A FRAMEWORK FOR MDL CLUSTERING

Petri Myllymäki

Helsinki Institute for Information Technology HIIT
P.O.Box 68, Department of Computer Science
FIN-00014 University of Helsinki, FINLAND
petri.myllymaki@cs.helsinki.fi

1. INTRODUCTION

Data clustering is one of the central concepts in the field of unsupervised data analysis and machine learning, but it is also a surprisingly controversial issue, and the very meaning of the concept "clustering" may vary a great deal between different scientific disciplines (see, e.g., [1] and the references therein). However, a common goal in all cases is that the objective is to find a structural representation of data by grouping (in some sense) similar data items together. In our work we have focused on non-hierarchical (flat) clustering, where clustering is regarded as a partitional data assignment or data labeling problem, and the goal is to partition the data into mutually exclusive clusters so that similar (in a sense that needs to be defined) data items are grouped together. The number of clusters is unknown, and determining the optimal number is part of the clustering problem. The data are assumed to be in a vector form so that each data item is a vector consisting of a fixed number of attribute values.

We can now identify two fundamental problems within this framework:

1. Scoring: what constitutes a reasonable optimality or "goodness" criterion; when is one clustering (data partitioning) better than another, and in what sense?

2. Optimization: how to find good clusterings with respect to the chosen scoring criterion?

Traditionally, the scoring problem has been approached by first fixing a distance metric, and then by defining a global goodness criterion based on pairwise distances of data items. However, while this approach can be intuitively appealing, from the theoretical point of view it introduces many problems, such as choosing a suitable distance metric and the handling of non-continuous attributes. An attempt to construct an axiomatic formalization for the desirable properties of this type of similarity-based clustering methods is given in [2] . Cilibrasi and Vitányi [3] on the other hand suggest a clustering method based on the universal *Normalized Compression Distance (NCD)* metric: the intuitive idea is that two data objects are close to each other if they can be compressed well together. In [4] we used a similar approach, but instead of thinking in terms of pairwise distances, we applied it to sets of data objects, so that a set of objects is clustered together if the items in the set can be compressed well together.

For encoding the data vectors inside a cluster, we use probabilistic parametric models. This can be seen as a *model-based clustering* approach, where for each cluster a data generating function (a probability distribution) is assumed, and the clustering problem is defined as the task to identify these distributions (see, e.g., [5, 6, 7]). In other words, the data are assumed to be generated by a sum of distributions, a *finite mixture model* [8, 9, 10]. In this framework the optimality of a clustering can be defined as a function of the fit of data with the finite mixture model, not as a function of the distances between the data vectors. See [4] for more discussion on the differences between similarity-based and model-based clustering approaches.

It should be noted that in the model-based approach we implicitly introduce "hidden data": for each data vector, we need to identify which model/distribution/code to use (to which cluster the data vector belongs?). This means that we need to encode the cluster labels (the "model index") *together* with the actual observed data so that the resulting total code length is minimized. The clustering criterion suggested [4] is based on the MDL principle [11, 12, 13] which aims at finding the shortest possible encoding for the data. For formalizing this intuitive goal, we adopted the *normalized maximum likelihood (NML)* coding approach [14], which can be shown to lead to a criterion with very desirable theoretical properties (see e.g. [13, 15, 16, 17, 18, 19]). Similar ideas have been explored in [20], and an application in signal denoising can be found in [21]. It should be noted that approaches based on either earlier formalizations of MDL or on more heuristic encoding schemes (see e.g. [22, 23, 24]) do not possess these theoretical properties.

In Section 2 we briefly summarize the main idea behind the scoring criterion suggested in [4], and in Section 2 discuss the optimization methods explored in [25].

2. SCORING

Let us assume that our problem domain consists of m discrete variables $X_1, \ldots, X_m$ and that the variable X_i has K_i values. The data $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ consists of

i.i.d. observations $\mathbf{x}_j = (x_{j0}, x_{j1}, \ldots, x_{jm}) \in \mathcal{X}$, where

$$\mathcal{X} = \{1, 2, \ldots, K_1\} \times \cdots \times \{1, 2, \ldots, K_m\}. \quad (1)$$

We assume that the possibly originally continuous variables have been discretized. One reason for focusing on discrete data is that in this case we can model the domain variables by multinomial distributions without having to make restricting assumptions about unimodality, normality etc., which is the situation we face in the continuous case.

