Chapter 4

Multi-source machine learning

Samuel Kaski, Mehmet Gönen, Arto Klami, Gayle Leen, Jaakko Peltonen, Ilkka Huopaniemi, Melih Kandemir, Suleiman A. Khan, Kristian Nybo, Juuso Parkkinen, Tommi Suvitaival, Jaakko Viinikanoja, Seppo Virtanen, Yusuf Yaslan

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. 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 given little data we need to seek alternative ways to bring in more information. Our approach is combining various sources of information.

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 and multi-way learning

Multi-view learning tells how several data sources, or views, can be combined to extract more relevant information. We build Bayesian latent variable models for the task of extracting statistical dependencies between multiple views of the same objects, for example to capture relationships between images and their captions, or between expressions of genes and chemical descriptors of drugs.

In the completely unsupervised case, we are given only the data matrices of co-occurring data, and the task is to mine for dependencies between them. For combining two views we have earlier introduced the **Bayesian Canonical Correlation Analysis** (CCA) model, which finds linear components capturing correlations between the views while modeling the variation specific to each view by separate noise components. To extend the range of potential applications, we have extended the Bayesian CCA model to mixtures of robust CCAs [1] and to generic exponential family noise models [2]. Recently, we introduced a considerably more efficient version of Bayesian CCA [3], which is applicable also to very large dimensionalities. Our novel solution builds on an efficient variational approximation, enabled by reformulating the problem as a group-wise sparse latent component model. Besides working with linear models, we have also presented a nonparametric Bayesian clustering model for similar setups [4].

The problem of analysing dependencies between more than two views is considerably mode difficult. Most solutions seek relationships between all views, whereas most practical applications will not satisfy that assumption. Recently we introduced the **Group Factor Analysis** problem, where the task is to find dependencies between all possible subsets of the views. By building on the group-wise sparsity assumption used for CCA we were able to derive a model that finds efficiently all types of dependencies present in the data collection, even though their potential number grows exponentially as a number of views [5]. The model is illustrated in Figure 4.1.

The task in the **matching** problem is to infer the co-occurrence of the samples from the data set itself. For example, given a collection of documents written in two languages, we might want to learn which documents correspond to each other. In [6] we show how such a match or alignment can be learned simultaneously while learning a model that maximizes the dependency between the two views, by an algorithm that alternates between learning the match and learning a subspace in which the samples can be compared with each other. We also demonstrated how multiple matching solutions can be combined to learn a consensus match over multiple data set instances, to learn a match between metabolites of two species.

The samples co-occurring in the multiple views can also be associated with covariates (labels); then the analysis problem becomes to discover how the different populations indicated by the labels differ from each other, akin to analysis of variance (ANOVA). The problem is particularly difficult in the "large p, small n" case ubiquitous in computational molecular biology, of having a high dimensionality p and a small sample size n. In [7] we introduced a solution combining both multi-view and multi-way learning, by building a Bayesian model that models the covariate effects in the latent space, assuming the views to be conditionally independent given the latent variable, similarly as in the above models. This kind of models and their applications in computational systems biology are discussed in detail in Chapter 5.

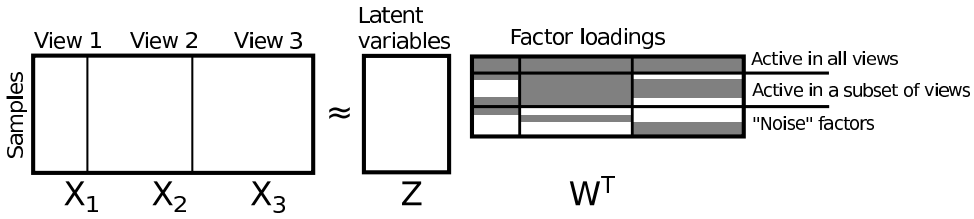

Figure 4.1: Illustration of the group factor analysis of three data sets or views. The feature-wise concatenation of the data sets $\mathbf{X}_i$ is factorized as a product of the latent variables $\mathbf{Z}$ and factor loadings $\mathbf{W}$. The factor loadings are group-wise sparse, so that each factor is active (gray shading, indicating $\mathbf{f}_{m,k} = 1$) only in some subset of views (or all of them). The factors active in just one of the views model the structured noise, variation independent of all other views, whereas the rest model the dependencies. The nature of each of the factors is learned automatically, without needing to specify the numbers of different factor types (whose number could be exponential in the number of views) beforehand.

