Mixture models and frequent sets: combining global and local methods for 0–1 data

Jaakko Hollmén* Jouni K. Seppänen* Heikki Mannila*

Abstract

We study the interaction between global and local techniques in data mining. Specifically, we study the collections of frequent sets in clusters produced by a probabilistic clustering using mixtures of Bernoulli models. That is, we first analyze 0–1 datasets by a global technique (probabilistic clustering using the EM algorithm) and then do a local analysis (discovery of frequent sets) in each of the clusters. The results indicate that the use of clustering as a preliminary phase in finding frequent sets produces clusters that have significantly different collections of frequent sets. We also test the significance of the differences in the frequent set collections in the different clusters by obtaining estimates of the underlying joint density. To get from the local patterns in each cluster back to distributions, we use the maximum entropy technique [17] to obtain a local model for each cluster, and then combine these local models to get a mixture model. We obtain clear improvements to the approximation quality against the use of either the mixture model or the maximum entropy model.

1 Introduction

Data mining literature contains examples of at least two research traditions. The first tradition, probabilistic modeling, views data mining as the task of *approximating the joint distribution*. In this tradition, the idea is to develop modeling and description methods that incorporate an understanding of the generative process producing the data; thus the approach is global in nature. The other tradition can be summarized by the slogan: *data mining is the technology of fast counting* [12]. The most prominent example of this type of work are association rules [1]. This tradition typically aims at discovering frequently occurring patterns. Each pattern and its frequency indicate only a local property of the data, and a pattern can be understood without having information about the rest of the data.

In this paper we study the interaction between global and local techniques. The general question we are interested in is whether global and local analysis methods can be combined to obtain significantly different sets of patterns. Specifically, we study the collections of frequent sets in clusters produced by a probabilistic clustering using mixtures of Bernoulli models. Given the dataset, we first build a mixture model of multivariate Bernoulli distributions using the EM algorithm, and use this model to obtain a clustering of the observations. Within each cluster, we compute frequent sets, i.e., sets of columns whose value is often 1 on the same row.

These techniques, association rules (frequent sets) and mixture modeling, are both widely used in data mining. However, their combination does not seem to have attracted much attention. The techniques are, of course, different. One is global, and the other is local. Association rules are clearly asymmetric with respect to 0s and 1s, while from the point of view of mixtures of multivariate Bernoulli distributions the values 0 and 1 have equal status. Our study aims at finding out whether there is something interesting to be obtained by combining the different techniques.

The clusters can be considered as potentially interesting subsets of the data, and frequent sets (or association rules computed from them) could be shown to the users as a way of characterizing the clusters. Our main interest is in finding out whether we can quantify the differences in the collections of frequent sets obtained for each cluster. We measure the difference in the collections by considering a simple distance function between collections of frequent sets, and comparing the observed values against the values obtained for randomly chosen clusters of observations. The results show that the patterns explicated by the different clusters are quite different from the ones witnessed in the whole data set. While this is not unexpected, the results indicate that the power of frequent set techniques can be improved by first doing a global partition of the data set.

To get some insight into how different the total information given by the frequent set collections is, we go back from the patterns to distributions by using the maximum entropy technique as in [17]. Given a collection of frequent sets, this technique builds a distribution

*Helsinki Institute for Information Technology, Basic Research Unit, and Laboratory of Computer and Information Science, P.O. Box 5400, 02015 Helsinki University of Technology, Finland
[Jaakko.Hollmen,Jouni.Seppanen,Heikki.Mannila]@hut.fi

that has the same frequent sets (and their frequencies) and has maximal entropy among the distributions that have this property. In this way, we can obtain a distribution from each collection of frequent sets. These distributions can be combined to a mixture distribution by using the original mixture weights given by the EM clustering. We can then measure the distance of this mixture distribution from the original, empirical data distribution using either the Kullback-Leibler or the L_1 distance. The drawback of this evaluation method is that as the maximum entropy technique has to construct the distribution explicitly, the method is exponential in the number of variables, and hence can be used only for a small number of variables. Nevertheless, the results show that the use of collections of frequent sets obtained from the clusters gives us very good approximations for the joint density.

