set of 193 image features has been used. Theses features consider statistics of JPEG compressed images such as histograms of DCT coefficients for different frequencies, histograms of DCT coefficients for different values, global histograms, blockiness measures and co-occurrence measures. The main purpose of this high number of features is to obtain a model able to detect about any steganographic process.

The usual process in steganalysis is then to train a classifier according to the extracted features. Consequently a set of 193 features for each image of the database is obtained, giving an especially high dimensionality space for classifiers to work on. Earlier research about these high dimensionality spaces has shown that a lot of issues come out when the number of features is as high as this one.

The main idea behind the carried out work [15, 16] is to give insights on proper handling and use of such high dimensionality datasets; indeed, these are very common in the steganography/steganalysis field and users tend not to respect basic principles (for example having a sufficient number of samples regarding the dimensionality of the problem).

17.9 Bankruptcy prediction

Yu Qi, Laura Kainulainen, Eric Severin, Olli Simula and Amaury Lendasse

Bankruptcies are not only financial but also individual crises which affect many lives. Although unpredictable things may happen, bankruptcies can be predicted to some extent.

This is important for both the banks and the investors that analyze the companies, and for the companies themselves. The aim of our research is to see, whether new machine learning models combined with variable selection perform better than traditional models: Linear Discriminant Analysis, Least Squares Support Vector Machines and Gaussian Processes. They form a good basis for comparison, since LDA is a widely spread technique in the financial tradition of bankruptcy prediction, LSSVM is an example of Support Vector Machine classifiers and Gaussian Processes is a relatively new Machine Learning method.

Since all the possible combinations of the variables cannot be evaluated due to time constraints, forward selection may offer a fast and accurate solution for finding suitable variables.

Our main results can be found in [17, 18, 19].

References

[1] A. Lendasse. European Symposium on Time Series Prediction, ESTSP'08, Amaury Lendasse editor, ISBN 978-951-22-9544-9.

[2] A. Guillén, D. Sovilj, F. Mateo, I. Rojas and A. Lendasse, Minimizing the Delta Test for Variable Selection in Regression Problems International Journal of High Performance Systems Architecture, Vol. 4, pp. 269-281, 2008.

[3] E. Liitiäinen, M. Verleysen, F. Corona and A. Lendasse, Residual variance estimation in machine learning, Neurocomputing, October 2009, pp. 3692-3703.

[4] E. Liitiäinen, F. Corona, A. Lendasse, On non-parametric residual variance estimation Neural Processing Letters, December 2008,pp. 155-167.

[5] E. Liitiäinen, A. Lendasse andF. Corona, Bounds on the mean power-weighted nearest neighbour distance, Proceedings of the Royal Society A, September 2008, pp. 2293-2301.

[6] A. Sorjamaa, A. Lendasse, Y. Cornet and E. Deleersnijder, An improved methodology for filling missing values in spatiotemporal climate data set, Computational Geosciences, January 2010, pp. 55-64.

[7] A. Sorjamaa, P. Merlin, B. Maillet and A. Lendasse, A Non-Linear Approach for Completing Missing Values in Temporal Databases, European Journal of Economic and Social Systems, November 2009, pp. 99-117.

[8] Y. Miche, A. Sorjamaa and A. Lendasse, OP-ELM: Theory, Experiments and a Toolbox, LNCS - Artificial Neural Networks - ICANN 2008 - Part I, September 2008, pp. 145-154.

[9] M. van Heeswijk, Y. Miche, Tiina Lindh-Knuutila, Peter A.J. Hilbers, Timo Honkela, E. Oja and A. Lendasse, Adaptive Ensemble Models of Extreme Learning Machines for Time Series Prediction, ICANN'09, Lecture Notes in Computer Science, pp. 305-314.

[10] F. Corona and E. Liitiäinen and A. Lendasse and L. Sassu and S. Melis and R. Baratti, A SOM-based approach to estimating product properties from spectroscopic measurements, Neurocomputing, 2008, pp. 71-79.

[11] T. Kärnä, F. Corona and A. Lendasse, Gaussian basis functions for chemometrics, Journal of Chemometrics, 2008, pp. 701–707.

[12] F. Corona, S.P. Reinikainen, K. Aaljoki, A. Perkkio, E. Liitiäinen, R. Baratti, A. Lendasse and O. Simula, Wavelength selection using the measure of topological relevance on the Self-Organizing Map, Journal of Chemometrics, 2008, pp. 610-620.

[13] F. Corona , M. Mulas, R. Baratti and J. A. Romagnoli, On the topological analysis of industrial process data using the SOM, Elsevier, Computer Aided Chemical Engineering: Proceedings of PSE 2009 International Symposium on Process Systems Engineering, Salvador Bahia (Brazil), August 16-20 2009, pp. 1173-1178.

[14] F. Corona, E. Liitiäinen, A. Lendasse, R. Baratti and L. Sassu, Delaunay tessellation and topological regression: An application to estimating product properties. , Elsevier, Computer Aided Chemical Engineering: Proceedings of PSE 2009 International

Symposium on Process Systems Engineering, Salvador Bahia (Brazil), August 16-20 2009, pp. 1179-1184.

[15] Y. Miche, P. Bas, A. Lendasse, C. Jutten and O. Simula, A Feature Selection Methodology for Steganalysis, Traitement du Signal, May 2009, pp. 13-30.

[16] Y. Miche and P. Bas and A. Lendasse and C. Jutten and O. Simula, Reliable Steganalysis Using a Minimum Set of Samples and Features, EURASIP Journal on Information Security, March 2009, 13 pages.

[17] Q. Yu, A. Sorjamaa, Y. Miche and E. Séverin, A methodology for time series prediction in Finance, ESTSP'08, September 17-19, 2008, pp. 285-293.

[18] Q. Yu, A. Sorjamaa, Y. Miche, A. Lendasse, A. Guillén, E. Séverin and F. Mateo, Optimal Pruned K-Nearest Neighbors: OP-KNN - Application to Financial Modeling , Eighth International Conference on Hybrid Intelligent Systems, September 2008, pp. 764-769.

[19] Q. Yu, A. Sorjamaa, Y. Miche, E. Séverin and A. Lendasse, OP-KNN for Financial regression problems, Mashs 08, Computational Methods for Modelling and learning in Social and Human Sciences, Creteil France, June 5-6, 2008.

Individual projects

A. On the quantization error in SOM vs. VQ: A critical and systematic study

Teuvo Kohonen, Ilari T. Nieminen, and Timo Honkela

The self-organizing map (SOM) is related to the classical vector quantization (VQ). Like in the VQ, the SOM represents a distribution of input data vectors using a finite set of models. In both methods, the quantization error (QE) of an input vector can be expressed, e.g., as the Euclidean norm of the difference of the input vector and the best-matching model.

Some attempts have been made to compare the quantization errors in the SOM vs. the same errors in VQ. It is usually taken as self-evident that if, for instance, the models or "codebook vectors" are optimized in the VQ so that the sum of the squared QEs is minimized for given training vectors, it will be impossible to find any other set of models that produces a smaller *rms QE* (square root of the mean square of QE over independent test data). Thus the rms QE also in the SOM is supposed to be larger. Therefore it has come as a surprise that *the rms QE of the SOM may sometimes be smaller than that of the VQ* (cf., e.g., [1] and [6]).

We have found out that *this effect depends most strongly on the ratio of the number of training vectors and the number of model vectors.* If this ratio is *small, on the order of small integers*, the rms QE of the SOM is usually smaller than that of the VQ. However, *the training vectors must also have a significant local variance in sufficiently many dimensions.*

Artificial data

The first experiment was carried out with artificially generated, normally distributed random data with zero mean and identity covariance, and a 10x14 SOM. The ratio of the QEs in the SOM and the VQ with 140 codebook vectors is displayed as a function of the number of training vectors per model. Fig. I.4 represents the results. With the input dimensionality 10 the rms QE of the VQ was always smaller. For the dimensionalities 20 and 50 , the "break even" point (where the rms QEs of the SOM and the VQ are equal) occurred at the argument value 12.2 Below this point, the rms QE was always smaller in the SOM. The lower limit of the input dimensionalities for this effect to occur seems to exist between 10 and 20.

ISOLET data

The input vectors in this experiment consisted of 617 acoustic features extracted from spoken letters of the English alphabet [2]. In order to achieve a sufficient statistical accuracy, the *repeated holdout validation* was used. In it, a number of training samples is picked up at random from the available data set, while the rest of it is set aside for testing.

