Chapter 13

Time series prediction

Amaury Lendasse, Timo Honkela, Federico Pouzols, Antti Sorjamaa, Yoan Miche, Qi Yu, Eric Severin, Mark van Heeswijk, Erkki Oja, Francesco Corona, Elia Liitiäinen, Zhanxing Zhu, Laura Kainulainen, Emil Eirola, Olli Simula

13.1 Introduction

Amaury Lendasse

The Environmental and Industrial Machine Learning (EIML) group is a sub-group of the Adaptive Informatics Applications (AIA) group. It is part of both the Department of Information and Computer Science and the Adaptive Informatics Research Centre, Centre of Excellence of the Academy of Finland.

The EIML group is based on the former Time Series Prediction and Chemoinformatics group, and is developing new Machine Learning techniques: 1) to model environment (using e.g. time series prediction, variable selection and ensemble modeling); 2) to solve industrial problems (for example in the fields of chemometrics, electricity production and distribution, bankruptcy prediction and information security. The EIML group has been created and is lead by Dr. Amaury Lendasse, Docent. The Industrial Machine Learning is under the responsibility of Dr. Francesco Corona, Docent. The information security is under the responsibility of Dr. Yoan Miche.

13.2 Environmental Modeling and Related Tools

Amaury Lendasse,Timo Honkela, Federico Pouzols, Olli Simula and Antti Sorjamaa

Research Environmental Modeling and Time Series prediction are the main research areas of the EIML group.

Environmental Sciences Environmental Sciences have seen a great deal of development and attention over the last few decades, fostered by an impressive improvement in observational capabilities and measurement procedures. The fields of environmental modeling and analysis seek to better understand phenomena ranging from Earth-Sun interactions to ecological changes caused by climatic factors. Traditional environmental modeling and analysis approaches emphasize deterministic models and standard statistical analyses, respectively. However, the application of further developed data-driven analysis methods has shown the great value of computational analysis in environmental monitoring research. Furthermore, these analyses have provided evidence for the feasibility of predicting environmental changes. Thus, linear and nonlinear methods and tools for the analysis and predictive modeling of environmental phenomena are sought. In interpreting biological monitoring data, there is an even stronger need to develop new modeling techniques, because biota does not respond in a linear manner to environmental changes. In addition, a time lag between a stimulus and a response is common, e.g., a change in nutrient concentration and subsequent changes in algal growth, or turbidity of water. Hence, there is a demand for predictive and causal models. In this context, we follow a multidisciplinary approach, involving diverse areas of machine learning. These include time series prediction, ensemble modeling, feature selection and dimension reduction. Our activity concentrates specially on developing new methods and tools motivated by real-world needs in close cooperation with experts in the field. Our current research spans a number of areas, including long-term prediction, spatial-temporal analysis, missing data, irregular and incomplete sampling, and time-frequency analysis. Among a number of application areas, we focus on Marine Biology. Applications of our research have a direct relevance for the Baltic Sea countries. Sophisticated environmental models are needed and directly or indirectly requested by policy makers, industry and citizens. This is of special relevance in the context of regulations such as the EU Water Framework Directive, among others. Our research aims to contribute to the scientific and technological challenges posed by such regulations as well as general challenges in Environmental Sciences worldwide. Time Series Prediction

What is Time series prediction? Time series forecasting is a challenge in many fields. In finance, one forecasts stock exchange or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the electric load and hydrologists forecast river floods. The common point to their problems is the following: how can one analyze and use the past to predict the future? 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 future values. A new challenge in time series prediction is the long-term prediction also known as multiple step-ahead prediction. Many methods designed for time series forecasting perform well (depending on the complexity of the problem) on a rather short-term horizon but are rather poor on a longer-term one. This is due to the fact that these methods are usually designed to optimize the performance at short term, their use at longer term being not optimized. Furthermore, they generally carry out the prediction of a single value while the real problem sometimes requires predicting a vector of future values in one step. One particular problem of long-term prediction is studied: the prediction of the electric load. This problem is very complex and is more and more crucial because of the liberalization of the electricity market. Electricity producers and network companies are looking for models to predict not only their needs for the next hours but also for next days and next weeks.

Our main results can be found in [1, 2, 3, 4, 5, 6, 7].

13.3 Extreme Learning Machine

Yoan Miche, Qi Yu, Eric Severin, Antti Sorjamaa, Mark van Heeswijk, Erkki Oja, Federico Pouzols, 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).

Our main results can be found in [8, 9, 10, 11, 12].

13.4 Process Informatics

Francesco Corona, Elia Liitiäinen, Olli Simula, Zhanxing Zhu and Amaury Lendasse

Process Informatics investigates the development and application of modeling methods from adaptive informatics on measurements from process industry. The methods aim at representing complex chemical and physical processes with models directly derived from the data collected by the automation systems present in the process plants, without an explicit regard to the first principles. We concentrate on algorithmic methods satistying properties like accuracy, robustness, computational efficiency and understandability. Accuracy, robustness and efficiency favor on-line implementations of the models in full-scale applications, whereas understandability permits the interpretation of the models from the aprioristic knowledge of the underlying phenomena. Such an approach to process modeling provides tools that can be used in real-time analysis and supervision of the processes and can be embedded in advanced model -based control strategies and optimization. Specific application domains are chemometrics, spectroscopy, chromatography and on-line analytical technologies in process and power industry. On the algorithmic side we concentrate on methods for nonlinear dimensionality reduction, variable selection, functional and regularized regression.

