Proc. IDEAL 2000, in press

Clustering by Similarity in an Auxiliary Space

Janne Sinkkonen and Samuel Kaski

Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Finland
janne.sinkkonen@hut.fi, samuel.kaski@hut.fi

Abstract. We present a clustering method for continuous data. It defines local clusters into the (primary) data space but derives its similarity measure from the posterior distributions of additional discrete data that occur as pairs with the primary data. As a case study, enterprises are clustered by deriving the similarity measure from bankruptcy sensitivity. In another case study, a content-based clustering for text documents is found by measuring differences between their metadata (keyword distributions). We show that minimizing our Kullback–Leibler divergence-based distortion measure within the categories is equivalent to maximizing the mutual information between the categories and the distributions in the auxiliary space. A simple on-line algorithm for minimizing the distortion is introduced for Gaussian basis functions and their analogs on a hypersphere.

1 Introduction

Clustering by definition produces localized groups of items, which implies that the results depend on the used similarity measure. We study the special case in which additional, stochastic information about a suitable similarity measure for the items $\boldsymbol{x}_k \in \mathbb{R}^n$ exists in the form of discrete auxiliary data c_k. Thus, the data consists of primary-auxiliary pairs $(\boldsymbol{x}_k, c_k)$. In the resulting clusters the data items $\boldsymbol{x}$ are similar by the associated conditional distributions $p(c|\boldsymbol{x})$. Still, because of their parameterization, the clusters are localized in the primary space in order to retain its (potentially useful) structure. The auxiliary information is only used to learn what distinctions are important in the primary data space.

We have earlier explicitly constructed an estimate $\hat{p}(c|\boldsymbol{x})$ of the conditional distributions, and a local Riemannian metric based on that estimate [5]. Metrics have additionally been derived from generative models that do not use auxiliary information [3, 4]. Both kinds of metrics could be used in standard clustering methods. In this paper we present a simpler method that directly minimizes the within-cluster dissimilarity, measured as distortion in the auxiliary space.

We additionally show that minimizing the within-cluster distortion maximizes the mutual information between the clusters and the auxiliary data. Maximization of mutual information has been used previously for constructing representations of the input data [1].

In another related work, the information bottleneck [7, 9], data is also clustered by maximizing mutual information with a relevance variable. Contrary to our work, the bottleneck treats discrete or prepartitioned data only, whereas we create the categories by optimizing a parametrized partitioning of a continuous input space.

2 The Clustering Method

We cluster samples $\boldsymbol{x} \in \mathbb{R}^n$ of a random variable X. The parameterization of the clusters keeps them local, and the similarity of the samples is measured as the similarity of the conditional distributions $p(c|\boldsymbol{x})$ of the random variable C.

Vector quantization (VQ) is one approach to categorization. In VQ the data space is divided into cells represented by prototypes or codebook vectors $\boldsymbol{m}_j$, and the average distortion between the data and the prototypes,

$$E = \sum_j \int y_j(\boldsymbol{x}) D(\boldsymbol{x}, \boldsymbol{m}_j)\, p(\boldsymbol{x})\, d\boldsymbol{x} \ , \tag{1}$$

is minimized. Here $D(\boldsymbol{x}, \boldsymbol{m}_j)$ denotes a dissimilarity between $\boldsymbol{x}$ and $\boldsymbol{m}_j$, and $y_j(\boldsymbol{x})$ is the cluster membership function for which $0 \le y_j(\boldsymbol{x}) \le 1$ and $\sum_j y_j(\boldsymbol{x}) = 1$. In the classic "hard" VQ the membership function is binary valued: $y_j(\boldsymbol{x}) = 1$ if $D(\boldsymbol{x}, \boldsymbol{m}_j) \le D(\boldsymbol{x}, \boldsymbol{m}_i)$, $\forall i$, and $y_j(\boldsymbol{x}) = 0$ otherwise. In the "soft" VQ, the $y_j(\boldsymbol{x})$ attain continuous values and they can be interpreted as conditional densities $p(v_j|\boldsymbol{x}) \equiv y_j(\boldsymbol{x})$ of a discrete random variable V that indicates the cluster identity. Given $\boldsymbol{x}$, C and V are conditionally independent: $p(c, v|\boldsymbol{x}) = p(c|\boldsymbol{x})p(v|\boldsymbol{x})$. It follows that $p(c, v) = \int p(c|\boldsymbol{x})p(v|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$.