2 Probabilistic modeling of binary data

To model multivariate binary data $\mathbf{x} = (x_1, \ldots, x_d)$ with a probabilistic model, we assume independence between observations and arrive at the multivariate Bernoulli distribution $P(\mathbf{x}|\theta) = \prod_{k=1}^{d} \theta_k^{x_k}(1-\theta_k)^{1-x_k}$. The independence assumption is very strong, however, and quite unrealistic in many situations. *Finite mixtures* of distributions provide a flexible method to model statistical phenomena, and have been used in various applications [13, 8]. A (finite) mixture is a weighted sum of component distributions $P(\mathbf{x}|\theta_j)$, weights or mixing proportions π_j satisfying $\pi_j \geq 0$ and $\sum \pi_j = 1$. A finite mixture of multivariate Bernoulli probability distributions is thus specified by the equation

$$P(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{J} \pi_j P(\mathbf{x}|\theta_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i}(1-\theta_{ji})^{1-x_i}$$

with the parameterization $\mathbf{\Theta} = \{\pi_1, \ldots, \pi_J, (\theta_{ji})\}$ containing $J(d+1)$ parameters for data with d dimensions.

Given a data set R with d binary variables and the number J of mixture components, the parameter values of the mixture model can be estimated using the Expectation Maximization (EM) algorithm [6, 19, 14]. The EM algorithm has two steps which are applied alternately in an iterative fashion. Each step is guaranteed to increase the likelihood of the observed data, and the algorithm converges to a local maximum of the likelihood function [6, 21]. Each component distribution of the mixture model can be seen as a cluster of data points; a point is associated with the component that has the highest posterior probability.

While the mixture modeling framework is very powerful, care must be taken in using it for data sets with high dimensionality. The solutions given by the EM algorithm are seldom unique. Much work has been done recently in improving the properties of the methods and in generalizing the method; see, e.g., [4, 15, 20, 7].

3 Frequent itemsets and maximum entropy distributions

We now move to the treatment of local patterns in large 0–1 datasets. Let R be a set of n observations over d variables, each observation either 0 or 1. For example, the variables can be the items sold in a supermarket, each observation corresponding to a basket of items bought by a customer. If many customers buy a set of items, we call the set frequent; in the general case, if some variables have value 1 in at least a proportion σ of observations, they form a frequent (item)set. The parameter σ must be chosen so that there are not too many frequent sets. Efficient algorithms are known for mining frequent itemsets [2, 9, 10].

Frequent sets are the basic ingredients in finding association rules [1]. While association rules have been a very popular topic in data mining research, they have produced fewer actual applications. One reason for this is that association rules computed from the whole dataset tend to give only fairly general and vague information about the interconnections between variables. In practice, one often zooms into interesting subsets. While association rules are fairly intuitive, in many applications the domain specialists are especially interested in the frequently occurring patterns, not just on the rules (see, e.g., [11]).

By themselves, the frequent sets provide only local information about the dataset. Given a collection of frequent sets and their frequencies, we can, however, use the maximum entropy principle in a similar way as in [16, 17] to obtain a global model. A model is a joint distribution of the variables, and in general there are many distributions that can explain the observed frequent sets. From these distributions, we want to find the one that has the maximum entropy. There is a surprisingly simple algorithm for this called *iterative scaling* [5, 18]. A drawback is that the joint distribution, when represented explicitly, consists of 2^d numbers.

4 Experimental data

We consider three data sets. The first data set *Checker* has $d = 9$ and $n = 10^4$, and the generative distribution is a mixture of 6 Bernoulli distributions with varying mixture proportions $P(j) \propto j$. The Bernoulli distributions form 3 horizontal and 3 vertical bars in the 3×3 grid of the 9 variables. To add noise to the data, ones were observed with probability 0.8 and in the clusters and with probability 0.2 elsewhere.

