edited by R. Ramakrishnan, S. Stolfo, R. Bayardo and I. Parsa. The Association for Computing Machinery, New York, NY, USA, 2000.

Unsupervised Bayesian Visualization of High-Dimensional Data

Petri Kontkanen, Jussi Lahtinen, Petri Myllymäki, Henry Tirri
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland
cosco@cs.Helsinki.FI, http://www.cs.Helsinki.FI/research/cosco/

ABSTRACT

We propose a data reduction method based on a probabilistic similarity framework where two vectors are considered similar if they lead to similar predictions. We show how this type of a probabilistic similarity metric can be defined both in a supervised and unsupervised manner. As a concrete application of the suggested multidimensional scaling scheme, we describe how the method can be used for producing visual images of high-dimensional data, and give several examples of visualizations obtained by using the suggested scheme with probabilistic Bayesian network models.

1. INTRODUCTION

Multidimensional scaling (see, e.g., [3, 2]) is a data compression or data reduction task where the goal is to replace the original high-dimensional data vectors with much shorter vectors, while losing as little information as possible. Intuitively speaking, it can be argued that a pragmatically sensible data reduction scheme is such that two vectors close to each other in the original multidimensional space are also close to each other in the lower-dimensional space. This raises the question of a distance measure — what is a meaningful definition of similarity when dealing with high-dimensional vectors in complex domains?

Traditionally, similarity is defined in terms of some standard geometric distance measure, such as the Euclidean distance. However, such distances do not generally reflect properly the properties of complex problem domains, where the data typically is not coded in a geometric or spatial form. In this type of domains, changing one bit in a vector may totally change the relevance of the vector, and make it in some sense a quite different vector, although geometrically the difference is only one bit. This issue is discussed in more detail in Section 2.

In [9] we proposed and analyzed a supervised, probabilistic model-based data reduction scheme where the similarity of two vectors was determined by using a formal model of the problem domain. In the suggested Bayesian framework, *two vectors are considered similar if they lead to similar predictive distributions*, when the corresponding attribute-value pairs are given as input to the same probabilistic model. The idea is related to the Bayesian distance metric suggested in [11] as a method for defining similarity in the case-based reasoning framework. The basic principles of the suggested supervised Bayesian data reduction scheme are reviewed in Section 3.

An obvious drawback with the approach suggested in [9] is that the scheme is inherently supervised in nature, and cannot be applied in unsupervised domains. To overcome this limitation, in Section 4 we introduce a novel, unsupervised Bayesian distance measure based on an extension of the earlier supervised approach. As a concrete application of the suggested unsupervised Bayesian data reduction scheme, we consider the problem of visualizing high-dimensional data on a 2D or 3D display. This type of visualizations can be exploited in finding regularities in complex domains, and applied in various data mining tasks, such as instance selection and clustering. A formal description of the visualization problem is given in Section 2.

In this paper we use probabilistic *Bayesian networks* [16, 14] as the formal model family required in our Bayesian data reduction framework. Intuitively speaking, a Bayesian (belief) network is a representation of a probability distribution over a set of (usually) discrete variables, consisting of an acyclic directed graph, where the nodes correspond to domain variables, and the arcs define a set of independence assumptions which allow the joint probability distribution for a data vector to be factorized as a product of simple conditional probabilities. Techniques for learning such models from sample data are discussed in [5]. One of the main advantages of the Bayesian network model is the fact that with certain technical assumptions it is possible to marginalize (integrate) over all parameter instantiations in order to produce the corresponding predictive distribution. As demonstrated in, e.g., [13], such marginalization improves the predictive accuracy of the Bayesian network model, especially in cases where the amount of sample data is small. Practical algorithms for performing predictive inference in general Bayesian networks are discussed for example in [16, 15, 7].

KDD 2000 Boston MA USA

After producing a compressed, two- or three-dimensional representation of our data, an obvious question concerns the quality of the result: how do we know whether the compressed data set represents the problem domain in a reasonable manner? This question can of course be partly answered through the results of a data mining process: if the user is capable of discovering new, interesting regularities in the data based on the visualization obtained through the data reduction scheme, we can say that the compression is in some sense reasonable, at least from a pragmatic point of view. However, we would like to come up with a more theoretically rigorous, statistical methodology for estimating the quality of different compressions. This important question is discussed in Section 5, where we also give illustrative examples of visualizations obtained with public-domain classification data sets.