Formally, we can notate a clustering by using a *clustering vector* $\mathbf{z}^n = (z_1, \ldots, z_n)$, where z_j denotes the index of the cluster to which the data vector $\mathbf{x}_j$ is assigned to. Denote the *clustering variable* by Z so that $\mathbf{z}^n$ is a sample from the distribution of Z. The number of clusters, say K, is implicitly defined in the clustering vector, as it can be determined by counting the number of different values appearing in $\mathbf{z}^n$. It is reasonable to assume that K is bounded by the size of our data set, so we can define the *clustering space* $\mathcal{Z}$ as the set containing all the clusterings $\mathbf{z}^n$ with the number of clusters being less or equal to n. Hence the clustering problem is now to find from all the $\mathbf{z}^n \in \mathcal{Z}$ the optimal clustering $\mathbf{z}^n$.

We now have to define what type of probabilistic models we aim to use for encoding our data. By introducing the cluster label Z we have implicitly assumed a *finite mixture model* where $P(\mathbf{x}) = \sum_Z P(\mathbf{x} \mid Z)P(Z)$. For the class-conditional probability distributions $P(\mathbf{x} \mid Z)$ we can of course consider any model, but as our data was assumed to be i.i.d., it is natural to consider models defined by independencies between our variables $X_1, \ldots, X_m$. The simplest of such models assumes no dependencies, and this is the model we have mostly used. This means that in this special case we use the same parametric model (form) for each cluster, but they each have separate sufficient statistics (determined by data vectors assigned to the cluster).

A motivating reason for using this simple model is that it is now easy to "explain" each cluster: as the variables are independent within a cluster, we can study the variable distributions separately of each other. There are also computational reasons, as the joint probability distribution now factorizes conveniently as a product (of multinomials):

$$P_{\text{FM}}(Z = z, X_1 = x_1, \ldots, X_m = x_m \mid \boldsymbol{\theta}) = P(Z = z \mid \boldsymbol{\theta}) \cdot \prod_{i=1}^{m} P(X_i = x_i \mid Z = z, \boldsymbol{\theta}). \quad (2)$$

It should be noted that although we assumed local (class-conditional) independence, the model allows variables to be globally dependent (when the value of the latent cluster variable is not known). Actually, we can clearly represent probability distributions of arbitrary complexity by just adding more mixture components.

Now given $\mathcal{M}(K)$, a finite mixture model with K component distributions (each assuming local independence), the normalized maximum likelihood of our observed data $\mathbf{x}^n$, together with the cluster indexes $\mathbf{z}^n$, is given by

$$P_{\text{NML}}(\mathbf{x}^n, \mathbf{z}^n \mid \mathcal{M}(K)) = \frac{P(\mathbf{x}^n, \mathbf{z}^n; \hat{\theta}(\mathbf{x}^n, \mathbf{z}^n), \mathcal{M}(K))}{\mathcal{C}(\mathcal{M}(K), n)}, \quad (3)$$

where $\mathcal{C}(\mathcal{M}(K), n)$ is the parametric complexity of the parametric model $\mathcal{M}(K)$ with sample size n, The maximum likelihood term in the numerator is easy to compute:

$$P(\mathbf{x}^n, \mathbf{z}^n; \hat{\theta}(\mathbf{x}^n, \mathbf{z}^n), \mathcal{M}(K)) = \prod_{k=1}^{K} \left(\frac{h_k}{n}\right)^{h_k} \prod_{i=1}^{m} \prod_{l=1}^{K_i} \left(\frac{f_{ikl}}{h_k}\right)^{f_{ikl}}, \quad (4)$$

where h_k is the number of times Z has value k in $\mathbf{z}^n$, f_{ikl} is the number of times X_i has value l when Z has value k. The parametric complexity term in the denominator is in principle an exponential sum,