The second data set is a subset of the *Reuters-21578* data collection, restricted to the words occurring in at least 50 of the documents ($d = 3310, n = 19043$). The third data set is the so called *Microsoft Web data* [3] that records users' visits on a collection of Web pages ($d = 285, n = 32711$).

5 Frequent sets in clusters: comparisons of the collections

Recall that our basic goal is to find out whether one could obtain useful additional information by using the frequent set discovery methods only after the data has been clustered. To test this hypothesis, we first cluster the data with the mixture model to k cluster sets, with methods described in Section 2. Then we calculate the collections of frequent sets separately for each cluster using some given threshold σ. This gives us k collections of frequent sets. To compare two different collections $\mathcal{F}_1$ and $\mathcal{F}_2$ of frequent sets, we define a dissimilarity measure that we call *deviation*,

$$d(\mathcal{F}_1, \mathcal{F}_2) = \frac{1}{|\mathcal{F}_1 \cup \mathcal{F}_2|} \sum_{I \in \{\mathcal{F}_1 \cup \mathcal{F}_2\}} |f_1(I) - f_2(I)|.$$

Here, we denote by $f_j(I)$ the frequency of the set I in $\mathcal{F}_j$, or σ if $I \notin \mathcal{F}_j$. The deviation is in effect an L_1 distance where missing values are replaced by σ.

We computed the mean deviation between each cluster and the whole dataset, varying the number of clusters k and the value of the support threshold σ. Moreover, we compared the results against the collections of frequent sets obtained by taking a random partitioning of the observations into k groups, of the same size as the clusters, and then computing the frequent set collections. The results are shown in Figure 1 for the Checker dataset. Other datasets exhibited similar behavior.

The results show that indeed the patterns explicated by the different clusters are quite different from the ones witnessed in the whole data set. The size of the difference was tested for statistical significance by using a randomization method, which shows a clear separation between the differences among the true clusters and random clusters. In 100 randomization trials the differences were never larger than in the true data. While this is not unexpected, one should note that the average error in the frequency of a frequent set is several multiples of the frequency threshold: the collections of frequent sets are clearly quite different. The results indicate that the power of frequent set techniques can be improved by first doing a global partition of the data set.

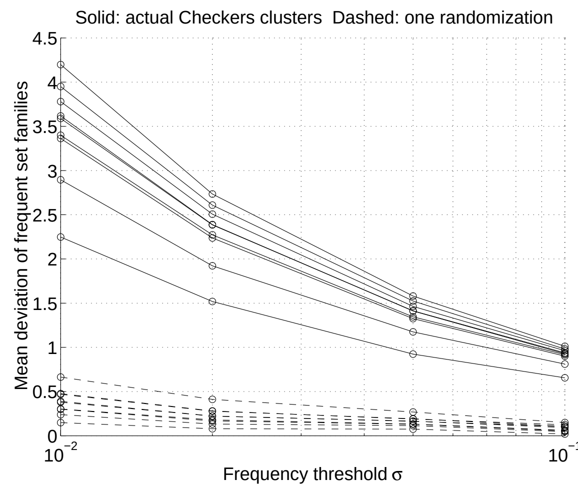

Figure 1: Mean of the difference in the Checker data between the frequencies of the frequent sets in the whole data set compared against the frequency in each of the clusters, divided by the support threshold. X-axis: frequency threshold σ. Y-axis: the difference in multiples of σ. Frequencies of sets that are not frequent on one of the datasets are approximated using $f = \sigma$. Dashed values: results of a single randomization round.

6 Maximum entropy distributions from frequent sets in the clusters

The comparison of collections of frequent sets given in the previous section shows that the collections are quite different. We would also like to understand how much the collections contain information about the joint distribution. For this, we need to be able to obtain a distribution from the frequent sets.

We go back from the patterns to distributions by using the maximum entropy method as in [16, 17]. Given a collection of frequent sets, this technique builds a distribution that has the same frequent sets (and their frequencies) and has maximum entropy among the distributions that have this property. In this way, we can obtain a distribution from each collection of frequent sets. These distributions can be combined to a mixture distribution by using the original mixture weights given by the EM estimation. We can then measure the distance of this mixture distribution from the original, empirical data distribution.