2. THE VISUALIZATION PROBLEM

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ denote a collection of N vectors, and let us assume that each vector $\mathbf{x}_i$ consists of values of m attributes $X_1, \ldots, X_m$. For simplicity, in the sequel we will assume the attributes X_i to be discrete.

Let $\tilde{\mathbf{X}}$ denote a new data matrix where each m-component data vector $\mathbf{x}_i$ is replaced by a two- or three-component data vector $\tilde{\mathbf{x}}_i$. As this new compressed data matrix can easily be plotted on a two- or three-dimensional display, the result can be used for producing a visual representation of a high-dimensional domain, and hence in the sequel we will call $\tilde{\mathbf{X}}$ the *visualization* of data $\mathbf{X}$.

For producing a visualization of a high-dimensional data set $\mathbf{X}$, we need to find a transformation (function) which maps each data vector $\mathbf{x}_i$ in the domain space to a vector $\tilde{\mathbf{x}}_i$ in the visual space. In order to guarantee the usefulness of the transformation used, an obvious requirement is that two vectors close to each other in the domain space should also be close to each other in the visual space. One way to express this condition formally is to aim at minimizing the following objective function $F(\mathbf{X}, \tilde{\mathbf{X}})$,

$$F(\mathbf{X}, \tilde{\mathbf{X}}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(d(\mathbf{x}_i, \mathbf{x}_j) - \tilde{d}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\right)^2, \qquad (1)$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the pairwise distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in the domain space, and $\tilde{d}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ the distance between the corresponding vectors in the visual space. This approach to the visualization problem is known as Sammon's mapping (see [8]).

Our practical goal is to apply the reduced data set $\tilde{\mathbf{X}}$ for data visualization purposes, so the geometric Euclidean distance is a natural choice for the distance metric $\tilde{d}(\cdot)$ in the visual space. Nevertheless, it is important to realize that there is no a priori reason why this distance measure would make a good similarity metric in the high-dimensional domain space. As a matter of fact, in many complex domains it is quite easy to see that geometric distance measures reflect poorly the significant similarities and differences between the data vectors. Handling discrete data is especially difficult: many data sets contain nominal or ordinal attributes, in which case finding a reasonable coding with respect to the Euclidean distance metric is a difficult task. Furthermore, the results are highly dependable on attribute scaling: as all attributes are treated as equal, it is obvious that an attribute with a scale of, say, between -1000 and 1000, is more influential than an attribute with a range between -1 and 1. What is more, although it seems at first sight that Euclidean distance is model-free in the sense that the similarities are not based on a any specific domain model, this view is flawed: when summing over the pairwise distances between different attribute values independently, we have already made an implicit global independence assumption, although we have not stated this (and other) assumptions explicitly. For these reasons, we argue that although the Euclidean distance is an obvious choice for the distance metric $\tilde{d}(\cdot)$, in general $d(\cdot)$ should be different from $\tilde{d}(\cdot)$.

There have been several attempts to circumvent the above weaknesses by using various coding schemes or variants of the Euclidean distance measure, such as the Mahalanobis distance (see, e.g., [2]). However, the proposed approaches either use ad hoc methodologies with no theoretical framework to support the solutions presented, or are based on relatively simple implicit assumptions that do not usually hold in practice. As an example of the latter case, it is easy to see that the Euclidean distance is based on an underlying model with normally distributed, independent variables, while the Mahalanobis distance assumes the multivariate normal model. These models are clearly too simple for modeling practically interesting, complex domains, especially without the explicit, formal theoretical framework that can be used for determining the model parameters.

3. A SUPERVISED BAYESIAN DISTANCE METRIC

We argue that that in order to overcome the problems listed in Section 2, our assumptions concerning the domain space should be explicitly listed and exploited by using a formal model of the problem domain. By a model M we mean here a parametric model form so that each parameterized instance (M, θ) of the model produces a probability distribution $P(X_1, \ldots, X_m | M, \theta)$ on the space of possible data vectors $\mathbf{x}$. To make our presentation more concrete, for the remainder of the paper we assume that the models M represent different *Bayesian network structures* (for an introduction to Bayesian network models, see e.g., [16, 15]).