Our measure of dissimilarity is the Kullback–Leibler divergence, defined for two multinomial distributions with event probabilities $\{p_i\}$ and $\{q_i\}$ as $D_{\mathrm{KL}}(p_i, q_i) \equiv \sum_i p_i \log(p_i/q_i)$. In our case, the first distribution corresponds to the data $\boldsymbol{x}$: $p_i \equiv p(c_i|\boldsymbol{x})$. The second distribution will be the prototype. It can be shown that the optimal prototype, given that the values of the $y_j(\boldsymbol{x})$ are fixed, is $q_j \equiv p(c_i|v_j) = p(c_i, v_j)/p(v_j)$. By plugging this prototype and the Kullback–Leibler distortion measure into the error function of VQ, equation (1), we get

$$E_{\mathrm{KL}} = \sum_j \int y_j(\boldsymbol{x}) D_{\mathrm{KL}}(p(c|\boldsymbol{x}), p(c|v_j)) p(\boldsymbol{x}) d\boldsymbol{x} \ . \tag{2}$$

Instead of computing the distortion between the vectorial samples and vectorial prototypes as in (1), we now have pointwise comparisons between the distributions $p(c|\boldsymbol{x})$ and the indirectly defined prototypes $p(c|v_j)$. The primary data space has been used to define the domain in the auxiliary space that is used for estimating each prototype.

If the membership functions are parametrized by $\boldsymbol{\theta}$ the average distortion becomes

$$E_{\mathrm{KL}} = -\sum_{i,j} \log p(c_i|v_j) \int y_j(\boldsymbol{x}; \boldsymbol{\theta}) p(c_i, \boldsymbol{x})\, d\boldsymbol{x} + \text{const.}, \tag{3}$$

where the constant is independent of the parameters. Note that minimizing the average distortion E_{KL} is equivalent to maximizing the mutual information between C and V, because $E_{\mathrm{KL}} = -I(C;V) + \mathrm{const.}$

The choice of parameterization of the membership functions depends on the data space. For Euclidean spaces Gaussians have desirable properties. When the data comes from an n-dimensional hypersphere, spherical analogs of Gaussians, the von Mises–Fisher (vMF) basis functions [6] are more approriate. Below we derive the algorithm for vMF's; the derivation for Gaussians is analogous.

Von Mises–Fisher Basis Functions. A normalized n-dimensional vMF basis function is defined for normalized data by

$$y_j(\boldsymbol{x}) = \frac{M(\boldsymbol{x};\boldsymbol{w}_j)}{\sum_k M(\boldsymbol{x};\boldsymbol{w}_k)} \ , \quad \text{where} \quad M(\boldsymbol{x};\boldsymbol{w}_j) = \frac{\kappa^{\frac{1}{2}n-1}}{(2\pi)^{\frac{1}{2}n} I_{\frac{1}{2}n-1}(\kappa)} \exp \kappa \frac{\boldsymbol{x}^T \boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} \ , \tag{4}$$

where $I_r(\kappa)$ denotes the modified Bessel function of the first kind and order r. The dispersion parameter κ is selected *a priori.* With the vMF basis functions the gradient of the average distortion (3) becomes

$$\nabla_{\boldsymbol{w}_j} E_{\mathrm{KL}} = \frac{1}{\sigma^2} \sum_i \sum_{l \neq j} \log \frac{p(c_i|v_j)}{p(c_i|v_l)} \int (\boldsymbol{x} - \boldsymbol{w}_j \boldsymbol{w}_j^T \boldsymbol{x}) y_j(\boldsymbol{x}) y_l(\boldsymbol{x}) p(c_i, \boldsymbol{x}) d\boldsymbol{x} \ , \tag{5}$$

where the $\boldsymbol{w}_j$ are assumed normalized (without loss of generality).

An on-line Algorithm can be derived using $y_j(\boldsymbol{x}) y_l(\boldsymbol{x}) p(c_i, \boldsymbol{x}) = p(v_j, v_l, c_i, \boldsymbol{x})$ as the sampling function for stochastic approximation. The following steps are repeated with $\alpha(t)$ gradually decreasing to zero:

1. At the step t of stochastic approximation, draw a data sample $(\boldsymbol{x}(t), c_i(t))$.
2. Draw independently two basis functions, j and l, according to the probabilities $\{y_k(\boldsymbol{x}(t))\}$.
3. Adapt the parameters $\boldsymbol{w}_j$ according to $\boldsymbol{w}_j(t+1) = \Delta \mathbf{w}_j / \|\Delta \boldsymbol{w}_j\|$, where

$$\Delta \boldsymbol{w}_j = \boldsymbol{w}_j(t) + \alpha(t) \log \frac{\hat{p}(c_i|v_j)}{\hat{p}(c_i|v_l)} \left(\boldsymbol{x}(t) - \boldsymbol{w}_j(t) \boldsymbol{w}_j(t)^T \boldsymbol{x}(t) \right) \ , \tag{6}$$

   and $\alpha(t)$ is the gradually decreasing step size. The $\hat{p}$ are estimates of the conditional probabilities. The parameters $\boldsymbol{w}_l$ can be adapted at the same step, by exchanging j and l in (6).
4. Adapt the estimates $\hat{p}(c_i|v_j)$ with stochastic approximation, using the expression

$$\hat{p}(c_i|v_j)(t+1) = (1-\lambda(t))\hat{p}(c_i|v_j)(t) + \lambda(t)$$
$$\hat{p}(c_k|v_j)(t+1) = (1-\lambda(t))\hat{p}(c_k|v_j)(t) \ , k \neq i$$

   where the rate of change $\lambda(t)$ should be larger than $\alpha(t)$. In practice, $2\alpha(t)$ seems to work.

3 Case Studies

We applied our model and two other models to two different data sets. The other models were the familiar mixture model $p(\boldsymbol{x}) = \sum_j p(\boldsymbol{x}|j)P(j)$, and the mixture discriminant model $p(c_i, \boldsymbol{x}) = \sum_j P(c_i|j)p(\boldsymbol{x}|j)P(j)$ (MDA2 [2]). The $P(j)$ are mixing parameters, and the $P(c_i|j)$ are additional parameters that model class distributions.

Clustering of text documents is useful as such, and the groupings can additionally be used to speed-up searches. We demonstrate that grouping based on textual content, with goodness measured by independent *topic* information, can be improved by our method utilizing (manually constructed) metadata (keywords). Thus, in this application our variable C corresponds to the keywords, and the variable X represents the textual content of the documents, encoded into a vector form.

Model performance was measured by the mutual information between the generated (soft) categories and nine *topic classes*, such as nuclear physics and optics, found independently by informaticians.

We carried out two sets of experiments with different preprocessing. The von Mises–Fisher kernels (4) were used both in our model and as the mixture components $p(\boldsymbol{x}|j) = M(\boldsymbol{x}; \boldsymbol{w}_j)$. To encode the textual content, the words in the abstracts and titles were used, converted to base form. The rarest words were discarded. Documents with less than 5 words remaining after the preprocessing were discarded, resulting in about 50,000 data vectors.

The first experiment utilized no prior relevance information of the words: we picked 500 random words and encoded the documents with the "vector space model" [8] with "TF" (term frequency) weighting. In the second experiment more prior information was utilized. Words belonging to a stop-list were removed, and the "TF-IDF" (term frequency times inverse document frequency) weighting was used. In the first experiment with 'random' feature selection, our method performed clearly better than the other models. With the improved feature extraction the margin reduced somewhat (Fig. 1).

Clustering enterprises by bankruptcy sensitivity. We clustered financial statements of small and medium-sized Finnish enterprises by bankruptcy sensitivity, a key issue affecting credit decisions. The data set consisted of 6195 financial statements of which 158 concerned companies later gone bankrupt. Multiple yearly statements from the same enterprise were treated as independent samples.

We compared the MDA2 with our model. The basis functions $M(\boldsymbol{x}; \boldsymbol{w}_j)$ of both models were Gaussians parametrized by their location, with the covariance matrices *a priori* set to $\sigma^2 \boldsymbol{I}$. Measured by the mutual information, our model clearly outperformed MDA2 (Fig. 2. Note that it is not feasible to estimate our model with the straightforward algorithm presented in this paper when σ is very small. The reason is that the gradient (5) becomes very small because of the products $y_j(\boldsymbol{x})y_l(\boldsymbol{x})$).