The drawback of this evaluation method is that as the maximum entropy technique has to construct the distribution explicitly, the method is exponential in the number of variables, and hence can be used only for a small number of variables. In the experiments we used the 9 most commonly occurring variables in both the

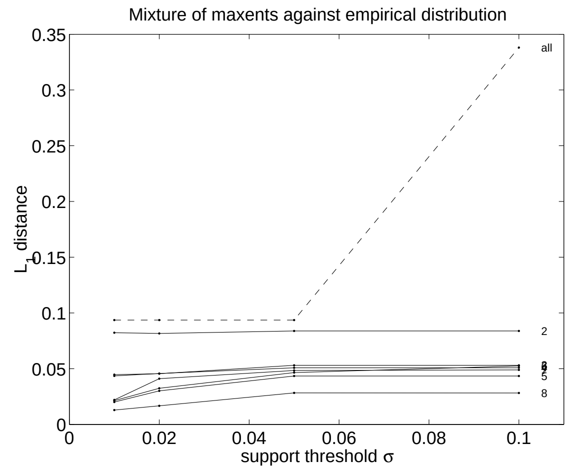

Figure 2: Reuters data, 9 most frequently occurring variables: L1-distance for mixtures of maxents with empirically estimated mixing proportions. X-axis: the frequency threshold. Y-axis: the L1-distance between the mixture distribution and the empirical distribution in the data.

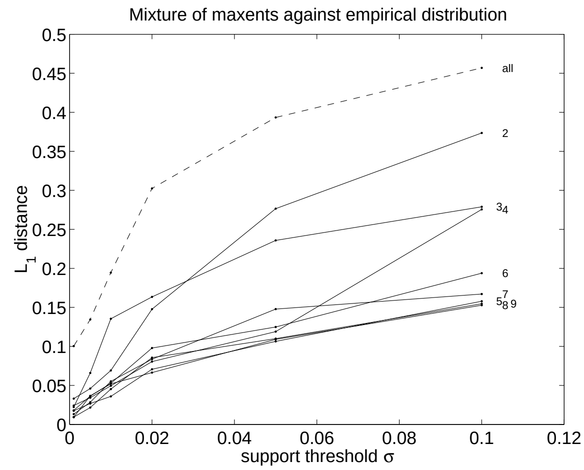

Figure 3: Microsoft web data, 9 most frequently occurring variables: L1-distance for mixtures of maxents with empirically estimated mixing proportions. X-axis: the frequency threshold. Y-axis: the L1-distance between the mixture distribution and the empirical distribution in the data.

Reuters and the Microsoft web datasets. The results are shown in Figure 2 for the Reuters dataset and in Figure 3 for the Microsoft web dataset. The Checker dataset exhibited similar behavior and is omitted.

In the figures the x-axis corresponds to the frequency threshold σ for the frequent set computation, and the y-axis shows the L_1 distance $\sum_x |g(x) - f(x)|$, where x is a 0–1 vector of length d, g is the maxent mixture, and f is the "real" distribution from which the data were generated. With the Kullback-Leibler measure $E_g[\log(g/f)] = \sum_x g(x) \log(g(x)/f(x))$, the resuls were similar and are omitted here.

We also compared the approximation of the joint distribution given by the initial mixture model against the empirical distribution; this distance is not dependent on the frequency threshold used. The results were in most cases clearly inferior to the others, and are therefore not shown.

We observe that the distance from the "true" empirical distribution is clearly smaller when we use the mixture of maximum entropy distributions obtained from the frequent sets of the clusters. The effect is especially strong in the case of the Microsoft web data. The results show that the use of collections of frequent sets obtained from the clusters gives us very good approximations for the joint density.

One could, of course, use out-of-sample likelihood techniques or BIC-type of methods to test whether the number of extra parameters involved in the mixtures of maximum entropy distributions is worth it. Our goal in this paper is, however, only to show that the clusters of observations produced by EM clustering do indeed have significantly different collections of frequent sets.