In the ISOLET experiment we had 7797 input vectors available. For instance, with the SOM array size 10x14 = 140 and the argument value 50, we selected M = 140x50 = 7000 samples at random from the basic data set for the construction of one SOM/VQ pair, while the rest was set aside for testing. At lower argument values, less data are needed for the construction of a SOM/VQ pair, whereupon more data can be reserved for testing. This random selection of the training vectors was repeated 20 times for every argument value, and a new SOM/VQ pair was constructed every time. The averages over the repeated evaluations of the rms QEs were then formed for every argument value.

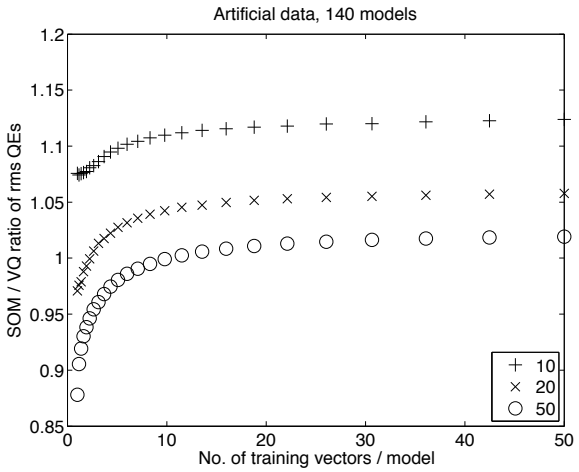


Figure I.4: Ratio of the rms QEs in the SOM and the VQ for the artificially generated random-data set, as a function of the number of training vectors per model, and for the dimensionalities 10, 20, and 50, respectively. The number of models was 140.

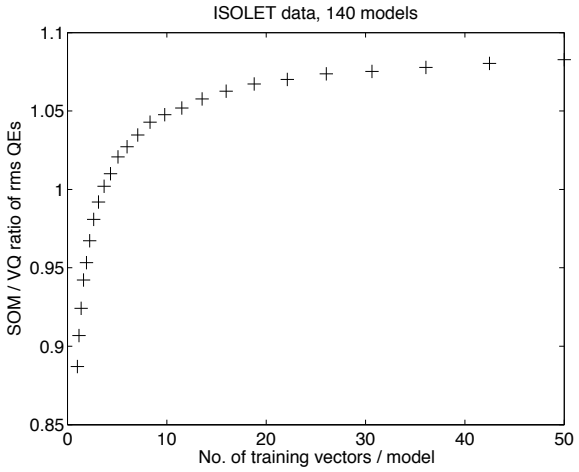


Figure I.5: Ratio of the rms QEs in the SOM and the VQ for the ISOLET data set, as a function of the number of training vectors per model and for 140 models.

In Fig. I.5 we display the ratio of the rms QEs in the SOM and in the VQ as a function of the number of training vectors per model. It can be seen that the "break even" point (at which the rms QEs in the SOM and the VQ are equal) is about 3.7. It is to be noted that the "effective dimensionality" or "fractal dimension" of real data is always much less than the true input dimensionality.

Reuters data

Our third experiment was based on the text corpus collected by the Reuters Corp. No original documents were available to us, but Lewis et al. [4] have prepared a test data set on the basis of this corpus for benchmarking purposes, preprocessing the textual data, removing the stop words, and reducing the words into their stems. The input vectors, with the true dimensionality of 233, were formed as weighted word histograms.

The general arrangement of this experiment was similar to that with the ISOLET data.

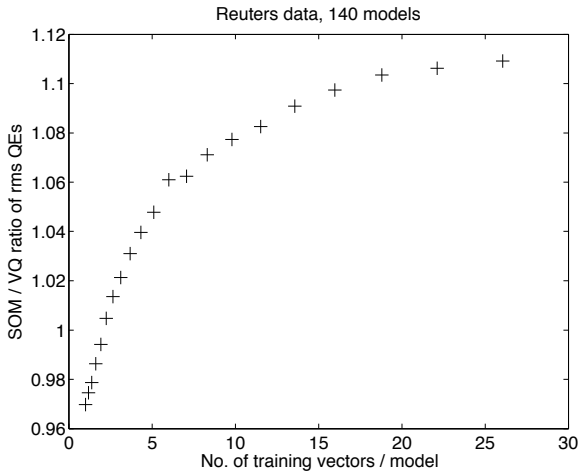


Figure I.6: Ratio of the rms QEs in the SOM and the VQ for the Reuters data set, as a function of the number of training vectors per model and for 140 models.

This time we had only 4000 input samples (documents) available. The averaged results are shown in Fig. I.6 for the SOM array size 10x14. The "break even" point was about 2.45.

Discussion

An explanation of the observed effect seems to ensue from statistics. Each model vector in the VQ is determined as the average of those training vectors that are mapped into the same Voronoi domain as the model vector. On the contrary, each model vector of the SOM is determined as a weighted average of all of those training vectors that are mapped into the "topological" neighborhood around the corresponding model. The number of training vectors mapped into the neighborhood of a SOM model is generally much larger than that mapped into a Voronoi domain around a model in the VQ. Since the SOM model vectors are then determined with a significantly higher statistical accuracy, the Voronoi domains of the SOM are significantly more regular, and the resulting rms QE may then be smaller than in the VQ. For a more detailed discussion, see [3].

References

[1] O. Bação, V. Lobo, and M. Painho, "Self-organizing maps as substitutes for k-means clustering," in *Computational Science - ICCS 2005, Lecture Notes in Computational Science,* Berlin, Heidelberg, Germany: Springer-Verlag, 2005, pp. 476-483.

[2] R. A. Cole, Y. Muthusamy, and M. A. Fanty, *The ISOLET Spoken Letter Database,* Technical Report 90-004, Computer Science Department, Oregon Graduate Institute, 1994.

[3] T. Kohonen, I. T. Nieminen, and T. Honkela, "On the Quantization Error in SOM vs. VQ:A Critical and Systematic Study," in *Advances in Self-Organizing Maps, Lecture Notes in Computational Science LNCS-5629,* Berlin, Heidelberg, Germany: Springer-Verlag, 2009, pp. 133-144.

[4] D.D. Lewis, Y. Yang, T.G. Rose, and T. Li, "RCV1: A new benchmark collection for text categorization research," *J. Machine Learning Research*. Vol. 5, pp. 361-397, 2004.

[5] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing,* Cambridge, MA: MIT Press, 1999.

[6] J. D. McAuliffe, L. E. Atlas, and C. Rivera, "A comparison of the LBG algorithm and Kohonen neural network paradigm for image vector quantization," in *Proc. ICASSP-90, Acoustics, Speech and Signal Processing,* Vol. IV, Piscataway, N.J.: IEEE Service Center, 1990, pp. 2293-2296.

Publications of the Adaptive Informatics Research Centre

[1] A. Ajanki, M. Billinghurst, M. Kandemir, S. Kaski, M. Koskela, M. Kurimo, J. Laaksonen, K. Puolamäki, and T. Tossavainen. Ubiquitous contextual information access with proactive retrieval and augmentation. Technical Report TKK-ICS-R27, Dec. 2009.

[2] A. Ajanki, D. R. Hardoon, S. Kaski, K. Puolamäki, and J. Shawe-Taylor. Can eyes reveal interest?—Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 19:307–339, 2009.

[3] T. Alhonnoro and M. Sirola. Feature selection on process fault detection and visualization. In *Proceedings of the 17th European Symposium On Artificial Neural Networks ESANN'08*, April 2008.

[4] M. Almeida and R. Vigário. Source separation of phase-locked subspaces. In *Proc. 8th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA'2009)*, pages 203–210, Paraty, Brazil, 2009.

[5] E. Ar'soy, M. Kurimo, M. Saraçlar, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and H. Sak. Statistical language modeling for automatic speech recognition of agglutinative languages. In *Speech Recognition*, pages 193–204, Vienna, Austria, 2008. I-Tech.

[6] F. Benachenhou, P. Jern, M. Oja, G. Sperber, V. Blikstad, P. Somervuo, S. Kaski, and J. Blomberg. Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data. *PLoS ONE*, 4(4):e5179, 2009.

[7] M. Blachnik and J. Laaksonen. Image classification by histogram features created with learning vector quantization. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'08)*, pages 827–836, Sept. 2008.

[8] G. J. Brown and K. J. Palomäki. A reverberation-robust automatic speech recognition system based on temporal masking. In *Acoustics 2008, J. Acoust. Soc. Am. 123*, page 2978, Paris, France, July 2008. abstract only publication.