$$\mathcal{C}(\mathcal{M}(K), n) = \sum_{\mathbf{y}^n} \sum_{\mathbf{v}^n} P(\mathbf{y}^n, \mathbf{v}^n; \hat{\theta}(\mathbf{y}^n, \mathbf{v}^n), \mathcal{M}(K)), \quad (5)$$

but in our case it fortunately can be simplified to

$$\mathcal{C}(\mathcal{M}(K), n) = \sum_{h_1 + \cdots + h_K = n} \frac{n!}{h_1! \cdots h_K!} \prod_{k=1}^{K} \left(\frac{h_k}{n}\right)^{h_k} \prod_{i=1}^{m} \mathcal{C}(K_i, h_k), \quad (6)$$

where the multinomial parametric complexities

$$\mathcal{C}(K_i, h) = \sum_{n_1 + \cdots + n_{K_i} = h} \frac{h!}{n_1! \cdots n_{K_i}!} \prod_{k=1}^{K_i} \left(\frac{n_k}{h}\right)^{n_k}, \quad (7)$$

can be computed efficiently using the recursive formula presented in [26]. All in all, the time complexity of computing the parametric complexity $\mathcal{C}(\mathcal{M}(K), n)$ for a fixed K is $\mathcal{O}\left(n^2 \log K\right)$ (for details, see [27]).

3. OPTIMIZATION

The clustering space $\mathcal{Z}$ is obviously exponential in size, which means that if we want to find a good clustering with respect to the NML clustering criterion presented in the previous section, in practice we need to resort to combinatorial search algorithms. However, it should be noted that if we keep the number of clusters constant, the denominator of the NML criterion (3) does not change, which leaves us with the maximum likelihood term given in the numerator. This means that we can use the toolbox of standard clustering algorithms developed for maximizing the likelihood, vary the number of clusters, and from the candidate solutions produced by these algorithms, pick the one that maximizes the NML criterion. Note that it is obviously easy to get high-likelihood solutions with more

clusters, but the normalizing coefficient of NML also increases with increasing number of clusters, penalizing for complexity.

Standard methods for maximizing the likelihood of a finite mixture model include the Expectation-Maximization (EM) algorithm, and the related K-means algorithm (KM) (or the CEM algorithm [28] as it is sometimes called). Both algorithms are computationally simple, and converge towards a local minimum of the likelihood, so the algorithms have to used iteratively from several different starting points. For more details of the EM algorithm, see e.g. [29, 30].

In [25] we showed how to use the EM and K-means algorithms, and their variants, for optimizing the NML clustering criterion, and empirically studied the performance of different optimization algorithms by using several real-world datasets from the UCI repository [31]. The results suggest that straghtforward application of these algorithms does not give good results, but the new variants developed seem to work much better.

4. CONCLUSIONS

We regarded clustering as a data assignment problem where the goal is to partition the data into non-hierarchical groups of items. We suggested an information-theoretic criterion, based on the minimum description length (MDL) principle, for defining a goodness criterion for this type of clustering of data. The basic idea behind this framework is to optimize the total code length over the data by encoding together data items belonging to the same cluster. In this setting efficient coding is possible only by exploiting underlying regularities that are common to the members of a cluster, which means that this approach produces an implicitly (but not explicitly) defined similarity metric between the data items.

Formally the global code length criterion to be optimized is defined by using the intuitively appealing universal normalized maximum likelihood (NML) code. We have also studied the optimization aspect of the clustering problem, and developed algorithms that can be used for efficiently searching the exponentially-sized clustering space. As the suggested NML clustering criterion can be used for comparing clusterings with different number of cluster labels, the number of clusters does not have to be fixed beforehand and determining it can also be handled as part of the optimization process.

5. ACKNOWLEDGMENTS

This work was supported in part by the Academy of Finland under the project ModeST, and by the IST Programme of the European Community under the PASCAL Network of Excellence.

6. REFERENCES

[1] A.k. Jain, M.N. Murty, and P.J Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[2] R. Zadeh and S. Ben-David, "A uniqueness theorem for clustering," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, J. Bilmes and A. Ng, Eds. AUAI Press, 2009.