The general idea can be summarized as follows: *two vectors are considered similar if they lead to similar predictive distributions*, when the corresponding attribute-value pairs are given as input to the same Bayesian network model M. To make this idea more precise, we must first define the predictive distribution used in the above informal definition. In [9] this predictive distribution was determined with respect to a special *target variable* X_k, resulting in the conditional distribution

$$P(X_k \mid X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_m, M). \qquad (2)$$

Data vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are now considered similar if the corresponding predictive distributions are similar, i.e., $P(X_k \mid \mathbf{x}_i^-, M) \approx P(X_k \mid \mathbf{x}_j^-, M)$, where $\mathbf{x}_i^-$ denotes the attribute values in vector $\mathbf{x}_i$ without the value of the target variable X_k.

This type of similarity measures lead to supervised distance measures, and we can easily change the focus of the metric by changing the target variable X_k. The scheme is also scale invariant as we have moved from the original attribute space to the probability space where all the numbers lie between 0 and 1. This also allows us to handle different type of attributes (discrete or continuous) in the same consistent framework. Furthermore, the framework fulfills the requirement stated earlier: the approach is theoretically on a solid basis as all our domain assumptions must be formalized in the model M.

The above scheme still leaves us with the question of defining a similarity measure between two predictive distributions. The standard solution for computing the distance between two distributions is to use the Kullback-Leibler divergence. However, this asymmetric measure is not (in its basic form) a distance metric in the geometric sense. In the empirical experiments reported in [9] it was observed that the simpler distance metric $d_k(\mathbf{x}_i, \mathbf{x}_j) = 1.0 - P(\mathrm{MAP}_k(\mathbf{x}_i) = \mathrm{MAP}_k(\mathbf{x}_j))$ yields good results in practice. Here $\mathrm{MAP}_k(\mathbf{x}_i)$ denotes the *maximum posterior probability* value of target variable X_k with respect to the predictive distribution (2). In this paper we however use a slightly different distance function $d(\mathbf{x}_i, \mathbf{x}_j)$ based on a straightforward logarithmic transformation of the predictive probabilities:

$$d_k(\mathbf{x}_i, \mathbf{x}_j) = -\log P(\mathrm{MAP}_k(\mathbf{x}_i) = \mathrm{MAP}_k(\mathbf{x}_j)). \quad (3)$$

As noted in [9], extending this general approach to cases with two or more target variables is straightforward.

4. AN UNSUPERVISED BAYESIAN DISTANCE METRIC

The distance metric described in Section 3 is inherently supervised in nature, as it requires us to choose one (or more) of the domain variables to be used as the target variable. Consequently, this approach cannot be directly used cases where no natural candidate for such a target variable exist, and we would like to visualize our data in a purely unsupervised manner. For such unsupervised domains, we propose the following unsupervised extension of the suggested supervised similarity metric:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{m} d_k(\mathbf{x}_i, \mathbf{x}_j). \quad (4)$$

Consequently, the distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed by taking each of the variables X_k in its turn as the target variable, and summing the resulting m supervised distance measures computed by formula (3). Intuitively speaking, this means that two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are considered similar, if the most probable outcome is the same in both cases in all the m individual supervised prediction tasks based on the m conditional distributions (2) for all $k \in \{1, \ldots, m\}$. As we are using a sum of logarithms in the definition of our overall unsupervised distance function (4), this means that we are basically treating all these separate supervised prediction tasks independently.

In the Bayesian framework, the model structure M used in the conditional distributions required can be determined by maximizing the posterior probability $P(M \mid \mathbf{X})$. Assuming the uniform prior for the model structures, this equals to using the model with the highest *marginal likelihood* or *evidence*. The required conditional distribution (2) can then be computed by marginalizing the joint probability distribution $P(\mathbf{x}_i \mid \mathbf{X}, M)$ appropriately.

