From learning metrics towards dependency exploration

Samuel Kaski
Laboratory of Computer and Information Science, Helsinki University of Technology, and
Department of Computer Science, University of Helsinki, Finland
samuel.kaski@hut.fi

Abstract - *We have recently introduced new kinds of data fusion techniques, where the goal is to find what is shared by data sets, instead of modeling all variation in data. They extend our earlier works on learning of distance metrics, discriminative clustering, and other supervised statistical data mining methods. In the new methods the supervision is symmetric, which translates to mining of dependencies. We have so far introduced methods for associative clustering and for extracting dependent components which generalize classical canonical correlations.*

Key words - Associative clustering, data fusion, dependent components, discriminative clustering, learning metrics

1 Introduction

We address the data-analytic problem of making a new kind of data fusion: How to extract common properties in several data sets. While this problem recurs in several fields, the setting is motivated in particular by new data-analysis challenges in bioinformatics.

One of the hardest challenges of modern biology and medicine, for which the need of bioinformatics methods is widely recognized, is how to best use the new genome-wide measurement technologies and the growing community-resource databanks of measurement data. They are needed especially in systems biology, in understanding how the parts of the cell, and groups of cells, function together. The data-analytic problems stem from measurement and biological noise, that is, irrelevant variation that may even have structure which is unknown *a priori*. Moreover, the research tasks are necessarily ill-defined: the ultimate goal is to make new scientific findings, and it is not clear yet what kinds of findings are possible with the new kinds of measurements.

The general problem setting, knowledge discovery from databases, has been recognized in a branch of computer science called data mining. Mainstream data mining methods are insufficient, however, because they cannot adequately handle the noise and the irrelevant variation in the data. Mainstream statistics is not sufficient either, since in the absence of sufficient prior knowledge the models need to be manually tailored to each task and databank, which is too laborious for widespread use in the various research laboratories. Machine learning addresses the problem of automating part of the modeling process, by aiming at more general-purpose methods, but so far mostly in specific kinds of restricted learning tasks.

Our main methodological goal is to combine the merits of the three different approaches into new kinds of modeling tools, within a field that could be called *statistical data mining.*

In traditional unsupervised learning, a statistical machine learning approach applicable to data mining, the data is a set of samples $\{\mathbf{x}\}$, which may be either vectorial or discrete-valued. The task is to find regularities in the data, usually in the density of the samples (clustering) or in the relationships of the variables (component models). Common to both is that, in probabilistic terms, the goal is to model the probability density $p(\mathbf{x})$ of the data. This goal is shared by much of statistical generative modeling; the distinctive feature in machine learning is that there usually exists very little prior knowledge of the data generating mechanism, and everything needs to be learned from data. The task of better understanding a data set by searching for regularities in it could be called data exploration or *unsupervised data mining.*

Since it is very hard if not impossible to derive everything from data, we have studied methods for inserting more knowledge into the mining task in a data-driven fashion. We work with settings where there are two or more data sets with *co-occurring samples*, that is, samples coming in pairs $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ belongs to the first set and $\mathbf{y}$ to the second set. Both samples may be either discrete or continuous-valued. For such data two different kinds of classical machine learning approaches are possible. If the samples are concatenated together to form a single long vector, the problem reduces to standard unsupervised learning for modeling the joint density $p(\mathbf{x}, \mathbf{y})$. The other task, called supervised learning, is to predict the values of the other, say $\mathbf{y}$, given $\mathbf{x}$. The classical supervised tasks are to find classes (in classification) or predicted values of a variable (in regression). In probabilistic terms the goal is to build a good model for the conditional distribution $p(\mathbf{y}|\mathbf{x})$.

Our task is not equivalent to either of the two classical ones, but related. The goal is to use the information contained in the relationship between well-chosen data set pairs $\{\mathbf{x}\}$ and $\{\mathbf{y}\}$, to direct mining of $\mathbf{x}$ (and later to symmetrically mine both $\mathbf{x}$ and $\mathbf{y}$). Methods capable of doing that are useful in situations where it is hard to specify prior knowledge into the model structure, but it is easier to point out a collection of data sets sharing the interesting properties and only them. Examples of such tasks are given in the following sections.

