Predictive Data Mining with Finite Mixtures

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

Abstract

In data mining the goal is to develop methods for discovering previously unknown regularities from databases. The resulting models are interpreted and evaluated by domain experts, but some model evaluation criterion is needed also for the model construction process. The optimal choice would be to use the same criterion as the human experts, but this is usually impossible as the experts are not capable of expressing their evaluation criteria formally. On the other hand, it seems reasonable to assume that any model possessing the capability of making good predictions also captures some structure of the reality. For this reason, in predictive data mining the search for good models is guided by the expected predictive error of the models. In this paper we describe the Bayesian approach to predictive data mining in the finite mixture modeling framework. The finite mixture model family is a natural choice for domains where the data exhibits a clustering structure. In many real world domains this seems to be the case, as is demonstrated by our experimental results on a set of public domain databases.

Introduction

Data mining aims at extracting useful information from databases by discovering previously unknown regularities from data (Fayyad *et al.* 1996). In the most general context, finding such interesting regularities is a process (often called knowledge discovery in databases) which includes the interpretation of the extracted patterns based on the domain knowledge available. Typically the pattern extraction phase is performed by a structure searching program, and the interpretation phase by a human expert. The various proposed approaches differ in the representation language for the structure to be discovered (association rules (Agrawal *et al.* 1996), Bayesian networks (Spirtes, Glymour, & Scheines 1993), functional dependencies (Mannila & Räihä 1991), prototypes (Hu & Cercone 1995) etc.), and in the search methodology used for discovering such structures. A large body of the data mining research is exploratory in nature, i.e., search for any kind of structure in the database in order to understand the domain better.

Akin to the practice of multivariate exploratory analysis in social sciences (Basilevsky 1994), much of the work in the data mining area relies on a task-specific expert assessment of the model goodness. We depart from this tradition, and assume that the discovery process is performed with the expected prediction capability in mind. Consequently, we are trying to answer the question "Which of the models best explains a given database?" by addressing the (in many practical cases more pertinent) question "Which of the models yields the best predictions for future observations from the same process which generated the given database?" In our work the evaluation criteria in the model construction process is based directly on the expected predictive capability of the models, not on more implicit criteria embedded in the search algorithm. The use of predictiveness as a model selection criteria can be justified by the observation that a model with a good predictive capability must have captured some regularities that also reflect properties of the data generating process. We call this approach *predictive data mining*. Predictive data mining is relevant in a wide variety of application areas from credit card fraud detection and sales support systems to industrial process control. Our current work is motivated by large scale configuration problems (e.g., building large generators) where properties of new configurations can be predicted using the regularities in the existing configuration database.

For estimating the expected predictive performance, there exist theoretical measures (see e.g., (Wallace & Freeman 1987; Rissanen 1989; Raftery 1993)) which offer a solid evaluation criterion for the models, but such measures tend to be hard to compute for high-dimensional spaces. In the case of large databases several approximations to these criteria could be used, but many of them are inaccurate with small databases as pointed out in (Kontkanen, Myllymäki, & Tirri 1996a). Alternatively we can choose some prediction problem, and evaluate prediction error empirically by using the available database. An example of such a prediction task would be to predict an unknown attribute value of a data item, given a set of some other instantiated attributes. It should be observed

that we do not assume that the set of predicted attributes are fixed in advance during the discovery process — prediction can be seen as a pattern completion task, where the errors in incomplete pattern completion can be used as a *model measure* for the goodness of the model. In this work we adopt the empirical approach and use the crossvalidation method (Stone 1974; Geisser 1975) for model selection on a set of public domain databases.

In the work presented below we have adopted the basic concepts from the general framework of exploring computational models of scientific discovery (Shrager & Langley 1990). Given a database, we do not attempt to discover arbitrary structures, but restrict the possible patterns (models) to be members of a predefined set, which we call *the model space.* Examples of such model spaces are the set of all possible association rules with a fixed set of attributes, or a set of all *finite mixture distributions* (Everitt & Hand 1981; Titterington, Smith, & Makov 1985). A choice of a model space necessarily introduces prior knowledge to the search process. We would like the model space to be simple enough to allow tractable search, yet powerful enough to include models with good prediction capabilities. Therefore in the current work we have restricted ourselves to a simple, computationally efficient set of probabilistic models from the family of finite mixtures. Intuitively this choice reflects our a priori assumption that the real life data is generated by several distinct processes, which is revealed as a cluster structure in the data.