7 Summary

We have studied the combination of mixture modeling and frequent sets in the analysis of 0–1 datasets. We used mixtures of multivariate Bernoulli distributions and the EM algorithm to cluster datasets. For each resulting cluster, we computed the collection of frequent sets. We computed the distances between the collections by an L_1-like measure. We also compared the information provided by the collections of frequent sets by computing maximum entropy distributions from the collections and combining them to a mixture model.

The results show that the information in the frequent set collections computed from clusters of the data is clearly different from the information given by the frequent sets on the whole data collection. One could view the result as unsurprising. Indeed, it is to be expected that computing a larger number of results (k collections of frequent sets instead of one) gives more information. What is noteworthy in the results is that the differences between the frequencies of the frequent sets are so large (see Figure 1). This indicates that the global technique of mixture modeling finds features that can actually be made explicit by looking at the frequent sets in the clusters.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, *Mining association rules between sets of items in large databases*, in Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'93), P. Buneman and S. Jajodia, eds., Washington, D.C., USA, May 1993, ACM, pp. 207 – 216.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, *Fast discovery of association rules*, in Advances in Knowledge Discovery and Data Mining, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., AAAI Press, Menlo Park, CA, 1996, pp. 307 – 328.

[3] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, *Visualization of navigation patterns on a web site using model-based clustering*, in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, R. Ramakrishnan, S. Stolfo, R. Bayardo, and I. Parsa, eds., 2000, pp. 280–289.

[4] I. V. Cadez, S. Gaffney, and P. Smyth, *A general probabilistic framework for clustering individuals and objects*, in KDD 2000, 2000, pp. 140–149.

[5] J. Darroch and D. Ratcliff, *Generalized iterative scaling for log-linear models*, The Annals of Mathematical Statistics, 43 (1972), pp. 1470–1480.

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

[7] J. G. Dy and C. E. Brodley, *Feature subset selection and order identification for unsupervised learning*, in Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000, pp. 247–254.

[8] B. Everitt and D. Hand, *Finite Mixture Distributions*, Monographs on Applied Probability and Statistics, Chapman and Hall, 1981.

[9] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000.

[10] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, 2001.

[11] K. Hatonen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen, *TASA: Telecommunication alarm sequence analyzer, or "How to enjoy faults in your network"*, in Proceedings of the 1996 IEEE Network Operations and Management Symposium (NOMS'96), Kyoto, Japan, Apr. 1996, IEEE, pp. 520 – 529.

[12] H. Mannila, *Global and local methods in data mining: basic techniques and open problems*, in ICALP 2002, 29th International Colloquium on Automata, Languages, and Programming, 2002.

[13] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics, John Wiley & Sons, 2000.

[14] G. J. McLachlan, *The EM Algorithm and Extensions*, Wiley & Sons, 1996.

[15] C. Ordonez, E. Omiecinski, and N. Ezquerra, *A fast algorithm to cluster high dimensional basket data*, in ICDM, 2001, pp. 633–636.

[16] D. Pavlov, H. Mannila, and P. Smyth, *Probabilistic models for query approximation with large sparse binary data sets*, in UAI-2000, 2000, pp. 465–472.

[17] D. Pavlov, H. Mannila, and P. Smyth, *Beyond independence: Probabilistic models for query approximation on binary transaction data*, Tech. Report ICS TR-01-09, Information and Computer Science Department, UC Irvine, 2001. To appear in IEEE TDKE.

[18] S. D. Pietra, V. J. D. Pietra, and J. D. Lafferty, *Inducing features of random fields*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (1997), pp. 380–393.

[19] R. Redner and H. Walker, *Mixture densities, maximum likelihood and the EM algorithm*, SIAM Review, 26 (1984), pp. 195–234.

[20] A. K. H. Tung, J. Han, V. S. Lakshmanan, and R. T. Ng, *Constraint-based clustering in large databases*, Lecture Notes in Computer Science, 1973 (2001), pp. 405–419.

[21] C. J. Wu, *On the convergence properties of the EM algorithm*, The Annals of Statistics, 11 (1983), pp. 95–103.