Our main results can be found in [16, 17, 18, 19, 20].

13.5 Bankruptcy prediction

Yu Qi, Laura Kainulainen, Eric Severin, Olli Simula, Yoan Miche, Emil Eirola 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 [13, 14, 15].

References

[1] A. Ventela, T. Kirkkala, A. Lendasse, M. Tarvainen, H. Helminen and J. Sarvala. Climate-related challenges in long-term management of Sakylan Pyhajarvi (SW Finland) In Hydrobiologia, volume 660, pages 49–58. 2011.

[2] A. Lendasse, T. Honkela and O. Simula. European Symposium on Times Series Prediction In Neurocomputing, volume 73, pages 1919–1922. June, 2010.

[3] A. Guillen, L. Herrera, G. Rubio, A. Lendasse and H. Pomares. New method for instance or prototype selection using mutual information in time series prediction In Neurocomputing, volume 73, pages 2030–2038. June, 2010.

[4] P. Merlin, A. Sorjamaa, B. Maillet and A. Lendasse. X-SOM and L-SOM: A Double Classification Approach for Missing Value Imputation In Neurocomputing, volume 73, pages 1103-1108. March, 2010.

[5] A. Sorjamaa, A. Lendasse, Y. Cornet and E. Deleersnijder. An improved methodology for filling missing values in spatiotemporal climate data set In Computational Geosciences, volume 14, pages 55-64. January, 2010.

[6] F. Pouzols and A. Barros. Automatic Clustering-Based Identification of Autoregressive Fuzzy Inference Models for Time Series In Journal of Multivariate Analysis, volume 101, pages 811–823 . April, 2010.

[7] F. Pouzols, A. Lendasse and A. Barros. Autoregressive Time Series Prediction by Means of Fuzzy Inference Systems Using Nonparametric Residual Variance Estimation In Fuzzy Sets and Systems, volume 161, pages 471–497. February, 2010.

[8] M. Heeswijk, Y.Miche, E. and A. Lendasse. GPU-Accelerated and Parallelized ELM Ensembles for Large-scale Regression In Neurocomputing, volume 74, pages 2430-2437. September, 2011.

[9] Y. Miche, M.Heeswijk, P. Bas, O. Simula and A. Lendasse. TROP-ELM: a Double-Regularized ELM using LARS and Tikhonov Regularization In Neurocomputing, volume 74, pages 2413-2421. September, 2011.

[10] F. Pouzols and A. Lendasse. Evolving fuzzy optimally pruned extreme learning machine for regression problems In Evolving Systems, volume 1, pages 43–58. August, 2010.

[11] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten and A. Lendasse. OP-ELM: Optimally-Pruned Extreme Learning Machine In IEEE Transactions on Neural Networks, volume 21, pages 158–162. January, 2010.

[12] Q. Yu, Y. Miche, A. Sorjamaa, A. Guillen, A. Lendasse and E. Severin. OP-KNN: Method and Applications In Advances in Artificial Neural Systems, volume 2010, pages 6 pages. February, 2010.

[13] L. Kainulainen, Y. Miche, E. Eirola, Q. Yu, B. Frenay, E. Severin and A. Lendasse. Ensembles of Local Linear Models for Bankruptcy Analysis and Prediction In Case Studies in Business, Industry and Government Statistics (CSBIGS), volume 4. November, 2011.

[14] Q. Yu, Y. Miche, E. Severin and A. Lendasse. Bankruptcy Prediction with Missing Data In Proceedings of the 2011 International Conference on Data Mining, pages 279-285. July, 2011.

[15] L. Kainulainen, Q. Yu, Y. Miche, E. Eirola, E. Severin and A. Lendasse. Ensembles of Locally Linear Models: Application to Bankruptcy Prediction In Proceedings of the 2010 International Conference on Data Mining, pages 280–286. July, 2010.

[16] Z. Zhu, F. Corona, A. Lendasse, R. Baratti and J. Romagnoli. Local linear regression for soft-sensor design with application to an industrial deethanizer In 18th World Congress of the International Federation of Automatic Control (IFAC). August, 2011.

[17] F. Corona, A. Lendasse and E. Liitiainen. A boundary corrected expansion of the moments of nearest neighbor distributions In Random Structures and Algorithms, volume 37, pages 223–247. September, 2010.

[18] M. Toiviainen, F. Corona, J. Paaso and P. Teppola. Blind source separation in diffuse reflectance NIR spectroscopy using independent component analysis In Journal of Chemometrics, volume 24, pages 514–522. May, 2010.

[19] E. Liitiainen, F. Corona and A. Lendasse. On the Curse of Dimensionality in Supervised Learning of Smooth Regression Functions In Neural Processing Letters, volume 34, pages 133–154. 2011.

[20] E. Liitiainen, A. Lendasse and F. Corona. Residual variance estimation using a nearest neighbor statistic In Journal of Multivariate Analysis, volume 101, pages 811–823 . April, 2010.