A finite mixture model for a set of random variables is a weighted sum of a relatively small number of independent mixing distributions. The main advantage of using finite mixture models lies in the fact that the computations for probabilistic reasoning can be implemented as a single pass computation (see the next section). Finite mixtures have also a natural means to model multimodal distributions and are universal in the sense that they can approximate any distribution arbitrarily close as long as a sufficient number of component densities can be used. Finite mixture models can also be seen to offer a Bayesian solution to the case matching and case adaptation problems in instance-based reasoning (see the discussion in (Tirri, Kontkanen, & Myllymäki 1996)), i.e., they can also be viewed as a theoretically sound representation language for a "prototype" model space. This is interesting from the a priori knowledge acquisition point of view, since in many cases the domain experts seem to be able to express their expert knowledge very easily by using prototypical examples or distributions, which can then be coded as mixing distributions in our finite mixture framework.

In order to find probabilistic models for making good predictions, we follow the Bayesian approach (Gelman *et al.* 1995; Cheeseman 1995), as it offers a solid theoretical framework for combining both (suitably coded) a priori domain information and information from the sample database in the model construction process. Bayesian approach also makes a clear separation between the search component and the model measure, and allows therefore modular combinations of different search algorithms and model evaluation criteria. Our approach is akin to the AutoClass system (Cheeseman *et al.* 1988), which has been successfully used for data mining problems, such as LandSat data clustering (Cheeseman & Stutz 1996).

In the case of finite mixtures, the model search problem can be seen as searching for the missing values of the unobserved latent clustering variable in the dataset. The model construction process consists of two phases: *model class selection* and *model class parameter selection.* The model class selection can be understood as finding the proper number of mixing distributions, i.e., the number of clusters in the data space, and the model class parameter selection as finding the attribute value probabilities for each mixture component. The model search problem in this framework is only briefly outlined in this paper — a more detailed exposition can be found in (Kontkanen, Myllymäki, & Tirri 1996b; 1996a). One should observe that theoretically the correct Bayesian approach for obtaining maximal predictive accuracy would be to use the sum of outcomes of all the possible different models, weighted by their posterior probability, i.e., in our case a "mixture of all the mixtures". This is clearly not feasible for data mining considerations, since such a model can hardly be given any useful semantic interpretation. We therefore use only a single, maximum a posteriori probability (MAP) model for making predictions. The feasibility of this approach is discussed in (Cheeseman 1995).

Bayesian inference by finite mixture models

In our predictive data mining framework the problem domain is modeled by m discrete random variables $X_1, \ldots, X_m$. A *data instantiation* $\vec{d}$ is a vector in which all the variables X_i have been assigned a value,

$$\vec{d} = (X_1 = x_1, \ldots, X_m = x_m),$$

where $x_i \in \{x_{i1}, \ldots, x_{in_i}\}$. Correspondingly we can view the database D as a *random sample* $(\vec{d}_1, \ldots, \vec{d}_N)$, i.e., a set of N i.i.d. (independent and identically distributed) data instantiations, where each $\vec{d}_j$ is sampled from $\mathcal{P}$, the joint distribution of the variables $(X_1, \ldots, X_m)$.

In our work we assume that the database D is generated by K different mechanisms, which all can have their own distributions, and that each data vector originates from exactly one of these mechanisms. Thus the instantiation space is divided into K *clusters*, each of which consists of the data vectors generated by the corresponding mechanism. From the assumptions above it follows that a natural candidate for a probabilistic model family is the family of *finite mixtures* (Everitt

& Hand 1981; Titterington, Smith, & Makov 1985), where the problem domain probability distribution is approximated as a weighted sum of mixture distributions:

$$(1) \qquad P(\vec{d}) = \sum_{k=1}^{K} \left(P(Y = y_k) P(\vec{d} \mid Y = y_k) \right).$$

Here the values of the discrete *clustering random variable* Y correspond to the separate clusters of the instantiation space, and each mixture distribution $P(\vec{d} \mid Y = y_k)$ models one data generating mechanism. Moreover, we assume that the problem domain data is tightly clustered so that the clusters can actually be regarded as points in the instantiation space, and data vectors belonging to the same cluster represent noisy versions of that (unknown) point. Therefore we can assume that the variables X_i inside each cluster are independent by which (1) becomes

$$\begin{aligned} P(\vec{d}) &= P(X_1 = x_1, \ldots, X_m = x_m) \\ &= \sum_{k=1}^{K} \left(P(Y = y_k) \prod_{i=1}^{m} P(X_i = x_i | Y = y_k) \right). \end{aligned}$$