However, as noted in e.g. [6], finding the maximal evidence model structure is not a feasible task in practice as the number of possible Bayesian network structures is super-exponential. This means that in practical situations we are dealing with a model structure M that is possibly only a poor model of the "true" joint domain probability distribution, and hence some of the probabilities obtained are not correct. As demonstrated in [12], instead of trying to find a good model of the joint probability distribution, in supervised classification domains it makes sense to try to find a model (or a set of models) so that the errors affect the accuracy of the conditional distribution (2) as little as possible, while we can allow the joint probability distribution to be such that the predictions concerning some other variable would be quite inaccurate. For this reason, we suggest that instead of using a single model structure M for determining the distance measure (4), we should use m supervised models $M_1, \ldots, M_m$, each chosen with respect to the corresponding predictive task. We return to this issue in Section 5.

5. EMPIRICAL RESULTS

5.1 The setup

To illustrate the validity of the suggested data reduction scheme, we performed a series of experiments with 20 publicly available classification data sets from the UCI data repository [1]. In the preprocessing phase of the experimental setup, all continuous attributes in the data sets were discretized by using a straightforward application of the k-means algorithm. Consequently, with respect to the empirical study reported here, all the data sets were discrete.

When producing a visualization of a data set $\mathbf{X}$, the pairwise distances between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ were determined by using the unsupervised distance metric (4). This requires determining m predictive distributions $P(X_k \mid \mathbf{x}_i^-, M_k)$. Unfortunately, as discussed in [6, 4, 12], finding accurate Bayesian network models for supervised prediction tasks is a difficult problem. On the other hand, as demonstrated in, for example, [17, 10], the structurally simple Naive Bayes classifier performs surprisingly well in many real-world classification domains, despite of the fact that the model is extremely fast to construct and use. For this reason, in this series of experiments the predictive models M_k used in the visualization scheme were Naive Bayes models, constructed with respect to the target variable X_k.

For constructing a visualization $\tilde{\mathbf{X}}$ for data $\mathbf{X}$, in the approach presented here we need an algorithm for finding a visualization $\tilde{\mathbf{X}}$ so that the objective function (1) is minimized. This objective function is typically quite complex and finding the optimal visualization in this sense is not possible in practice, so we are left with approximative solutions. However, it should be pointed out that in this context it is not necessary to aim at the absolutely optimal visualization $\tilde{\mathbf{X}}$ — for visualization purposes a reasonable approximation is usually quite sufficient. How to find effectively good approximations of the optimal visualization is however a wide

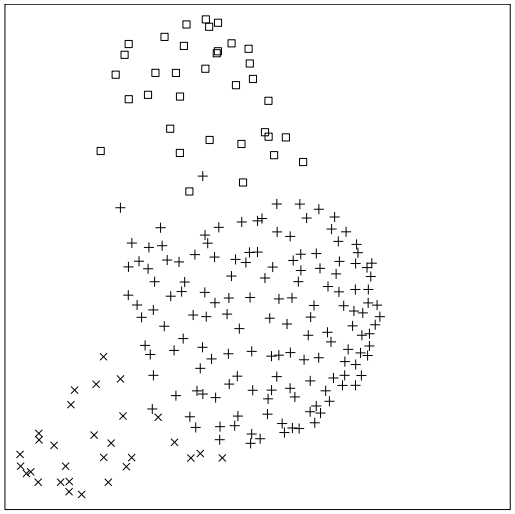

Figure 1: The Thyroid Disease data set: an example of the unsupervised visualizations obtained with the suggested method.

research problem on its own, and is not discussed in detail here. In the experiments reported here we used a simple iterative stochastic greedy algorithm where at each step the visual locations of two randomly chosen data vectors are optimized along the connecting line so that the objective function (1) is optimized locally.

The practical relevance of a visualization $\tilde{\mathbf{X}}$ can be indirectly measured through a data mining process, where domain experts try to capture interesting regularities from the visual image. In our case, however, no such domain experts were available, so we had to evaluate our visualization technique in a different manner. In this set of experiments, this was done by assuming that the clustering provided by the class labels in the UCI data sets is a reasonable clustering, in the sense that this clustering can be regarded as something that we should come up with, had we not seen the class labels originally. Following this line of reasoning, each classification data set was first pruned by removing the class labels, i.e., the column containing the values of the class variable X_m. The remaining data $\mathbf{X}$ was then visualized in two-dimensional space by using the unsupervised approached presented here. Finally, the produced visual images were colored according to the class labels that were not used at all in the visualization process. If the resulting image was visually pleasing in the sense that the different classes (different colors) were nicely separated in the picture, it can be said that we were able to recover the original clustering in a totally unsupervised manner, without using the class label information.