References

[1] Jaakko Viinikanoja, Arto Klami, and Samuel Kaski. Variational Bayesian mixture of robust CCA models. In Aristides Gionis José Luis Balcázar, Francesco Bonchi and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases. Proceedings of European Conference, ECML PKDD 2010*, volume III, pages 370–385, Berlin, 2010. Springer.

[2] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian exponential family projections for coupled data sources. In Peter Grunwald and Peter Spirtes, editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (2010)*, pages 286–293, 2010. AUAI Press.

[3] Seppo Virtanen, Arto Klami, and Samuel Kaski. Bayesian CCA via group sparsity. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 457–464, New York, NY, 2011. ACM.

[4] 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.

[5] Seppo Virtanen, Arto Klami, Suleiman A. Khan, and Samuel Kaski. Bayesian group factor analysis. In *Proceedings of AISTATS'12*, 2012. Preliminary version available as arXiv:1110.3204.

[6] Abhishek Tripathi, Arto Klami, Matej Orešič, and Samuel Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23:300–321, 2011.

[7] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391–i398, 2010. (ISMB 2010).

4.3 Multi-task learning

We have introduced two new multi-task learning setups, suitable for different scenarios, and solutions for them: *asymmetric multi-task learning* and *multi-task multiple kernel learning*.

4.3.1 Asymmetric multi-task learning

Multi-task learning is the setting where several collections of data samples are analyzed together; each collection represents a different learning task and comes from a different underlying distribution: for example, measurements of student performances in different schools, or scientific documents collected from different venues. The task is usually a supervised task, classification or regression, but may also be an unsupervised task such as clustering or density estimation. Often the data are high-dimensional and the number of data points in each individual task is too small for learning well the subtle distinctions necessary for good performance in the task.

Unlike in multi-view learning, in multi-task learning the individual samples in the data sources do not typically co-occur. Instead it is assumed that there are connections on the population level: if the underlying distributions in the tasks have similar properties (similar trends, groupings, manifolds, etc.) then learning the tasks together allows sharing the data between tasks, making possible learning of more complex models.

Typical multi-task learning solutions are based on treating all of the learning tasks symmetrically (with equal interest), for example by learning a hierarchical probabilistic model from all of the data collections where the models share parameters or priors of parameters; then all the data collections affect learning the shared parameters with an equal role. However, in many settings there is instead a task of interest (such as gene expression measurements of the current patient) where we wish to perform well and where test samples will come from, and other tasks are simply additional sources of information (such as historical records of earlier patients). In such settings the learning should be *asymmetric multi-task learning*: it should focus on learning the task of interest as well as possible, avoiding the danger of skewing the model of the task of interest in favor of modeling other tasks, which can happen in some symmetric approaches.

We have introduced a formalism for asymmetric multi-task learning, focusing on learning a classification task or regression task of interest with the help from auxiliary tasks that are related but are of less interest. On an intuitive level the idea is to *extract only the relevant information* of earlier data sets to help the learning of the task of interest. Technically, we use an intelligent mixture model, where each earlier task is explained partly by a *shared model* and partly by a task-specific *explaining-away model*. The task-of-interest, where everything is relevant, only uses the shared model, while other tasks are partly explained away by the explaining-away model.

Two kinds of methods were derived from this approach: a method for asymmetric multitask logistic regression [2], and a method for asymmetric multitask Gaussian process regression or classification [1]. In the logistic regression case, the model was formulated as

$$p_S(c|\mathbf{x}) = (1 - \pi_S)p^{shared}(c|\mathbf{x}) + \pi_S p_S^{explaining-away}(c|\mathbf{x})$$

where $\mathbf{x}$ are samples c are class labels, $p^{shared}(c|\mathbf{x})$ is a model shared between all tasks,

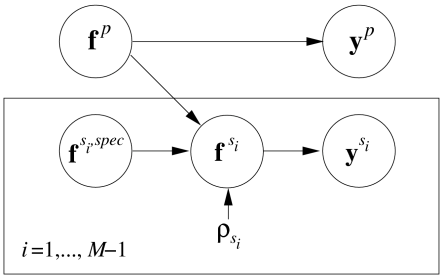

Figure 4.2: Graphical model of an asymmetric multi-task Gaussian process regression model, showing the relationship between the function values of the primary task (task of interest) and secondary tasks (other tasks).