In our model both the cluster distribution $P(Y)$ and the intra-class conditional distributions $P(X_i|Y = y_k)$ are multinomial. Thus a finite mixture model can be defined by first fixing K, the *model class* (the number of the mixing distributions), and then by determining the values of the model parameters $\Theta = (\alpha, \Phi), \Theta \in \Omega$, where $\alpha = (\alpha_1, \ldots, \alpha_K), \alpha_k = P(Y = y_k)$, and

$$\begin{aligned} \Phi &= (\Phi_{11}, \ldots, \Phi_{1m}, \ldots, \Phi_{K1}, \ldots, \Phi_{Km}), \\ \Phi_{ki} &= (\phi_{ki1}, \ldots, \phi_{kin_i}), \end{aligned}$$

where $\phi_{kil} = P(X_i = x_{il}|Y = y_k)$.

Given a finite mixture model Θ that models the cluster structure of the database, *predictive inference* can be performed in a computationally efficient manner. The Bayesian approach to predictive inference (see e.g., (Bernardo & Smith 1994)) aims at predicting unobserved future quantities by means of already observed quantities. More precisely, let $\mathcal{I} = \{i_1, \ldots, i_t\}$ be the indices of the instantiated variables, and let $\mathcal{X} = \{X_{i_s} = x_{i_s l_s}, s = 1, \ldots, t\}$ denote the corresponding assignments. Now we want to determine the distribution

$$P(X_i = x_{il} | \Theta, \mathcal{X}) = \frac{\sum_{k=1}^{K} \left(\alpha_k \phi_{kil} \prod_{s=1}^{t} \phi_{ki_s l_s} \right)}{\sum_{k=1}^{K} \left(\alpha_k \prod_{s=1}^{t} \phi_{ki_s l_s} \right)}.$$

The conditional predictive distribution of X_i can clearly be calculated in time $\mathcal{O}(Ktn_i)$, where K is the number of clusters, t the number of instantiated variables and n_i the number of values of X_i. Observe that K is usually small compared to the sample size N, and thus the prediction computation can be performed very efficiently (Myllymäki & Tirri 1994).

The predictive distributions can be used for classification and regression tasks. In classification problems, we have a special class variable X_c which is used for classifying data. In more general regression tasks, we have more than one variable for which we want to compute the predictive distribution, given that the values of the other variables are instantiated in advance. As in the configuration problems mentioned earlier, finite mixture models can also be used for finding the most probable value assignment combination for all the uninstantiated variables, given the values of the instantiated variables. These assignment combinations are useful when modeling actual objects such as machines, where probability information is in any case used to select a proper configuration with instantiated values for all the attributes.

Learning finite mixture models from data

In the previous section we described how the prediction of any variable could be made given a finite mixture model. Here we will briefly outline how to learn such models from a given database D. Let $D = (\vec{d}_1, \ldots, \vec{d}_N)$ be a database of size N. By learning we mean here the problem of constructing a single finite mixture model $M_K(\Theta)$ which represents the problem domain distribution $\mathcal{P}$ as accurately as possible in terms of the prediction capability. This learning process can be divided into two separate phases: in the first phase we wish to determine the optimal value for K, the number of mixing distributions (the model class), and in the second phase we wish to find MAP parameter values $\hat{\Theta}$ for the chosen model class.

In the Bayesian framework, the optimal number of mixing distributions (clusters) can be determined by evaluating the posterior probability for each model class $\mathcal{M}_K$ given the data:

$$P(\mathcal{M}_K|D) \propto P(D|\mathcal{M}_K)P(\mathcal{M}_K), K = 1, \ldots, N,$$

where the normalizing constant $P(D)$ can be omitted since we only need to compare different model classes. The number of clusters can safely be assumed to be bounded by N, since otherwise the sample size is clearly too small for the learning problem in question. Assuming equal priors for the model classes, they can be ranked by evaluating the *evidence* $P(D|\mathcal{M}_K)$ (or equivalently the stochastic complexity (Rissanen 1989)) for each model class. This term is defined as a multidimensional integral and it is usually very hard to evaluate, although with certain assumptions, the evidence can in some cases be determined analytically (Heckerman, Geiger, & Chickering 1995; Kontkanen, Myllymäki, & Tirri 1996a). In the experimental results presented in the next section we chose another approach and estimated the prediction error empirically by using the crossvalidation algorithm (Stone 1974; Geisser 1975).