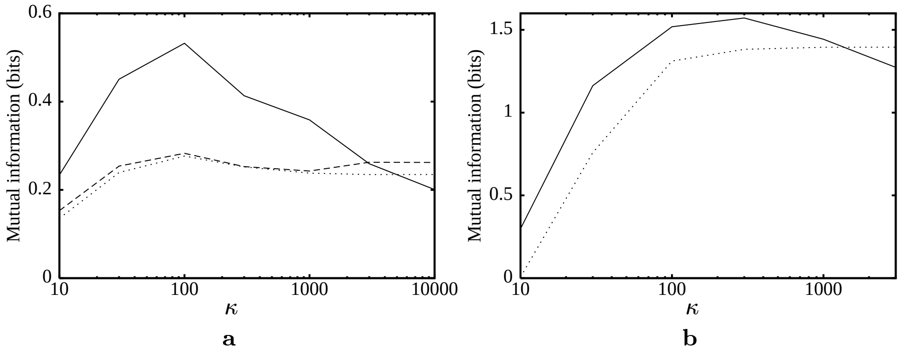

Fig. 1. Mutual information between the nine document clusters and the topics (*not* used in learning). **a** Random feature extraction, dimensionality: 500. **b** Informed feature extraction, dimensionality: 4748. Solid line: our model, dashed line: MDA2, dotted line: mixture model. (Due to slow convergence of MDA2, it was infeasible to compute it for part b. Another comparison with MDA2 is shown in Fig. 2)

4 Conclusions

We have demonstrated that clusters obtained by our method are more informative than clusters formed by a generative mixture model, MDA2 [2], for two kinds of data: textual documents and continuous-valued data derived from financial statements of enterprises. In (unpublished) tests for two additional data sets the results have been favorable to our model, although for one set the margin to MDA2 was narrow compared to the cases presented here.

For the first demonstration with textual documents, it would be interesting to compare the present method with the information bottleneck [7, 9] and metrics derived from generative models [3]. For the continuous data of the second experiment the bottleneck is not (directly) applicable. A generative model could be constructed, and we will compare our approach with such "unsupervised" generative models in subsequent papers.

When the feature extraction was improved using prior knowledge, the margin between our method and the "unsupervised" mixture model reduced. This suggests that our algorithm may be particularly useful when good feature extraction stages are not available but there exists auxiliary information that induces a suitable similarity measure.

Acknowledgments The authors would like to thank Finnvera Ltd., Pentti Bergius and Kimmo Kiviluoto for the bankruptcy data set, and Jaakko Peltonen for helping with the simulations. Academy of Finland provided financial support.

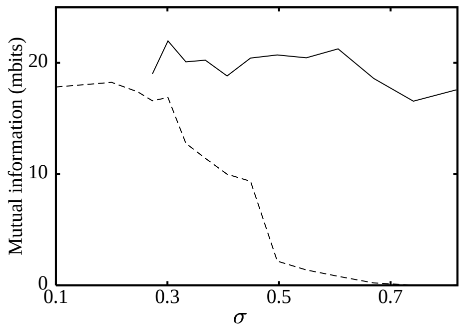

Fig. 2. Mutual information between the posterior probabilities of the ten enterprise clusters and the binary bankruptcy indicator. Solid line: our model, dashed line: MDA2. A set of 25 financial indicators was used as the primary data. The binary variable C indicated whether the statement was followed by a bankruptcy within 3 years

References

1. Becker, S.: Mutual information maximization: models of cortical self-organization. Network: Computation in Neural Systems **7** (1996) 7–31
2. Hastie, T., Tibshirani, R., Buja, A.: Flexible discriminant and mixture models. In: Kay, J., Titterington, D. (eds.): Neural Networks and Statistics. Oxford University Press (1995)
3. Hofmann, T.: Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In: Solla, S. A., Leen, T. K., Müller, K.-R. (eds.): Advances in Neural Information Processing Systems 12. MIT Press, Cambridge MA (2000) 914–920
4. Jaakkola, T. S., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Kearns, M. S., Solla, S. A., Cohn, D. A. (eds.): Advances in Neural Information Processing Systems 11. Morgan Kauffmann Publishers, San Mateo CA (1999) 487–493
5. Kaski, S., Sinkkonen, J.: Metrics that learn relevance. In: Proc. IJCNN-2000, International Joint Conference on Neural Networks. IEEE (2000) V:547–552
6. Mardia, K. V.: Statistics of directional data. Journal of the Royal Statistical Society, series B **37** (1975) 349–393
7. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of English words. In: Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (1983) 183–190.
8. Salton, G., McGill, M. J.: Introduction to modern information retrieval. McGraw-Hill, New York (1983)
9. Tishby, N., Pereira, F. C., Bialek, W.: The information bottleneck method. In: 37th Annual Allerton Conference on Communication, Control, and Computing. Illinois (1999)