[9] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25:i145–i153, 2009. (ISMB/ECCB 2009).

[10] J. Caldas and S. Kaski. Bayesian biclustering with the plaid model. In J. Príncipe, D. Erdogmus, and T. Adali, editors, *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing XVIII*, pages 291–296. IEEE, 2008.

[11] F. Corona, E. Liitiäinen, A. Lendasse, R. Baratti, and L. Sassu. Delaunay tessellation and topological regression: An application to estimating product properties. In E. B. Rita de Brito Alves, Claudio Oller do Nascimento, editor, *Computer Aided Chemical Engineering: Proceedings of PSE 2009 International Symposium on Process Systems Engineering, Salvador Bahia (Brazil)*, volume 27 of *Computer Aided Chemical Engineering*, pages 1179–1184. Elsevier, August 16-20 2009.

[12] F. Corona, E. Liitiäinen, A. Lendasse, L. Sassu, S. Melis, and R. Baratti. A SOM-based approach to estimating product properties from spectroscopic measurements. *Neurocomputing*, 73(1–3):71–79, December 2009.

[13] F. Corona, M. Mulas, R. Baratti, and J. Romagnoli. Data analysis and inference for an industrial deethanizer. In S. Pierucci, editor, *Chemical Engineering Transactions: Proceedings of ICHEAP9 International Conference on Chemical and Process Engineering, Rome (Italy)*, volume 17 of *Chemical Engineering Transactions*, pages 1197–1202. AIDIC, May 10-13 2009.

[14] F. Corona, M. Mulas, R. Baratti, and J. Romagnoli. Data derived analysis and inference for an industrial deethanizer. In *Proceedings of IFAC/ADCHEM 2009 International Symposium on Advanced Control of Chemical Processes, Istanbul (Turkey)*, pages 717–723, July 12-15 2009.

[15] F. Corona, M. Mulas, R. Baratti, and J. Romagnoli. On the topological analysis of industrial process data using the SOM. In E. B. Rita de Brito Alves, Claudio Oller do Nascimento, editor, *Computer Aided Chemical Engineering: Proceedings of PSE 2009 International Symposium on Process Systems Engineering, Salvador Bahia (Brazil)*, volume 27 of *Computer Aided Chemical Engineering*, pages 1173–1178. Elsevier, August 16-20 2009.

[16] F. Corona, S.-P. Reinikainen, K. Aaljoki, A. Perkkiö, E. Liitiäinen, R. Baratti, A. Lendasse, and O. Simula. Wavelength selection using the measure of topological relevance on the self-organizing map. *Journal of Chemometrics*, 22(11–12):610–620, November-December 2008.

[17] M. Creutz, S. Virpioja, and A. Kovaleva. Web augmentation of language models for continuous speech recognition of SMS text messages. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 157–165, Athens, Greece, March 2009. Association for Computational Linguistics.

[18] A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73–76, Boulder, USA, June 2009. Association for Computational Linguistics.

[19] T. Deselaers, A. Hanbury, V. Viitaniemi, J. D. R. Farquhar, M. Brendel, B. Daróczy, H. J. Escalante Balderas, T. Gevers, C. A. H. Gracidas, S. C. H. Hoi, J. Laaksonen, M. Li, H. M. Marin Castro, H. Ney, X. Rui, N. Sebe, J. Stöttinger, and L. Wu.

Overview of the ImageCLEF 2007 object retrieval task. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *Lecture Notes in Computer Science*, pages 445–471, Budapest, Hungary, 2008. Springer.

[20] E. Eirola. Variable selection with the delta test in theory and practice. Master's thesis, Helsinki University of Technology, Espoo, Finland, November 2009.

[21] E. Eirola, E. Liitiäinen, A. Lendasse, F. Corona, and M. Verleysen. Using the delta test for variable selection. In M. Verleysen, editor, *Proceedings of ESANN 2008, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 25–30. d-side publ. (Evere, Belgium), April 23-25 2008.

[22] D. Ellis, M. Creutz, T. Honkela, and M. Kurimo. Speech to speech machine translation: Biblical chatter from Finnish to English. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 123–130, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing.

[23] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, 24(16):i70–i75, 2008. In Proceedings of ECCB 2008.

[24] R. Girdziušas. *Stability and Inference in Discrete Diffusion Scale-Spaces*. D25, Feb. 2008.

[25] N. Gonçalves, J. Nikkilä, and R. Vigário. Partial clustering for tissue segmentation in MRI. In M. K. et al., editor, *Proceedings of ICONIP 2008*, volume LNCS 5507, pages 559–566, 2009.

[26] N. Gonçalves and R. Vigário. Semi-automatic approach for brain tissue segmentation using MRI. In *1st INCF Congress of Neuroinformatics: Databasing and Modeling the Brain (Neuroinformatics 2008)*, page 106, Stockholm, Sweden, September 2008. Poster.

[27] A. Guillén, L. Herrera, G. Rubio, A. Lendasse, H. Pomares, and I. Rojas. Instance or prototype selection for function approximation using mutual information. In A. Lendasse, editor, *ESTSP'08 Proceedings*, pages 67–75, September 2008.

[28] A. Guillén, A. Sorjamaa, Y. Miche, A. Lendasse, and I. Rojas. Efficient parallel feature selection for steganography problems. In J. Cabestany, F. S. A. Prieto, and J. Corchado, editors, *LNCS - Bio-Inspired Systems: Computational and Ambient Intelligence – IWANN 2009, Part I*, volume 5517/2009 of *Lecture Notes in Computer Science*, page 1224–1231. IWANN, Springer Berlin / Heidelberg, June 2009.

[29] A. Guillén, A. Sorjamaa, G. Rubio, A. Lendasse, and I. Rojas. Mutual information based initialization of forward-backward search for feature selection in regression problems. In C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, editors, *LNCS - Artificial Neural Networks - ICANN 2009 – Part I*, volume 5768 of *Lecture Notes in Computer Science*, pages 1–9. ICANN, Springer Berlin / Heidelberg, September 2009.

[30] A. Guillén, D. Sovilj, F. Mateo, I. Rojas, and A. Lendasse. Minimizing the delta test for variable selection in regression problems. *International Journal of High Performance Systems Architecture*, 1(4):269–281, 2008.

[31] A. Guillén, D. Sovilj, F. Mateo, I. Rojas, and A. Lendasse. New methodologies based on delta test for variable selection in regression problems. In *Workshop on Parallel Architectures and Bioinspired Algorithms*, Toronto, Canada, October 25-29 2008.

[32] M. Harva. *Algorithms for Approximate Bayesian Inference with Applications to Astronomical Data Analysis*. TKK-ICS-D3, May 2008.

[33] T. Hirsimäki and M. Kurimo. Analysing recognition errors in unlimited-vocabulary speech recognition. In *Proceedings of the 2009 Annual Conference of the North american Chapter of the association for Computational Linguistics, NAACL 2009*, Boulder, CO, May 31 - June 5 2009.

[34] A. Honkela. Identifying targets of transcriptionally regulated transcription factors using dynamical models. In *Mathematical and Statistical Aspects of Molecular Biology: 19th Annual MASAMB Workshop*, Imperial College London, UK, 2009.

[35] A. Honkela, M. Harva, T. Raiko, and J. Karhunen. Variational inference and learning for continuous-time nonlinear state-space models. In *Proc. of PASCAL 2008 Workshop on Approximate Inference in Stochastic Processes and Dynamical Systems*, Cumberland Lodge, UK, May 2008.

[36] A. Honkela, N. D. Lawrence, and M. Rattray. Identifying targets of transcriptionally regulated transcription factors using dynamical models. In *17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB)*, Stockholm, Sweden, 2009.

[37] A. Honkela, J. Seppä, and E. Alhoniemi. Agglomerative independent variable group analysis. *Neurocomputing*, 71(7–9):1311–1320, 2008.

[38] A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In *Proceedings of the 14th International Conference on Neural Information Processing (ICONIP 2007)*, volume 4985 of *Lecture Notes in Computer Science*, pages 305–314, Kitakyushu, Japan, 2008. Springer-Verlag, Berlin.

[39] T. Honkela. Conceptual autonomy of agents. In *ICAART 2009 - Proceedings of the International Conference on Agents and Artificial Intelligence*, page 9. INSTICC Press, 2009.