$p_S^{explaining-away}(c|\mathbf{x})$ is a model to explain away non-relevant parts of task S and π_S is a mixture weight. In a Gaussian process regression context this can be similarly written as

$$y = f_S(\mathbf{x}) = f^{shared}(\mathbf{x}) + f_S^{explaining-away}(\mathbf{x})$$

where the y are regression targets, $f^{shared}(\mathbf{x})$ is a shared regression function and $f_S^{explaining-away}(\mathbf{x})$ is a function to explain away non-relevant regression variation of task S, and the functions are drawn independently from Gaussian process priors.

The methods were shown to outperform both naive approaches, such as single-task learning or pooling together all tasks, and also the nearest comparable symmetric multi-task learning approaches.

4.3.2 Multi-task multiple kernel learning

Empirical success of kernel-based learning algorithms is very much dependent on the kernel function used. Instead of using a single fixed kernel function, multiple kernel learning algorithms learn a combination of different kernel functions in order to obtain a similarity measure that better matches the underlying problem. We study multi-task learning problems and formulate a novel multi-task learning algorithm [3] that trains coupled but nonidentical multiple kernel learning models across the tasks. The proposed algorithm is especially useful for tasks that have different input and/or output space characteristics and is computationally very efficient. Empirical results on three data sets validate the generalization performance and the efficiency of our approach.

References

[1] Gayle Leen, Jaakko Peltonen, and Samuel Kaski. Focused Multi-task Learning Using Gaussian Processes. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases (proceedings of ECML PKDD 2011)*, Part II, pages 310–325, Berlin Heidelberg, 2011. Springer-Verlag. Winner of the ECML PKDD 2011 Best Paper Award in Machine Learning.

[2] Jaakko Peltonen, Yusuf Yaslan, and Samuel Kaski. Relevant subtask learning by constrained mixture models. *Intelligent Data Analysis*, 14:641–662, 2010.

[3] Mehmet Gönen, Melih Kandemir, and Samuel Kaski. Multitask Learning Using Regularized Multiple Kernel Learning. In Bao-Liang Lu, Liqing Zhang, and James Kwok, editors, *Proceedings of 18th International Conference on Neural Information Processing (ICONIP)*, Part II, pages 500–509, Berlin Heidelberg, 2011. Springer-Verlag.

4.4 Information visualization

Information visualization is an essential part of analysis of new data, especially in the first stages when strong hypotheses about the data have not been made yet. Many dimensionality reduction methods have been designed for tasks such as manifold learning and are not suitable for reducing the data much beyond the effective dimensionality of data. A visualization on a low-dimensional display cannot represent all aspects of high-dimensional data: it is then crucial to be able to quantify the errors that unavoidably occur in any visualization.

We have *formalized information visualization as a task of visual information retrieval* [1, 2], focusing on the specific task of retrieving similar items (retrieving neighborhood relationships) for a query item based on the visual display. In this task, all visualizations naturally end up with two kinds of errors, false neighbors and misses. The accuracy of such retrieval can be rigorously quantified using the information retrieval measures *precision* and *recall*. The analyst needs to specify a tradeoff between precision and recall (tradeoff between the costs of false neighbors and misses) to evaluate the goodness of visualizations. Moreover, generalizations of such measures can be directly set as an optimization goal, to produce visualizations that are optimal for information retrieval. We have also shown that optimizing visualizations for information retrieval can be done in the framework of generative modeling [3]. We have created nonlinear embeddings optimal for information retrieval, and have shown that they outperform existing visualization methods in the information retrieval tasks, and also by commonly used indirect measures.

We have applied the approach to visualization of graphs (graph layout) [4], fMRI data (Fig 4.3), and gene expression measurements (e.g., [5]).

References

[1] 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.

[2] Samuel Kaski and Jaakko Peltonen. Dimensionality Reduction for Data Visualization. *IEEE Signal Processing Magazine*, 28(2):100–104, 2011.