[3] R. Cilibrasi and P. Vitányi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4 (April), pp. 1523–1545, 2005.

[4] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri, "An MDL framework for data clustering," in *Advances in Minimum Description Length: Theory and Applications*, P. Grünwald, I.J. Myung, and M. Pitt, Eds. The MIT Press, 2006.

[5] P. Smyth, "Probabilistic model-based clustering of multivariate and sequential data," in *Proceedings of the Seventh International Conference on Artificial Intelligence and Statistics*, D. Heckerman and J. Whittaker, Eds. 1999, pp. 299–304, Morgan Kaufmann Publishers.

[6] Chris Fraley and Adrian E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.

[7] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, "Autoclass: A Bayesian classification system," in *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, June 1988, pp. 54–64.

[8] B.S. Everitt and D.J. Hand, *Finite Mixture Distributions*, Chapman and Hall, London, 1981.

[9] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons, New York, 1985.

[10] G.J. McLachlan, Ed., *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.

[11] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 445–471, 1978.

[12] J. Rissanen, "Stochastic complexity," *Journal of the Royal Statistical Society*, vol. 49, no. 3, pp. 223–239 and 252–265, 1987.

[13] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[14] Yu M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3–17, 1987.

[15] A. Barron, J. Rissanen, and B. Yu, "The minimum description principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, October 1998.

[16] P. Grünwald, *The Minimum Description Length Principle and Reasoning under Uncertainty*, Ph.D. thesis, CWI, ILLC Dissertation Series 1998-03, 1998.

[17] J. Rissanen, "Hypothesis selection and testing by the MDL principle," *Computer Journal*, vol. 42, no. 4, pp. 260–269, 1999.

[18] Q. Xie and A.R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, March 2000.

[19] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1712–1717, July 2001.

[20] P.-H Lai, J. O'Sullivan, and R. Pless, "Minimum description length and clustering with exemplars," in *2009 IEEE International Symposium on Information Theory*, R. Calderbank, H. Chung, and A. Orlitsky, Eds. IEEE, 2009.

[21] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," *IEEE Transactions on Signal Processing (to appear)*, http://arxiv.org/abs/cs/0609138.

[22] B. Dom, "An information-theoretic external cluster-validity measure," Tech. Rep. RJ 10219, IBM Research, 2001.

[23] M. Plumbley, "Clustering of sparse binary data using a minimum description length approach," Tech. Rep., Department of Electrical Engineering, Queen Mary, University of London, 2002, Unpublished manuscript.

[24] M-C. Ludl and G. Widmer, "Clustering criterion based on minimum length encoding," in *Proceedings of the 13th European Conference on Machine Learning*, Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, Eds. 2002, vol. 2430 of *Lecture Notes in Computer Science*, pp. 258–269, Springer.

[25] P. Kontkanen and P. Myllymäki, "An empirical comparison of NML clustering algorithms," in *Proceedings of the 2008 International Conference on Information Theory and Statistical Learning (ITSL-08)*, M. Dehmer, M. Drmota, and F. Emmert-Streib, Eds., pp. 125–131. CSREA Press, 2008.

[26] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.

[27] P. Kontkanen, *Computationally Efficient Methods for MDL-Optimal Clustering and Density Estimation*, Ph.D. thesis, Department of Computer Science, University of Helsinki, 2009 (to appear).

[28] M. Meila and D. Heckerman, "An experimental comparison of several clustering and initialization methods," in *UAI'98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Gregory F. Cooper and Serafín Moral, Eds., 1998, pp. 386–395.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[30] P. Kontkanen, P. Myllymäki, and H. Tirri, "Constructing Bayesian finite mixture models by the EM algorithm," Tech. Rep. NC-TR-97-003, ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT), 1996.

[31] S. Hettich, C.L. Blake, and C.J. Merz, "UCI repository of machine learning databases, University of California, Irvine, Dept. of Information and Computer Sciences," 1998, http://www.ics.uci.edu/∼mlearn/MLRepository.html.