[40] T. Honkela, N. Janasik, K. Lagus, T. Lindh-Knuutila, M. Pantzar, and J. Raitio. Modeling communities of experts. Technical Report TKK-ICS-R24, Dec. 2009.

[41] T. Honkela, N. Janasik, K. Lagus, T. Lindh-Knuutila, M. Pantzar, and J. Raitio. Modeling communities of experts – conceptual grounding of expertise. Technical Report TKK-ICS-R24, Helsinki University of Technology, 2009.

[42] T. Honkela, V. Könönen, T. Lindh-Knuutila, and M.-S. Paukkeri. Simulating processes of concept formation and communication. *Journal of Economic Methodology*, 15(3):245–259, 2008.

[43] T. Honkela, M.-S. Paukkeri, M. Pöllä, and O. Simula, editors. *Proceedings of AKRR'08, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Helsinki University of Technology, Espoo, Finland, September 2008.

[44] T. Honkela and M. Pöllä. Concept mining with Self-Organizing Maps for the Semantic Web. In *Proceedings of WSOM'09*, pages 98–106. Springer, 2009.

[45] T. Honkela, S. Virpioja, and J. Väyrynen. Adaptive translation: Finding interlingual mappings using self-organizing maps. In V. Kurková, R. Neruda, and J. Koutnik, editors, *Proceedings of ICANN'08*, volume 5163 of *Lecture Notes in Computer Science*, pages 603–612. Springer, 2008.

[46] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19:261–276, 2009.

[47] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski. Multi-way, multi-view learning. In *NIPS 2009 workshop on Learning from Multiple Sources with Applications to Robotics*, 2009. Extended Abstract.

[48] A. Ilin and A. Kaplan. Bayesian PCA for reconstruction of historical sea surface temperatures. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2009)*, pages 1322–1327, Atlanta, USA, June 2009.

[49] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. Technical Report TKK-ICS-R6, Helsinki University of Technology, TKK reports in information and computer science, Espoo, Finland, 2008.

[50] N. Janasik, T. Honkela, and H. Bruun. Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods*, 12(3):436–460, 2009.

[51] H. Kallasjoki, K. Palomäki, C. Magi, P. Alku, and M. Kurimo. Noise robust lvcsr feature extraction based on stabilized weighted linear prediction. In *Proceedings of the 13th International Conference Speech and Computer, SPECOM 2009*, pages 221–225, St. Petersburg, Russia, June 21-25 2009.

[52] T. Kärnä, F. Corona, and A. Lendasse. Gaussian basis functions for chemometrics. *Journal of Chemometrics*, 22(11–12):701–707, November-December 2008.

[53] E. Karp, L. Parkkonen, and R. Vigário. Denoising single trial event related magnetoencephalographic recordings. In T. A. et al., editor, *Proceedings of 8th International Conference on Independent Component Analysis and Signal Separation (ICA 2009)*, volume LNCS 5441, pages 427–434, Paraty, Brazil, 2009. Springer-Verlag.

[54] A. Klami. *Modeling of Mutual Dependencies*. TKK-ICS-D6, Sept. 2008.

[55] A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72:39–46, 2008.

[56] A. Klami, C. Saunders, T. de Campos, and S. Kaski. Can relevance of images be inferred from eye movements? In *MIR '08: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 134–140. ACM, New York, NY, USA, 2008.

[57] O. Kohonen, S. Virpioja, and M. Klami. Allomorfessor: Towards unsupervised morpheme analysis. In *In Working Notes of the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[58] O. Kohonen, S. Virpioja, and M. Klami. Allomorfessor: Towards unsupervised morpheme analysis. In *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008 Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, pages 975–982. Springer, 2009.

[59] O. Kohonen, S. Virpioja, and K. Lagus. A constructionist approach to grammar inference. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, Whistler, Canada, December 2009. Extended abstract.

[60] T. Kohonen, I. T. Nieminen, and T. Honkela. On the quantization error in SOM vs. VQ: A critical and systematic study. In J. C. Príncipe and R. Miikkulainen, editors, *Proceedings of WSOM'09*, volume 5629 of *Lecture Notes in Computer Science*, pages 133–144. Springer, 2009.

[61] M. Koskela and J. Laaksonen. Specification of information interfaces in PinView. Technical Report TKK-ICS-R12, Nov. 2008.

[62] M. Koskela, J. Laaksonen, T. Jantunen, R. Takkinen, P. Rainò, and A. Raike. Content-based video analysis and access for finnish sign language – a multidisciplinary research project. In *Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages at 6th Language Resources and Evaluation Conference (LREC 2008)*, pages 101–104, Marrakech, Morocco, May-June 2008.

[63] M. Koskela, M. Sjöberg, and J. Laaksonen. Improving automatic video retrieval with semantic concept detection. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 480–489, Oslo, Norway, 2009. Springer Verlag.

[64] M. Koskela, M. Sjöberg, V. Viitaniemi, and J. Laaksonen. PicSOM experiments in TRECVID 2008. In *Proceedings of the TRECVID 2008 Workshop*, Gaithersburg, MD, USA, November 2008. Available online at `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`.

[65] L. Kozma, A. Ilin, and T. Raiko. Binary principal component analysis in the Netflix collaborative filtering task. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Grenoble, France, September 2009.

[66] L. Kozma, A. Klami, and S. Kaski. GaZIR: Gaze-based zooming interface for image retrieval. In *Proc. ICMI-MLMI 2009, The Eleventh International Conference on Multimodal Interfaces and The Sixth Workshop on Machine Learning for Multimodal Interaction*, pages 305–312, New York, NY, USA, 2009. ACM.

[67] E. Kurimo, L. Lepistö, J. Nikkanen, J. Grén, I. Kunttu, and J. Laaksonen. The effect of motion blur and signal noise on image quality in low light imaging. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 81–90, Oslo, Norway, 2009. Springer Verlag.

[68] M. Kurimo. Puheentunnistus. *Puhe ja Kieli*, 28(2):73–83, 2008.

[69] M. Kurimo. Puheentunnistus. In *Puhuva ihminen*, pages 336–343. Otava, 2009.

[70] M. Kurimo, M. Creutz, and V. Turunen. Morpho Challenge evaluation by information retrieval. In *Advances in Multilingual and MultiModal Information Retrieval, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Lecture Notes in Computer Science. Springer, 2009.

[71] M. Kurimo, M. Creutz, and M. Varjokallio. Morpho Challenge evaluation using a linguistic Gold Standard. In *Advances in Multilingual and MultiModal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, Lecture Notes in Computer Science , Vol. 5152, pages 864–873. Springer, 2008.

[72] M. Kurimo, T. Hirsimäki, V. Turunen, S. Virpioja, and N. Raatikainen. Unsupervised decomposition of words for speech recognition and retrieval. In *Proceedings of the 13th International Conference Speech and Computer, SPECOM 2009*, pages 23–28, St. Petersburg, Russia, June 21-25 2009.

[73] M. Kurimo and V. Turunen. Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[74] M. Kurimo, V. Turunen, and M. Varjokallio. Overview of Morpho Challenge 2008. In *Advances in Multilingual and MultiModal Information Retrieval, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Lecture Notes in Computer Science. Springer, 2009.

[75] M. Kurimo and M. Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic Gold Standard – Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[76] M. Kurimo, S. Virpioja, V. Turunen, and T. Hirsimäki. Morpho Challenge - evaluation of algorithms for unsupervised learning of morphology in various tasks and languages. In *Proceedings of the 2009 Annual Conference of the North american Chapter of the association for Computational Linguistics, NAACL 2009*, Boulder, CO, May 31 - June 5 2009.

[77] M. Kuusela. Algorithms for variational learning of mixture of Gaussians, 2009. Bachelor's thesis.

[78] M. Kuusela, T. Raiko, A. Honkela, and J. Karhunen. A gradient-based algorithm competitive with variational Bayesian EM for mixture of Gaussians. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2009*, pages 1688–1695, Atlanta, Georgia, June 2009.

[79] J. Laaksonen. Definition of enriched relevance feedback in PicSOM. Technical Report TKK-ICS-R13, Nov. 2008.

[80] K. Lagus, M. Creutz, S. Virpioja, and O. Kohonen. Morpheme segmentation by optimizing two-part MDL codes. In *2009 Workshop on Information Theoretic Methods in Science and Engineering (WITMSE)*, Tampere, Finland, August 2009. Extended abstract.

[81] K. Lagus, O. Kohonen, and S. Virpioja. Towards unsupervised learning of constructions from text. In M. Sahlgren and O. Knutsson, editors, *Proceedings of the Workshop on Extracting and Using Constructions in NLP of 17th Nordic Conference on Computational Linguistics, NODALIDA*, May 2009. SICS Technical Report T2009:10.