The general research problem is to find *common properties* in the set of pairs; statistically speaking, the goal is to find statistical dependencies between the $\mathbf{x}$ and $\mathbf{y}$. When applied to statistical data mining, the models could be called *dependency exploration methods*; they model the statistical dependencies between several data sets in order to find properties shared by them. The methods are generalizations of our data analysis principle called *learning metrics* [7, 9], where an adaptive distance measure is computed to focus on only relevant or important differences between data samples, and neglect the irrelevant. The metric was learned from paired data sets, and relevance was defined as relevance for the other data set.

A seemingly related task is to search for differences between data sets [3]. There the relative order of the samples within the two sets is not significant, both sets are within the same space, and the goal is to find differences between data *distributions.* The fundamental difference is that our data are paired and we search for commonalities between the pairs of *samples* that can have different variables (attributes) and different dimensionalities.

2 Learning metrics and discriminative models

Completely unsupervised data mining is a tough task, since in the absence of prior knowledge of model structures it is very hard or even impossible to distinguish interesting regularities from noise. The reason is that interestingness is often related to unexpectedness, and noise is by definition unpredictable and hence unexpected. Normally in probabilistic modeling this problem is circumvented by restricting the search space by incorporating prior knowledge either to the model structure or as the prior distribution of the parameters.

In data mining or data exploration prior knowledge usually is not available, since the goal is to make new findings. Yet, it is clear that prior knowledge needs to be input to the analysis process somehow—otherwise the analyst will put garbage in and get garbage out. The usual way in unsupervised learning is through variable selection: When the data vector contains only relevant or interesting variables, the unsupervised search for regularities in the data will be carried out only among those interesting variables. This will presumably result in finding of interesting regularities.

Variable selection may be hard, however, and moreover the variables need to be properly normalized and put into a common scale, weighted by how important they are. In completely unsupervised settings this all needs to be done manually. Realizing that the problem is much less severe in supervised learning, where the variable selection and scaling can be done to maximize prediction ability of the supervised model, we started thinking whether supervision could be applicable in data mining as well.

The task of variable selection and preprocessing effectively boils down to selection of the metric of the data space. Variable selection corresponds to binary weighting of the variables, and weighting corresponds to having a specific kind of a distance matrix. Linear pre-processing, on its part, corresponds to having a more general distance matrix. The most general kind of distance matrix, which gives a Riemannian distance, is different in different locations of the data space.

Given the task of mining the data $\mathbf{x}$, without applicable prior knowledge but with co-occurring data $(\mathbf{x}, \mathbf{y})$ available, we then wanted to extract useful information from the pairs to better focus the analysis of $\mathbf{x}$. The "auxiliary data" $\mathbf{y}$ had the same role of giving the "ground truth" of what is important as in supervised learning, but the difference was that the task was to explore $\mathbf{x}$. The solution, coined the *learning metrics principle* [7], was to optimize the metric of the data space to measure only the differences relevant to changes in the auxiliary data. After this supervised preprocessing the mining of $\mathbf{x}$ can be unsupervised. Such data mining using the learning metrics principle could be called *supervised data mining* since one data set, the $\{\mathbf{y}\}$, supervises the mining by telling what is important in the other, $\{\mathbf{x}\}$.

The metric was defined by using information-geometric concepts, several methods were developed for estimating it, and the metric was applied to several unsupervised data analysis algorithms [5, 9, 13], including the Self-Organizing Maps. The metric is generally applicable—when used with a data mining method which assumes spherically symmetric noise, it effectively transforms the model such that it assumes inhomogenous noise. There will be more noise in the relatively less important directions, and hence the model will pay less attention to them.

An information-theoretically derived method cannot rigorously take into account all uncertainties in small data sets. For this reason we developed alternative methods that maximize a rigorous finite-data criterion, the likelihood of predicting the auxiliary data $\mathbf{y}$ given $\mathbf{x}$. The methods called *discriminative clustering* [8, 15] and *relevant component analysis* or "informative" discriminant analysis [12] can be shown to asymptotically operate using the learning metrics.

3 Models of dependencies

In the learning metrics principle and the related discriminative models, the mining is asymmetric with respect to $\mathbf{x}$ and $\mathbf{y}$. The $\mathbf{y}$ is assumed to give the gold standard for selecting the metric for $\mathbf{x}$, that is, to supervise the mining of $\mathbf{x}$. Next we wanted to make the supervision symmetric.

Note that basic unsupervised learning, when applied to vectors $\mathbf{x}$ and $\mathbf{y}$ concatenated together, is symmetric in a trivial sense. All variation in the combination—including the individual variation in $\mathbf{x}$ and in $\mathbf{y}$—is modeled, and there is no mechanism for separating between-data-set variation from within-data-set variation.