dataset	size	#attrs	#CVfolds	#clusters	success rate (%)
Australian	690	15	10	17	87.2
Diabetes	768	9	12	20	76.8
German credit	1000	21	10	23	74.1
Glass	214	10	7	30	87.4
Heart disease	270	14	9	8	84.8
Hepatitis	150	20	5	9	88.0
Iris	150	5	5	4	97.3
Lymphography	148	19	5	19	86.6
Primary tumor	339	18	10	21	50.4

Table 1: Description of the experiments.

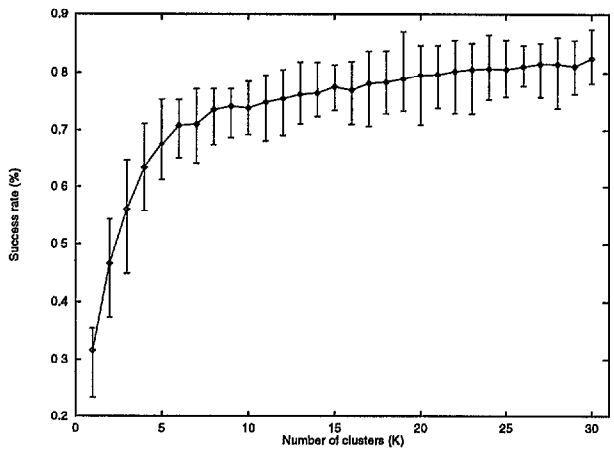

Figure 4: Crossvalidation results with the Glass database.

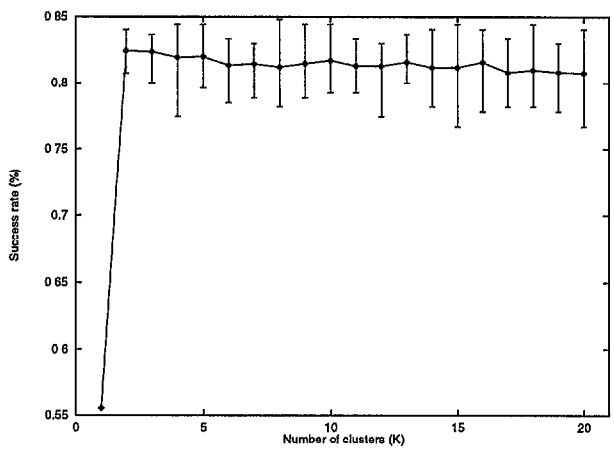

Figure 5: Crossvalidation results with the Heart Disease database.

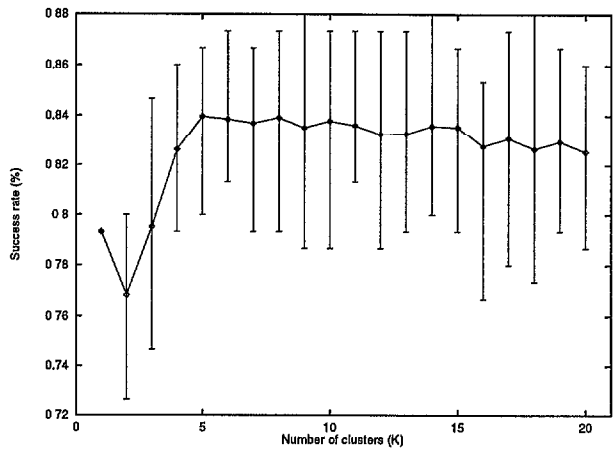

Figure 6: Crossvalidation results with the Hepatitis database.