[82] L. Lahti. RPA: probe reliability and differential gene expression analysis. BioConductor 2.5, October 2009. Computer program.

[83] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski. Dependency detection with similarity constraints. In *Proc. MLSP 2009, IEEE International Workshop on Machine Learning for Signal Processing*, pages 89–94. IEEE, 2009.

[84] J. Laiho, A. Höglund, K. Raivio, J. Henriksson, K. Hätönen, A. Hämäläinen, and O. Simula. Method for configuring a network by defining clusters, 2008. EP1374486.

[85] G. Leen, D. R. Hardoon, and S. Kaski. Automatic choice of control measurements. In Z.-H. Zhou and T. Washio, editors, *Advances in Machine Learning (Proc. ACML'09, The 1st Asian Conference on Machine Learning)*, volume 5828 of *Lecture Notes in Computer Science*, pages 206–219. Springer, 2009.

[86] P. Lehtimäki. *Data Analysis Methods for Cellular Network Performance Optimization*. TKK-ICS-D1, Apr. 2008.

[87] P. Lehtimäki and K. Raivio. Combining measurement data and Erlang-B formula for blocking prediction in GSM networks. In *Proceedings of The 10th Scandinavian Conference on Artificial Intelligence (SCAI)*, pages 98–105, Stockholm, Sweden, May 26 - 28 2008.

[88] A. Lendasse, editor. *ESTSP 2008: Proceedings*. Multiprint Oy / Otamedia, 2008. ISBN: 978-951-22-9544-9.

[89] A. Lendasse and F. Corona. Linear projection based on noise variance estimation: Application to spectral data. In M. Verleysen, editor, *Proceedings of ESANN 2008, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 457–462. d-side publ. (Evere, Belgium), April 23-25 2008.

[90] E. Liitiäinen, F. Corona, and A. Lendasse. A boundary corrected expansion of the moments of nearest neighbor distributions. Technical Report TKK-ICS-R9, Helsinki University of Technology, October 18 2008.

[91] E. Liitiäinen, F. Corona, and A. Lendasse. On non-parametric residual variance estimation. *Neural Processing Letters*, 28(3):155–167, December 2008.

[92] E. Liitiäinen, A. Lendasse, and F. Corona. Bounds on the mean power-weighted nearest neighbour distance. *Proceedings of the Royal Society A*, 464(2097):2293–2301, September 2008.

[93] E. Liitiäinen, A. Lendasse, and F. Corona. On the statistical estimation of rényi entropies. In *Proceedings of IEEE/MLSP 2009 International Workshop on Machine Learning for Signal Processing, Grenoble (France)*, September 2-4 2009.

[94] E. Liitiäinen, M. Verleysen, F. Corona, and A. Lendasse. Residual variance estimation in machine learning. *Neurocomputing*, 72(16–18):3692–3703, October 2009.

[95] T. Lindh-Knuutila, J. Raitio, and T. Honkela. Combining self-organized and Bayesian models of concept formation. In J. Mayor, N. Ruh, and K. Plunkett, editors, *Connectionist Models of Behaviour and Cognition II Proceedings of the Eleventh Neural Computation and Psychology Workshop*, number 18 in Progress in Neural Processing, pages 193–204. World Scientific, April 2009.

[96] J. Luttinen and A. Ilin. Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Advances in Neural Information Processing Systems 22*, pages 1177–1185. MIT Press, Cambridge, MA, USA, 2009.

[97] J. Luttinen, A. Ilin, and J. Karhunen. Bayesian robust PCA for incomplete data. In *Proceedings of the 8th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2009)*, pages 66–73, Paraty, Brazil, March 2009.

[98] J. Luttinen, A. Ilin, and T. Raiko. Transformations for variational factor analysis to speed up learning. In *Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN 2009)*, pages 77–82, Bruges, Belgium, April 2009.

[99] E. Malmi, J. Raitio, and T. Honkela. Modeling practice diffusion with an agent-based social simulation framework. In *Proceedings of the 6th European Social Simulation Association Conference, ESSA 2009*, page 53, Guildford, U.K., September 2009. Extended abstract.

[100] F. Mateo and A. Lendasse. A variable selection approach based on the delta test for extreme learning machine models. In M. Verleysen, editor, *Proceedings of the European Symposium on Time Series Prediction*, pages 57–66. d-side publ. (Evere, Belgium), September 2008.

[101] F. Mateo, D. Sovilj, R. Gadea, and A. Lendasse. RCGA-S/RCGA-SP methods to minimize the delta test for regression tasks. In *IWANN 2009*, volume 5517 of *Lecture Notes in Computer Science*, pages 359–366, Salamanca, Spain, June 10-12 2009. Springer.

[102] N. Matsuda, J. Laaksonen, F. Tajima, and H. Sato. Classification of fundus images for diagnosing glaucoma by self-organizing map and learning vector quantization. In *Proceedings of 15th International Conference on Neural Information Processing (ICONIP 2008)*, volume 5507 of *Lecture Notes in Computer Science*, pages 703–710, Auckland, New Zealand, 2009. Springer Verlag.

[103] N. Matsuda, H. Tokutaka, J. Laaksonen, F. Tajima, N. Miyatake, and H. Sato. Spherical SOM and its application to fundus image analysis. *Journal of Biomedical Fuzzy Systems Association*, 11(1):29–34, June 2009. In Japanese.

[104] P. Merlin, A. Sorjamaa, B. Maillet, and A. Lendasse. X-SOM and l-SOM: a nested approach for missing value imputation. In M. Verleysen, editor, *ESANN2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning*, ESANN Proceedings, pages 83–88, Brugge, Belgium, April 2009. ESANN, d-side publications.

[105] Y. Miche, P. Bas, C. Jutten, O. Simula, and A. Lendasse. A methodology for building regression models using extreme learning machine: OP-ELM. In M. Verleysen, editor, *ESANN 2008, European Symposium on Artificial Neural Networks, Bruges, Belgium*, pages 247–252. d-side publ. (Evere, Belgium), April 23-25 2008.

[106] Y. Miche, P. Bas, A. Lendasse, C. Jutten, and O. Simula. A feature selection methodology for steganalysis. *Traitement du Signal*, 26(1):13–30, May 2009.

[107] Y. Miche, P. Bas, A. Lendasse, C. Jutten, and O. Simula. Reliable steganalysis using a minimum set of samples and features. *EURASIP Journal on Information Security*, 2009(1):1–13 (Article ID 901381), March 2009. http://www.hindawi.com/journals/is/2009/901381.html.

[108] Y. Miche and A. Lendasse. A faster model selection criterion for OP-ELM and OP-KNN: Hannan-quinn criterion. In M. Verleysen, editor, *ESANN'09: European Symposium on Artificial Neural Networks*, pages 177–182. d-side publications, April 22-24 2009.

[109] Y. Miche, A. Sorjamaa, and A. Lendasse. OP-ELM: Theory, experiments and a toolbox. In R. N. Vera Kurková and J. Koutník, editors, *LNCS - Artificial Neural Networks - ICANN 2008 - Part I*, volume 5163/2008 of *Lecture Notes in Computer Science*, pages 145–154. Springer Berlin / Heidelberg, September 2008.

[110] M. Molinier, V. Viitaniemi, M. Koskela, J. Laaksonen, Y. Rauste, A. Lönnqvist, and T. Häme. Improving content-based target and change detection in Alos Palsar images with efficient feature selection. In *Proceedings of ESA-EUSC 2008: Image Information Mining*. ESA, March 2008.

[111] F. Montesino-Pouzols, A. Barriga, D. R. Lopez, and S. Sánchez-Solano. Encyclopedia of networked and virtual organizations. volume II, pages 1210–1215, Hershey, New York, USA, Mar. 2008. Information Science Reference (an imprint of IGI Global). ISBN: 978-1-59904-885-7.

[112] F. Montesino-Pouzols, A. Barriga, D. R. Lopez, and S. Sánchez-Solano. Encyclopedia of networked and virtual organizations. volume II, pages 1216–1222, Hershey, New York, USA, Mar. 2008. Information Science Reference (an imprint of IGI Global). ISBN: 978-1-59904-885-7.

[113] F. Montesino-Pouzols, A. Barriga, D. R. Lopez, and S. Sánchez-Solano. Linguistic Summarization of Network Traffic Flows. In *17th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE´2008), IEEE World Congress on Computational Intelligence*, pages 619–624, Hong Kong, China, June 2008.