5.2 The results

The empirical results show that the suggested unsupervised visualization method work very well: most of the produced visual images pass the class coloring clarity test explained above. A library of colored 2D examples of the produced visualizations can be found at URL: http://www.cs.Helsinki.FI/research/cosco/Projects/Visual/KDD2000.

There is however a possible caveat in the above experimental procedure: basically there is no a priori reason why the unsupervised visual image produced should reflect the clustering provided by the class labels, especially if the original clustering is poor from the probabilistic modeling point of view. One way to measure the "goodness" of the clustering provided by the class labels is to evaluate the predictive accuracy of the Naive Bayes model, as this model is essentially based on clustering the data according to the class variable X_m. Leave-one-out crossvalidated classification results of the Naive Bayes classifier can be found in the second column of Table 1.

We can now conjecture that the above "class-color clarity test" for the resulted visual images may fail with data sets where the performance of the Naive Bayes classifier is poor. The experimental results confirm this hypothesis: in cases where the leave-one-out crossvalidated classification accuracy of the NB classifier is poor in the absolute sense (as with the Liver Disorders data set), or in the relative sense with respect to the default classification accuracy (as with the Postoperative Patient data set), the class labeled colored images are somewhat blurred. Nevertheless, we would like to emphasize that this does not mean that the unsupervised visualization technique has "failed" in these cases, but that in these (relatively few) cases the somewhat artificial empirical setup used here is not practically sensible in the first place. This means that if the data would in these cases be clustered according to the visual image produced, this could result in a probabilistic model producing more accuracy predictions than the Naive Bayes classifier. This interesting idea is however not studied further here.

We believe that most people agree that the produced images (with the exception of the few cases discussed above) are visually pleasing in the sense that the original classes are clearly separable in the image. This however raises the question of whether the quality of the visualization could be measured more objectively. Intuitively, we would like to measure how well the data in the visual image is clustered according to the (hidden) class label. We suggest that this can be done by using, for example, a simple k-NN (nearest neighbor) method, where each data point is classified as a member of the class containing the most representatives in the k nearest data points. The results with this type of k-NN (with k=9) classification method are summarized in Table 1. The method 9-NN refers to 9-nearest neighbor classification, where the distances of the (m-1)-dimensional vectors are computed by formula (4). The method 9-NN^2 means the corresponding method with the distances computed by using the Euclidean distance in the 2-dimensional visual image produced by the unsupervised visualization method suggested here.

From Table 1 we can make the following observations. First, the quality of the crossvalidated k-NN classification accuracy does not degrade significantly as we move from the high-dimensional space to the 2-dimensional space. This is further evidence for the fact that the produced visualizations are of good quality. Second, we can see that the classification accuracy is quite comparable (and in some cases even better) to the accuracy of the Naive Bayes classifier, even though the visualization was done in a purely unsupervised

Table 1: Crossvalidated classification results.

Data	Default	NB	9-NN	9-NN^2
Australian Credit	55.5	87.1	82.0	81.2
Breast Cancer (Wisconsin)	65.5	97.4	97.3	96.9
Breast Cancer	70.3	72.3	71.7	74.1
Credit Screening	55.5	86.2	82.9	83.0
Pima Indians Diabetes	65.1	77.9	72.7	72.1
German Credit	70.0	74.9	67.4	66.1
Heart Disease (Cleveland)	54.1	57.8	57.4	57.4
Heart Disease (Hungarian)	63.9	83.3	82.7	81.3
Heart Disease (Statlog)	55.6	85.2	83.7	82.2
Hepatitis	79.4	83.2	82.6	82.6
Ionosphere	64.1	92.9	91.2	87.2
Iris Plant	33.3	94.0	88.7	87.3
Liver Disorders	58.0	63.2	58.8	58.6
Lymphography	54.7	85.8	81.8	78.4
Mole Fever	67.1	87.8	89.2	82.4
Postoperative Patient	71.1	67.8	71.1	68.9
Thyroid Disease	69.8	99.1	97.2	95.3
Vehicle Silhouettes	25.8	64.7	66.5	45.2
Congressional Voting Records	61.4	90.1	88.0	87.1
Wine Recognition	39.9	97.2	96.1	96.1