The task addressed in our current work is modeling only the variation in $\mathbf{x}$ and $\mathbf{y}$ that is *common* to both variables. In other words, we search for *dependencies* between the $\mathbf{x}$ and $\mathbf{y}$. This symmetric goal can be formalized as maximizing the dependency between two representations, $\hat{\mathbf{x}} \equiv \mathbf{f}_x(\mathbf{x}; \boldsymbol{\theta}^x)$ and $\hat{\mathbf{y}} \equiv \mathbf{f}_y(\mathbf{y}; \boldsymbol{\theta}^y)$, of $\mathbf{x}$ and $\mathbf{y}$, respectively. Here the $\boldsymbol{\theta}$ are the parameters to be optimized, and the traditional measure of dependencies is the mutual information. A familiar example is canonical correlation analysis [4] where both the $\mathbf{f}_x$ and $\mathbf{f}_y$ are linear projections, and the data are assumed to be normally distributed. This idea has earlier been generalized to nonlinear functions [1], and to finding clusters of $\mathbf{x}$ informative of a nominal-valued y [2, 15] (discussed in the previous section). It has been formalized in the information bottleneck framework [19, 17], resulting in efficient algorithms for two nominal-valued variables [18, 14].

These kinds of dependency models focus on the variation that is common in the data sets, and skip data set-specific variation. The theoretical difference from a commonly used modeling paradigm, Bayes networks, is that in the absence of prior knowledge, these latter types of models do not distinguish between data set-specific variation and variation common to both data sets, and hence in this sense they make a compromise in our task.

Modeling of dependency between data sets is a new data mining target that is considerably better-focused target than the common, fully unsupervised search for clusters and other regularities. Yet it is general and data-driven. It is intended to be applicable in problems when prior knowledge either is not available or cannot easily be incorporated in the model, but where it is easier to collect a set of relevant data sets. An example is analysis of yeast stress response. It is much harder to define stress response than it is to name stressful treatments. Each data set consists of measurements of the yeast's response to a different stressful treatment, and while the response specific to any of the treatments is not interesting, the response shared by all of them is. It defines stress.

The above-mentioned information-theoretic derivations for maximizing dependency between representations of $\mathbf{x}$ and $\mathbf{y}$ suffer from the same problem as the learning metrics: Since they are defined for distributions, they are not rigorously applicable to finite data sets. We are currently developing methods for dependency exploration or mining for finite data.

3.1 Associative clustering

The standard unsupervised clustering methods, reviewed for gene expression clustering for instance in [11], aim at finding clusters where genes have similar expression profiles. Our goal is different: to cluster the $\mathbf{x}$ and the $\mathbf{y}$ separately such that the dependencies between the two clusterings capture as much as possible of the statistical dependencies between two sets of clusters. In this sense the clustering is *associative*; it finds associations between samples of different spaces.

Analogously to the two linear projections in canonical correlation analysis, we use two sets of clusters as the representations in the dependency search. Clusters are more flexible than linear projections, and they have a definite role in exploratory data analysis, that is, in "looking at the data:" Clustering reveals outliers, finds groups of similar data, and simply compresses numerous samples into a more manageable and even visualizable summary. Clusters and other kinds of unsupervised models are of particular importance as the first step of microarray data analysis, where data are often noisy and even erroneous, and in general not well-known *a priori*.

Note that this very legitimate and necessary use of clustering in the beginning of the research process should not be confused with the widespread use of clusterings as a general-purpose tool in all possible research tasks, which could better be solved by other methods.

Associative clustering [6] builds on the insight [16] that discriminative clustering optimizes the dependency between the clusters and the (multinomial) auxiliary data, measured from their cross-tabulation into a *contingency table*. Dependencies in a contingency table can be measured by a certain kind of a Bayes factor, which discriminative clustering optimizes.

The difference between discriminative clustering and associative clustering is that while in the former the other margin of the contingency table is fixed by the fixed auxiliary data and only the margin formed by the clusters is optimized, associative clustering optimizes both margins. While this may seem to be a small change, the conceptual change is large. Discriminative clustering can be interpreted as maximization of the conditional likelihood for modeling $p(\mathbf{y}|\mathbf{x})$. We are not aware of any such finite-data probabilistic interpretation for the symmetric task of associative clustering.