[114] F. Montesino-Pouzols, D. R. Lopez, A. Barriga, and S. Sánchez-Solano. Encyclopedia of networked and virtual organizations. volume II, pages 1196–1203, Hershey, New York, USA, Mar. 2008. Information Science Reference (an imprint of IGI Global). ISBN: 978-1-59904-885-7.

[115] F. Montesino-Pouzols, D. R. Lopez, A. Barriga, and S. Sánchez-Solano. Encyclopedia of networked and virtual organizations. volume II, pages 1204–1209, Hershey, New York, USA, Mar. 2008. Information Science Reference (an imprint of IGI Global). ISBN: 978-1-59904-885-7.

[116] M. Multanen, K. Raivio, and P. Lehtimäki. Outlier detection in cellular network data exploration. In *Proceedings of the 22nd International Conference on Advanced Information Networking and Applications (AINA), 3rd International Workshop on Performance Analysis and Enhancement of Wireless Networks (PAEWN)*, pages 1323–1328, Okinawa, Japan, March 25 - 28 2008.

[117] P. Niemelä and T. Honkela. Analysis of parliamentary election results and socio-economic situation using self-organizing map. In *Proceedings of WSOM'09*, pages 209–218, 2009.

[118] I. T. Nieminen. Combining tag recommendations based on user history. In F. Eisterlehner, A. Hotho, and R. Jäschke, editors, *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, pages 229–234, Sept. 2009.

[119] J. Nikkilä, M. Sysi-Aho, A. Ermolov, T. Seppänen-Laakso, O. Simell, S. Kaski, and M. Orešič. Gender dependent progression of systemic metabolic states in early childhood. *Molecular Systems Biology*, 4:197, 2008.

[120] K. Nybo, J. Parkkinen, and S. Kaski. Graph visualization with latent variable models. Technical Report TKK-ICS-R20, Sept. 2009.

[121] J. Pajarinen, J. Peltonen, M. A. Uusitalo, and A. Hottinen. Latent state models of primary user behavior for opportunistic spectrum access. In *Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Tokyo, Japan, Sept. 2009.

[122] K. J. Palomäki, U. Remes, and M. K. (Eds.). Studies on noise robust automatic speech recognition. Technical Report TKK-ICS-R19, Sept. 2009.

[123] K. J. Palomäki and G. J. Brown. A computational model of binaural speech intelligibility level difference. In *Acoustics 2008, J. Acoust. Soc. Am. 123*, page 3715, Paris, France, July 2008. abstract only publication.

[124] J. Parkkinen. Generative probabilistic models of biological and social network data. Master's thesis, Helsinki University of Technology, Department of Information and Computer Science, September 2009.

[125] J. Parkkinen, J. Sinkkonen, A. Gyenge, and S. Kaski. A block model suitable for sparse graphs. In *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009)*, Leuven, Belgium, July 2-4 2009. Extended Abstract.

[126] K. Pasupa, C. Saunders, S. Szedmak, A. Klami, S. Kaski, and S. Gunn. Learning to rank images from eye movements. In *IEEE International Workshop on Human-Computer Interaction (HCI2009), October 4, 2009, Kyoto, Japan*, pages 2009–2016, 2009.

[127] M.-S. Paukkeri and T. Kotro. Framework for analyzing and clustering short message database of ideas. In *Proceedings of the 9th International Conference on Knowledge Management and Knowledge Technologies (I-KNOW'09)*, pages 239–247, Graz, Austria, September 2009.

[128] M.-S. Paukkeri, I. T. Nieminen, M. Pöllä, and T. Honkela. A language-independent approach to keyphrase extraction and evaluation. In *Coling 2008: Companion volume: Posters*, pages 83–86, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[129] J. Peltonen. Visualization by linear projections as information retrieval. In J. Príncipe and R. Miikkulainen, editors, *Advances in Self-Organizing Maps (proceedings of WSOM 2009)*, pages 237–245, Berlin Heidelberg, 2009. Springer.

[130] J. Peltonen, H. Aidos, and S. Kaski. Supervised nonlinear dimensionality reduction by neighbor retrieval. In *Proceedings of ICASSP 2009, the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1809–1812. IEEE, 2009.

[131] J. Peltonen, M. A. Uusitalo, and J. Pajarinen. Nano-scale fault tolerant machine learning for cognitive radio. In J. C. Principe, D. Erdogmus, and T. Adali, editors, *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, pages 163–168, Cancún, Mexico, Oct. 2008.

[132] J. Peltonen, J. Venna, and S. Kaski. Visualizations for assessing convergence and mixing of Markov chain Monte Carlo simulations. *Computational Statistics and Data Analysis*, 53:4453–4470, 2009.

[133] J. Pohjalainen, H. Kallasjoki, P. Alku, K. Palomäki, and M. Kurimo. Weighted linear prediction for speech analysis in noisy conditions. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009*, pages 1315–1318, Brighton, UK, September 6-10 2009. ISCA.

[134] M. Pöllä and T. Honkela. Change detection of text documents using negative first-order statistics. In *Poceedings of AKRR'08, The Second International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 48–55, Porvoo, Finland, September 2008.

[135] M. Pöllä, T. Honkela, and T. Kohonen. Bibliography of self-organizing map (SOM) papers: 2002–2005 addendum. Technical Report TKK-ICS-R23, Dec. 2009.

[136] F. M. Pouzols. *Mining and Control of Network Traffic by Computational Intelligence*. Doctoral dissertation, University of Seville, Seville, Spain, May 2009.

[137] F. M. Pouzols and A. B. Barros. Regressive fuzzy inference models with clustering identification: Application to the ESTSP08 competition. In M. Verleysen, editor, *2nd European Symposium on Time Series Prediction (ESTSP08)*, pages 205–214, Porvoo, Finland, September 2008. d-side publ. (Evere, Belgium).

[138] F. M. Pouzols, A. Lendasse, and A. B. Barros. Fuzzy inference based autoregressors for time series prediction using nonparametric residual variance estimation. In *17th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'2008), IEEE World Congress on Computational Intelligence*, pages 613–618, Hong Kong, China, June 2008.

[139] F. M. Pouzols, A. Lendasse, and A. B. Barros. xftsp: a tool for time series prediction by means of fuzzy inference systems. In *4th IEEE International Conference on Intelligent Systems (IS08)*, volume 1, pages 2–2–2–7, Varna, Bulgaria, September 2008.

[140] P. Prentis, M. Sjöberg, M. Koskela, and J. Laaksonen. Image theft detection with self-organising maps. In *Proceedings of 19th International Conference on Artificial Neural Networks (ICANN 2009)*, volume 5768 of *Lecture Notes in Computer Science*, pages 495–504, Limassol, Cyprus, 2009. Springer Verlag. Available online at: `http://dx.doi.org/10.1007/978-3-642-04274-4\_52`.

[141] K. Puolamäki, A. Ajanki, and S. Kaski. Learning to learn implicit queries from gaze patterns. In A. McCallum and S. Roweis, editors, *Proceedings of ICML 2008,*

Twenty-Fifth International Conference on Machine Learning, pages 760–767, Madison, WI, 2008.

[142] K. Puolamäki and S. Kaski. Bayesian solutions to the label switching problem. In N. Adams, C. Robardet, A. Siebes, and J.-F. Boulicaut, editors, *Advances in Intelligent Data Analysis VIII, Proceedings of the 8th International Symposium on Intelligent Data Analysis, IDA 2009*, pages 381–392, Berlin, 2009. Springer.

[143] J. Pylkkönen. Investigations on discriminative training in large scale acoustic model estimation. In *Proceedings of Interspeech*, pages 220–223, 2009.

[144] T. Raiko, P. Haikonen, and J. V. editors, editors. *AI and Machine Consciousness, Proceedings of the 13th Finnish Artificial Intelligence Conference (STeP 2008)*. Finnish Artificial Intelligence Society, Espoo, Finland, August 2008.

[145] T. Raiko, A. Ilin, and J. Karhunen. Principal component analysis for sparse high-dimensional data. In *Proceedings of the 14th International Conference on Neural Information Processing (ICONIP 2007)*, pages 566–575, Kitakyushu, Japan, 2008.

[146] T. Raiko and J. Peltonen. Application of uct search to the connection games of hex, y, *star, and renkula! In *Proc. of the Finnish Artificial Intelligence Conference (STeP 2008)*, pages 89–93, Espoo, Finland, August 2008.