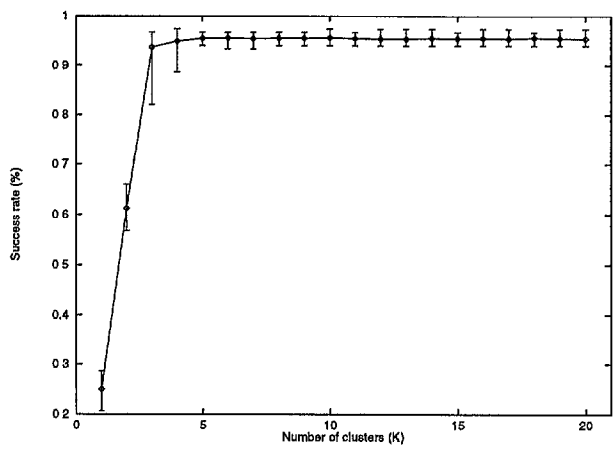

Figure 7: Crossvalidation results with the Iris database.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. 1996. Fast discovery of association rules. In Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining.* MIT Press.

Basilevsky, A. 1994. *Statistical Factor Analysis and Related Methods. Theory and Applicatioms.* New York: John Wiley & Sons.

Bernardo, J., and Smith, A. 1994. *Bayesian theory.* John Wiley.

Cheeseman, P., and Stutz, J. 1996. Bayesian classification (AutoClass): Theory and results. In Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining.* Menlo Park: AAAI Press. chapter 6.

Cheeseman, P.; Kelly, J.; Self, M.; Stutz, J.; Taylor, W.; and Freeman, D. 1988. Autoclass: A Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, 54–64.

Cheeseman, P. 1995. On Bayesian model selection. In Wolpert, D., ed., *The Mathematics of Generalization*, volume XX of *SFI Studies in the Sciences of Complexity.* Addison-Wesley. 315–330.

DeGroot, M. 1970. *Optimal statistical decisions.* McGraw-Hill.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38.

Everitt, B., and Hand, D. 1981. *Finite Mixture Distributions.* London: Chapman and Hall.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds. 1996. *Advances in Knowledge Discovery and Data Mining.* Cambridge, MA: MIT Press.

Geisser, S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70(350):320–328.

Gelman, A.; Carlin, J.; Stern, H.; and Rubin, D. 1995. *Bayesian Data Analysis.* Chapman & Hall.

Heckerman, D.; Geiger, D.; and Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20(3):197–243.

Hu, X., and Cercone, N. 1995. Rough sets similarity-based learning from databases. In Fayyad, U., and Uthurusamy, R., eds., *Proceedings of the First International Conference on Knowledge Discovery & Data Mining*, 162–167.

Kontkanen, P.; Myllymäki, P.; and Tirri, H. 1996a. Comparing Bayesian model class selection criteria by discrete finite mixtures. In *Proceedings of the ISIS (Information, Statistics and Induction in Science) Conference.* (To appear.).

Kontkanen, P.; Myllymäki, P.; and Tirri, H. 1996b. Constructing Bayesian finite mixture models by the EM algorithm. Technical Report C-1996-9, University of Helsinki, Department of Computer Science.

Mannila, H., and Räihä, K.-J. 1991. *The design of relational databases.* Addison-Wesley.

Michie, D.; Spiegelhalter, D.; and Taylor, C., eds. 1994. *Machine Learning, Neural and Statistical Classification.* London: Ellis Horwood.

Myllymäki, P., and Tirri, H. 1994. Massively parallel case-based reasoning with probabilistic similarity metrics. In Wess, S.; Althoff, K.-D.; and Richter, M., eds., *Topics in Case-Based Reasoning*, volume 837 of *Lecture Notes in Artificial Intelligence.* Springer-Verlag. 144–154.

Raftery, A. 1993. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Technical Report 255, Department of Statistics, University of Washington.

Rissanen, J. 1989. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company.

Shrager, J., and Langley, P., eds. 1990. *Computational Models of Scientific Discovery and Theory Formation.* San Mateo, CA: Morgan Kaufmann Publishers.

Spirtes, P.; Glymour, C.; and Scheines, R., eds. 1993. *Causation, Prediction and Search.* Springer-Verlag.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)* 36:111–147.

Tirri, H.; Kontkanen, P.; and Myllymäki, P. 1996. Probabilistic instance-based learning. In Saitta, L., ed., *Machine Learning: Proceedings of the Thirteenth International Conference (to appear).* Morgan Kaufmann Publishers.

Titterington, D.; Smith, A.; and Makov, U. 1985. *Statistical Analysis of Finite Mixture Distributions.* New York: John Wiley & Sons.

Wallace, C., and Freeman, P. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society* 49(3):240–265.