3.2 Dependent components

In associative clustering the $\mathbf{x}$ and the $\mathbf{y}$ are represented by indexes of their respective clusters, and the dependencies can be measured from their cross-tabulation, the contingency table. In component models the representations are continuous-valued linear projections.

Standard canonical correlation analysis assumes that both $\mathbf{x}$ and $\mathbf{y}$ are normally distributed, which implies that their projections are as well, and the dependencies measured by mutual information reduce to correlations between the projected values.

We have extended canonical correlation to non-normally distributed data by measuring the dependencies with non-parametric estimates [10]. The *non-parametric dependent components* are computed by optimizing a finite-data criterion, the likelihood ratio, between a model that assumes that the components of the two data sets are dependent vs. independent. The dependency is measured with non-parametric Parzen estimators in the low-dimensional space formed of the components. Asymptotically, for large data set, the criterion still converges to mutual information, and in this sense the method generalizes canonical correlations.

4 Conclusion

This paper is a summary of our recent work from data mining supervised by another data set to symmetric supervision between several data sets. We have developed a set of methods which search for dependencies between data sets, applicable in a task which could be called dependency mining.

The methods are applicable in settings where the variation specific to any single data set is not interesting, and only the variation common to the sets counts. If a good model for the data generating mechanism exists, including all variation in all data sets, the common variation can be sought with standard probabilistic modeling of the joint density of the data. If not, new methods targeting the common variation are needed.

While the supervised mining methods have already been applied to several information visualization methods, including the Self-Organizing Maps, the work in the dependency exploration methods has only been started. The dependent components generalize canonical correlations by removing the restriction to normally distributed data. They can be used for visualizing data, but the components are still linear. The next step is to generalize them to non-linear components, where a Self-Organizing Map-type discrete approximation of the visualized manifold is a viable option.

Acknowledgments

This paper is a summary of the motivations underlying the recent work by my research group. More details at `http://www.cis.hut.fi/projects/mi`. Contributors to the works have been listed as authors of the cited references. The work has been funded by the Academy of Finland and the National Technology Agency of Finland, and for a small part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

[1] S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, 1992.

[2] Suzanna Becker. Mutual information maximization: models of cortical self-organization. *Network: Computation in Neural Systems*, 7:7–31, 1996.

[3] V. Ganti, J. Gehrke, R. Ramakrishnan, and W.-Y. Loh. A framework for measuring changes in data characteristics. In *Proceedings of ACM PODS 1999, 18th Symposium on Principles of Database Systems*, pages 126–137. 1999.

[4] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[5] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4:48, 2003.

[6] Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Juha Knuuttila, and Christophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Accepted for publication.

[7] Samuel Kaski and Janne Sinkkonen. Principle of learning metrics for exploratory data analysis. *Journal of VLSI Signal Processing, special issue on Machine Learning for Signal Processing*, 37:177–188, 2004.

[8] Samuel Kaski, Janne Sinkkonen, and Arto Klami. Discriminative clustering. *Neurocomputing*. Accepted for publication.

[9] Samuel Kaski, Janne Sinkkonen, and Jaakko Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

[10] Arto Klami and Samuel Kaski. Non-parametric dependent components. In *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages V–209–V–212. IEEE, 2005.

[11] G. J. McLachlan, Kim-Anh Do, and Christophe Ambroise. *Analyzing microarray gene expression data*. Wiley, New York, 2004.

[12] Jaakko Peltonen and Samuel Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16:68–83, 2005.

[13] Jaakko Peltonen, Arto Klami, and Samuel Kaski. Improved learning of Riemannian metrics for exploratory analysis. *Neural Networks*, 17:1087–1100, 2004. Invited paper.

[14] Jaakko Peltonen, Janne Sinkkonen, and Samuel Kaski. Sequential information bottleneck for finite data. In Russ Greiner and Dale Schuurmans, editors, *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 647–654. Omnipress, Madison, WI, 2004.

[15] Janne Sinkkonen and Samuel Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[16] Janne Sinkkonen, Samuel Kaski, and Janne Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the ECML'02, 13th European Conference on Machine Learning*, pages 418–430. Springer, Berlin, 2002.

[17] Noam Slonim. *The information bottleneck: theory and applications.* PhD thesis, Hebrew University, Jerusalem, 2002.

[18] Noam Slonim, Nir Friedman, and Naftali Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 129–136. ACM Press, New York, NY, USA, 2002.

[19] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In Bruce Hajek and R. S. Sreenivas, editors, *Proceedings of The 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377. University of Illinois, Urbana, Illinois, 1999.