[147] T. Raiko, K. Puolamäki, J. Karhunen, J. Hollmén, A. Honkela, S. Kaski, H. Mannila, E. Oja, and O. Simula. Macadamia: Master's programme in machine learning and data mining. In *Teaching Machine Learning: Workshop on open problems and new directions*, Saint-Étienne, France, May 2008.

[148] T. Raiko and M. Tornio. Variational bayesian learning of nonlinear hidden state-space models for model predictive control. *Neurocomputing*, 72:3704–3712, October 2009.

[149] Y. Rauste, A. Lönnqvist, M. Molinier, H. Ahola, J. Praks, T. Tares, J. Laaksonen, K. Karila, and M. Karjalainen. NewSAR project, final report 1.7.2006–31.5.2008. Technical Report VTT-R-05074-08, VTT, June 2008.

[150] T. R. Raviv, K. Van Leemput, W. Wells, and P. Golland. Joint segmentation of image ensembles via latent atlases. In *Lecture Notes in Computer Science*, volume 5761, pages 272–280, 2009. Proceedings of MICCAI2009, September 20-14, 2009, London, UK.

[151] T. R. Raviv, B. Menze, K. Van Leemput, B. Stieltjes, M. Weber, N. Ayache, W. Wells, and P. Golland. Joint segmentation via patient-specific latent anatomy model. In *Proceedings of the MICCAI 2009 Workshop on Probabilistic Models for Medical Image Analysis (PMMIA 2009)*, pages 244–255, 2009. September 20, 2009, London, UK.

[152] U. Remes, K. J. Palomäki, and M. Kurimo. Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition. In *Proceedings of the 16th European Signal Processing Conference, EUSIPCO 2008*, Lausanne, Switzerland, August 25–29 2008.

[153] U. Remes, K. J. Palomäki, and M. Kurimo. Robust automatic speech recognition using acoustic model adaptation prior to missing feature reconstruction. In *Proceedings of the 17th European Signal Processing Conference, EUSIPCO 2009*, Glasgow, Scotland, UK, August 24–28 2009.

[154] U. Remes, K. J. Palomäki, and M. Kurimo. Speaker adaptation combined with missing data reconstruction. In *Acoustics 2008, J. Acoust. Soc. Am. 123*, page 3181, Paris, France, July 2008. abstract only publication.

[155] R. Ritala, E. Alhoniemi, T. Kauranne, K. Konkarikoski, A. Lendasse, and M. Sirola. Nonlinear temporal and spatial forecasting: modeling and uncertainty analysis (NoTeS) - MASIT20. In *MASI Programme 2005 - 2009 Yearbook 2008*, 2008. TEKES review 228/2008.

[156] S. Rogers, J. Sinkkonen, A. Klami, M. Girolami, and S. Kaski. Two-level infinite mixture for multi-domain data. In *Learning from Multiple Sources Workshop, 13 December 2008, Whistler Canada*, 2008. Proceedings at `http://web.mac.com/davidrh/LMSworkshop08/Schedule.html`.

[157] A.-M. Rusanen, O. Lappi, T. Honkela, and M. Nederström. Conceptual coherence in philosophy education - visualizing initial conceptions of philosophy students with self-organizing maps. In B. C. Love, K. McRae, and V. M. Sloutsky, editors, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages pp. 64–70, Austin, TX, 2008. Cognitive Science Society.

[158] M. Sabuncu, B. Yeo, K. Van Leemput, B. Fischl, and P. Golland. Nonparametric mixture models for supervised image parcellation. In *Proceedings of the MICCAI 2009 Workshop on Probabilistic Models for Medical Image Analysis (PMMIA 2009)*, pages 301–313, 2009. September 20, 2009, London, UK.

[159] M. Sabuncu, B. Yeo, K. Van Leemput, B. Fischl, and P. Golland. Supervised nonparametric image parcellation. In *Lecture Notes in Computer Science*, volume 5762, pages 1075–1083, 2009. Proceedings of MICCAI2009, September 20-14, 2009, London, UK.

[160] M. Sabuncu, B. Yeo, T. Vercauteren, K. Van Leemput, and P. Golland. Asymmetric image-template registration. In *Lecture Notes in Computer Science*, volume 5761, pages 565–573, 2009. Proceedings of MICCAI2009, September 20-14, 2009, London, UK.

[161] M. Sadeniemi, K. Kettunen, T. Lindh-Knuutila, and T. Honkela. Complexity of European Union languages: A comparative approach. *Journal of Quantitative Linguistics*, 15(2):185–211, 2008.

[162] J. Salojärvi. *Inferring Relevance from Eye Movements with Wrong Models*. TKK-ICS-D8, Nov. 2008.

[163] E. Savia, A. Klami, and S. Kaski. Fast dependent components for fMRI analysis. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1737–1740, 2009.

[164] E. Savia, K. Puolamäki, and S. Kaski. Latent grouping models for user preference prediction. *Machine Learning*, 74:75–109, 2009. Published online: 3 September 2008.

[165] E. Savia, K. Puolamäki, and S. Kaski. On two-way grouping by one-way topic models. Technical Report TKK-ICS-R15, May 2009.

[166] E. Savia, K. Puolamäki, and S. Kaski. Two-way grouping by one-way topic models. In N. A. et. al., editor, *Proceedings of IDA 2009, The 8th International Symposium*

on Intelligent Data Analysis, Lecture Notes in Computer Science, pages 178–189. Springer Berlin / Heidelberg, 2009.

[167] S. Savola, A. Klami, A. Tripathi, T. Niini, M. Serra, P. Picci, S. Kaski, D. Zambelli, K. Scotlandi, and S. Knuutila. Combined use of expression and CGH arrays pinpoints novel candidate genes in ewing sarcoma family of tumors. *BMC Cancer*, 9:17, 2009.

[168] J. Simola, J. Salojärvi, and I. Kojo. Using hidden markov models to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9:237–251, 2008.

[169] O. Simula, F. Corona, A. Lendasse, M.-L. Riekkola, K. Hartonen, P. Minkkinen, S.-P. Reinikainen, J. Kohonen, I. Vuorinen, J. Hänninen, and J. Silén. Developing chemometrics with the tools of information sciences (CHESS) – MASIT23. In *MASI Programme 2005-2009, Yearbook 2008*, pages 189–222. Libris Oy, May 2008.

[170] J. Sinkkonen, J. Aukia, and S. Kaski. Infinite mixtures for multi-relational categorical data. In S. Kaski, S. V. N. Vishwanathan, and S. Wrobel, editors, *6th International Workshop on Mining and Learning with Graphs (MLG 2008)*, 2008. Proceedings at `http://www.cis.hut.fi/MLG08`.

[171] J. Sinkkonen, J. Parkkinen, J. Aukia, and S. Kaski. A simple infinite topic mixture for rich graphs and relational data. In *Proceedings of the NIPS 2008 Workshop on Analyzing Graphs: Theory and Applications*, December 12 2008. Extended Abstract.

[172] M. Sirola, J. Parviainen, J. Talonen, G. Lampi, T. Alhonnoro, and R. Hakala. Early fault detection with SOM based methods and visualizations - new contents for wide monitoring screens. In *Proceedings of the EHPG-Meeting of OECD Halden Reactor Project*, May 2008.

[173] M. Sirola, J. Talonen, and G. Lampi. SOM based methods in early fault detection of nuclear industry. In *Proceedings of the 17th European Symposium On Artificial Neural Networks ESANN'09*, April 2009.

[174] M. Sjöberg and J. Laaksonen. Optimal combination of SOM search in best-matching units and map neighborhood. In *Proceedings of 7th International Workshop on Self-Organizing Maps (WSOM 2009)*, volume 5629 of *Lecture Notes in Computer Science*, pages 281–289, St. Augustine, Florida, USA, 2009. Springer. Available online at: `http://dx.doi.org/10.1007/978-3-642-02397-2\_32`.

[175] M. Sjöberg, J. Laaksonen, T. Honkela, and M. Pöllä. Inferring semantics from textual information in multimedia retrieval. *Neurocomputing*, 71(13–15):2576–2586, 2008.

[176] M. Sjöberg, V. Viitaniemi, M. Koskela, and J. Laaksonen. PicSOM experiments in TRECVID 2009. In *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009. Available online at `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html`.