[3] Jaakko Peltonen and Samuel Kaski. Generative Modeling for Maximizing Precision and Recall in Information Visualization. In Geoffrey Gordon, David Dunson, and Miroslav Dudik, editors, *Proceedings of AISTATS 2011, the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP, vol. 15*, 2011.

[4] Juuso Parkkinen, Kristian Nybo, Jaakko Peltonen, and Samuel Kaski. Graph Visualization With Latent Variable Models. In *Proceedings of MLG 2010, the Eighth Workshop on Mining and Learning with Graphs*, pages 94-101, New York, NY, USA, 2010. ACM.

[5] 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*, pages 2178-2181, 2010. IEEE.

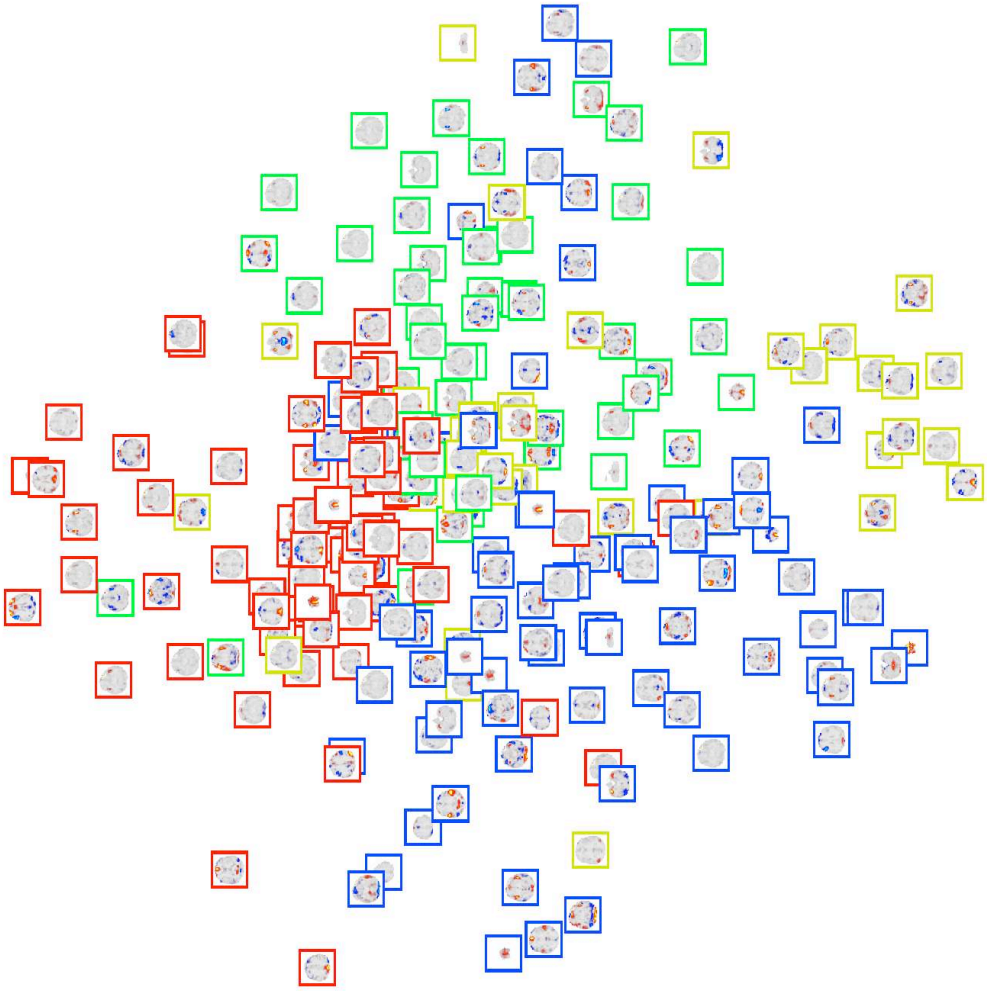

Figure 4.3: Visualization of fMRI whole-head volumes from an experiment with several people experiencing multiple stimuli. The visualization has been optimized for information retrieval of similar (neighbor) images from the visualization. The four stimuli types (red: tactile, yellow: auditory tone, green: auditory voice, blue: visual) have become separated in the visualization; the two auditory stimuli types are arranged close-by as is intuitively reasonable. An axial slice is shown for each whole-head volume, chosen so that the shown slice contains the highest-activity voxel.