manner, and classification was not at all the goal of the visualization process. This is even more surprising considering the fact that the 9-NN classifier used here was a random and naive choice for performing the classification, meant to be used only for illustrating the quality of the visual results. Consequently, the results suggest that the distance-based approach used would offer an interesting framework for producing accurate classifiers, if that would be the primary goal of the research.

6. CONCLUSIONS

We have described a data reduction scheme based on the idea of defining a probabilistic distance metric with respect to predictions obtained by a formal, probabilistic domain model. The concrete model used in this paper was a Bayesian network, or more precisely, a pool of supervised Naive Bayes classifiers, although the general approach can be used with any type of a probabilistic model. We gave examples of how this type of distance measures can be defined in both supervised and unsupervised manner. As a concrete application of the suggested scheme, we considered the problem of producing visual images of high-dimensional data. The method suggested in this paper is based on a Sammon's mapping technique with respect to the proposed probabilistic distance measure.

The suggested visualization method was empirically tested by using publicly available UCI data sets. To study how well the resulting visual images can reflect the hidden structure of the data, the class labels were not used at all in the visualization process, and the resulting images were colored afterwards according to this latent information. The results demonstrate that the suggested unsupervised visualization method can be used for effectively discovering the underlying structure of the data. We also discussed more objective methods for measuring the quality of the produced visualizations, and performed experiments by using a simple classification method for this purpose. The results confirm the visual observations about the good quality of the resulting images.

7. ACKNOWLEDGMENTS

This research has been supported by the National Technology Agency (Tekes), and by the Academy of Finland.

8. REFERENCES

[1] C. Blake, E. Keogh, and C. Merz. UCI repository of machine learning databases, 1998. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] C. Chatfield and A. Collins. *Introduction to Multivariate Analysis*. Chapman and Hall, New York, 1980.

[3] R. Duda and P. Hart. *Pattern classification and scene analysis*. John Wiley, 1973.

[4] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[5] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, One Microsoft Way, Redmond, WA 98052, 1996.

[6] D. Heckerman and C. Meek. Models and selection criteria for regression and classification. In D. Geiger and P. Shenoy, editors, *Uncertainty in Arificial Intelligence 13*, pages 223–228. Morgan Kaufmann Publishers, San Mateo, CA, 1997.

[7] F. Jensen. *An Introduction to Bayesian Networks*. UCL Press, London, 1996.

[8] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[9] P. Kontkanen, J. Lahtinen, P. Myllymäki, T. Silander, and H. Tirri. Using Bayesian networks for visualizing high-dimensional data. *Intelligent Data Analysis*, 2000. To appear.

[10] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. BAYDA: Software for Bayesian classification and feature selection. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 254–258. AAAI Press, Menlo Park, 1998.

[11] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On Bayesian case matching. In B. Smyth and P. Cunningham, editors, *Advances in Case-Based Reasoning, Proceedings of the 4th European Workshop (EWCBR-98)*, volume 1488 of *Lecture Notes in Artificial Intelligence*, pages 13–24. Springer-Verlag, 1998.

[12] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On supervised selection of Bayesian networks. In K. Laskey and H. Prade, editors, *Proceedings of the 15th International Conference on Uncertainty in Artificial Intelligence (UAI'99)*, pages 334–342. Morgan Kaufmann Publishers, 1999.

[13] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. On predictive distributions and Bayesian networks. *Statistics and Computing*, 10:39–54, 2000.

[14] S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Stat. Soc., Ser. B*, 50(2):157–224, 1988.

[15] R. Neapolitan. *Probabilistic Reasoning in Expert Systems*. John Wiley & Sons, New York, NY, 1990.

[16] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[17] H. Tirri, P. Kontkanen, and P. Myllymäki. Probabilistic instance-based learning. In L. Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference (ICML'96)*, pages 507–515. Morgan Kaufmann Publishers, 1996.