[177] A. Sorjamaa, F. Corona, A. Lendasse, Y. Miche, and E. Séverin. Linear combination of SOMs for data imputation: Application to financial problems. In *Proceedings of MASHS 2009, Modèles et Apprentissage en Sciences Humaines et Sociale, Lille (France)*. MASHS, June 8-9 2009.

[178] A. Sorjamaa, F. Corona, Y. Miche, P. Merlin, B. Maillet, E. Séverin, and A. Lendasse. Sparse linear combination of SOMs for data imputation: Application to financial database. In R. Principe, J.C.; Miikkulainen, editor, *Lecture Notes in Computer Science: Advances in Self-Organizing Maps - Proceedings of WSOM 2009 International Workshop on Self-Organizing Maps, Saint Augustine (Florida)*, volume 5629/2009 of *Lecture Notes in Computer Science*, pages 290–297. Springer Berlin / Heidelberg, June 8-10 2009.

[179] A. Sorjamaa, P. Merlin, B. Maillet, and A. Lendasse. A non-linear approach for completing missing values in temporal databases. *European Journal of Economic and Social Systems*, 22(1):99–117, November 2009.

[180] A. Sorjamaa, Y. Miche, R. Weiss, and A. Lendasse. Long-term prediction of time series using NNE-based projection and OP-ELM. In *IEEE World Conference on Computational Intelligence*, pages 2675–2681, Hong Kong, June 2008. Research Publishing Services, Chennai, India.

[181] D. Sovilj, A. Sorjamaa, and Y. Miche. Tabu search with delta test for time series prediction using OP-KNN. In A. Lendasse, editor, *ESTSP, European Symposium on Time Series Prediction*, pages 187–196, Porvoo, Finland, September 17-19 2008. Multiprint Oy / Otamedia , Espoo, Finland.

[182] M. Sulkava. *Learning from environmental data: methods for analysis of forest nutrition time series*. D.Sc. thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D24, Espoo, Finland, January 2008.

[183] T. Suvitaival. Bayesian two-way analysis of high-dimensional collinear metabolomics data. Master's thesis, Helsinki University of Technology, Department of Information and Computer Science, October 2009.

[184] S. B. Taieb, G. Bontempi, A. Sorjamaa, and A. Lendasse. Long-term prediction of time series by combining direct and MIMO strategies. In *International Joint Conference on Neural Networks*, Atlanta, Georgia, USA, June 2009.

[185] J. Talonen and M. Sirola. Abnormal Process State Detection by Cluster Center Point Monitoring in BWR Nuclear Power Plant. In *Proceedings of the 2009 International Conference on Data Mining (DMIN)*, volume I, II, July 2009.

[186] J. Talonen and M. Sirola. Generated Control Limits as a Basis of Operator-Friendly Process Monitoring. In *Proceedings of the 5th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, 2009.

[187] J. Talonen, M. Sirola, and J. Parviainen. Leakage Detection by Adaptive Process Modeling. In *Proceedings of DMIN'08*, pages 49 – 52, 2008.

[188] A. Tripathi, A. Klami, and S. Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

[189] A. Tripathi, A. Klami, and S. Kaski. Using dependencies to pair samples for multiview learning. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1561–1564, 2009.

[190] V. T. Turunen. Reducing the effect of OOV query words by using morph-based spoken document retrieval. In *Proceedings of the 9th Annual Conference of the*

International Speech Communication Association (Interspeech 2008), pages 2158–2161, September 2008.

[191] M. van Heeswijk, Y. Miche, T. Lindh-Knuutila, P. A. J. Hilbers, T. Honkela, E. Oja, and A. Lendasse. Adaptive ensemble models of extreme learning machines for time series prediction. In C. Alippi, M. M. Polycarpou, C. G. Panayiotou, and G. Ellinas, editors, *ICANN (2)*, volume 5769 of *Lecture Notes in Computer Science*, pages 305–314. Springer, 2009.

[192] M. Varjokallio, J. Pylkkönen, and M. Kurimo. Äännemallien diskriminatiivinen opettaminen puheentunnistuksessa. In *Proceedings of the Fonetiikan Päivät – Phonetics Symposium 2008 in Finland*, Tampere, Finland, January 2008.

[193] V. Viitaniemi and J. Laaksonen. Evaluation of pointer click relevance feedback in PicSOM. Technical Report TKK-ICS-R11, Nov. 2008.

[194] V. Viitaniemi and J. Laaksonen. Experiments on selection of codebooks for local image feature histograms. In M. Sebillo, G. Vitiello, and G. Schaefer, editors, *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, volume 5188 of *Lecture Notes in Computer Science*, pages 126–137, Salerno, Italy, September 2008. Springer Verlag.

[195] V. Viitaniemi and J. Laaksonen. Techniques for image classification, object detection and object segmentation. In M. Sebillo, G. Vitiello, and G. Schaefer, editors, *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, volume 5188 of *Lecture Notes in Computer Science*, pages 231–234, Salerno, Italy, September 2008. Springer.

[196] V. Viitaniemi and J. Laaksonen. Techniques for image classification, object detection and object segmentation. Technical Report TKK-ICS-R2, June 2008.

[197] V. Viitaniemi and J. Laaksonen. Combining local feature histograms of different granularities. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 636–645, Oslo, Norway, 2009. Springer Verlag.

[198] V. Viitaniemi and J. Laaksonen. Representing images with χ^2 distance based histograms of SIFT descriptors. In *Proceedings of 19th International Conference on Artificial Neural Networks (ICANN 2009)*, volume 5768 of *Lecture Notes in Computer Science*, pages 636–645, Limassol, Cyprus, 2009. Springer Verlag.

[199] V. Viitaniemi and J. Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of ACM International Conference on Image and Video Retrieval (CIVR 2009)*, Fira, Greece, July 2009.

[200] V. Viitaniemi, M. Sjöberg, M. Koskela, and J. Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, Austria, May 2008.

[201] S. Virpioja and O. Kohonen. Unsupervised morpheme analysis with Allomorfessor. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September 2009.

[202] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo. Thousands of voices for hmm-based speech synthesis. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009*. ISCA, September 2009.

[203] K. Yamazaki and S. Kaski. On asymmetric generalization error of asymmetric multitask learning. In *Learning from Multiple Sources Workshop, 13 December 2008, Whistler Canada*, 2008. Proceedings at `http://web.mac.com/davidrh/LMSworkshop08/Schedule.html`.

[204] K. Yamazaki and S. Kaski. An analysis of generalization error in relevant subtask learning. In M. Köppen, N. Kasabov, and G. Coghill, editors, *Advances in Neuro-Information Processing, 15th International Conference, ICONIP 2008*, pages 629–637, Berlin Heidelberg, 2009. Springer-Verlag.

[205] Z. Yang. *Discriminative Learning with Application to Interactive Facial Image Retrieval*. TKK-ICS-D9, Nov. 2008.

[206] Z. Yang, I. King, Z. Xu, and E. Oja. Heavy-tailed symmetric stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 22 (NIPS2009)*, pages 2169–2177, Vancouver, Canada, 2009.

[207] Z. Yang and J. Laaksonen. Principal whitened gradient for information geometry. *Neural Networks*, 21(2-3):232–240, 2008.

[208] Z. Yang and J. Laaksonen. Informative Laplacian projection. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 359–368, Oslo, Norway, 2009. Springer Verlag.

[209] Z. Yang and E. Oja. Projective nonnegative matrix factorization with α-divergence. In *Proceedings of 19th International Conference on Artificial Neural Networks*, pages 20–29, Limassol, Cyprus, 2009. Springer.

[210] J. Ylipaavalniemi, E. Savia, S. Malinen, R. Hari, R. Vigário, and S. Kaski. Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. *NeuroImage*, 48(1):176–185, October 2009.

[211] J. Ylipaavalniemi and R. Vigário. Analyzing consistency of independent components: An fMRI illustration. *NeuroImage*, 39(1):169–180, January 2008.

[212] J. Ylipaavalniemi and R. Vigário. Matching complex activation patterns with features of natural stimuli. In *1st INCF Congress of Neuroinformatics: Databasing and Modeling the Brain (Neuroinformatics 2008)*, page 71, Stockholm, Sweden, September 2008. Poster.

[213] Q. Yu, A. Lendasse, and E. Séverin. Ensemble KNNs for bankruptcy prediction. In *CEF 09, 15th International Conference: Computing in Economics and Finance, Sydney (Australia)*, June 15-17 2009.

[214] H. Zhang, M. Koskela, and J. Laaksonen. Report on forms of enriched relevance feedback. Technical Report TKK-ICS-R10, Nov. 2008.