Discriminative clustering in Fisher metrics

Jarkko Salojärvi, Samuel Kaski and Janne Sinkkonen
Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
{jarkko.salojarvi,samuel.kaski,janne.sinkkonen}@hut.fi

***Abstract*— Discriminative clustering (DC) finds a Voronoi partitioning of a primary data space that, while consisting of local partitions, simultaneously maximizes information about auxiliary data categories. DC is useful in exploration and in finding more coarse or refined versions of already existing categories. Theoretical results suggest that Voronoi partitions in the so-called Fisher metric would outperform partitions in the Euclidean metric. Here we use a local quadratic approximation of the Fisher metric, derived from a conditional density estimator, in defining the partitions and show that the resulting algorithms outperform the conventional variants.**

I. Introduction

Based on paired data $(\mathbf{x}, c)$ of values of random variables (X, C), discriminative clustering (DC; [1], [2]) partitions vectorial data $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$ into Voronoi regions V that are internally maximally homogeneous in terms of some categorical-valued auxiliary data c and, equivalently, mutually maximally different. The partitioning will retain information of c in the sense of asymptotically maximizing the mutual information between X and C [1]. DC is useful in explorative clustering, in predictive compression, and as a method for generating partitionings finer or coarser than the c-categories. DC has a generative interpretation as a maximizer of the likelihood of a simple piecewise constant model of the conditional distributions $p(c|\mathbf{x})$ [1], and a Bayesian version exists that properly takes into account the effects of sampling variation in finite data [2].

It has been suggested [3] that asymptotically, with a large data set and many local Voronoi regions, DC performs vector quantization using a distortion measured with a *learning metric* [3], [4], the analogue of *Fisher metric* or *information metric* [5] of the parameter spaces of generative models applied to the primary data space $\mathbb{X}$. The learning metric is defined for local neighbourhoods by

$$d_L(\mathbf{x}, \mathbf{x} + d\mathbf{x})^2 = d\mathbf{x}^T \mathbf{J}_F(\mathbf{x}) d\mathbf{x} \ , \tag{1}$$

where

$$\mathbf{J}_F(\mathbf{x}) = -E_{p(c|\mathbf{x})} \left[\frac{\partial^2}{\partial \mathbf{x}^2} \log p(c|\mathbf{x}) \right] \tag{2}$$

is a matrix analogous to the Fisher information matrix, familiar from the context of parameter estimation in generative models. Just like the Fisher metric is invariant to reparameterizations, the (local) metric (1) is invariant to smooth transformations of the data $\mathbf{x}$, and would therefore be preferable in data analysis as it would be relatively insensitive to feature extraction.

This asymptotic property of DC, however, comes with an additional constraint: although the distortion within the clusters is measured in the Fisher metric, the Voronoi regions are defined in the original Euclidean metric, making the partitioning less immune to feature transformations than would seem theoretically possible (Sect. II-B).

In this paper we attempt to construct Voronoi regions in approximated Fisher metrics, by plugging a local quadratic approximation of the Fisher metric, derived from a conditional Parzen density estimator, into the ordinary k-means and into a discriminative clustering algorithm. This modification of DC, motivated by the asymptotical properties of the method, improves practical clustering results as well.

Local metric modifications similar in spirit have earlier been applied to k-NN classifiers [6], [7], and isotropic magnification of the space near the class boundary has been tried with SVMs [8].

II. Discriminative Clustering

A. The Principle

We will begin by reviewing the original discriminative clustering model [1], [2]. Given pairs $(\mathbf{x}, c)$ of primary data $\mathbf{x}$ and auxiliary data c, the goal is to partition the primary data space into clusters that are (i) local with respect to $\mathbf{x}$ and (ii) homogeneous and predictive in terms of auxiliary data c. (The connection between homogeneity and predictivity of the clusters is detailed below.) Locality is enforced by defining the clusters as Voronoi regions: $\mathbf{x}$ belongs to cluster j, $\mathbf{x} \in V_j$, if $\|\mathbf{x} - \mathbf{m}_j\| \leq \|\mathbf{x} - \mathbf{m}_k\|$ for all k. The Voronoi regions are uniquely determined by the parameters $\{\mathbf{m}_j\}$.

Homogeneity is enforced by assigning a *distributional prototype* $\boldsymbol{\psi}_j \equiv \{\psi_{jc}\}_c$ (with $\sum_c \psi_{jc} = 1$) to each Voronoi region j, and searching for partitionings capable of predicting auxiliary data with the prototypes. The resulting model for $p(c|\mathbf{x})$, with the log likelihood

$$L(\{\mathbf{m}_j\}, \{\psi_{jc}\}) = \sum_j \sum_{\mathbf{x} \in V_j} \log \psi_{j,c(\mathbf{x})} \ , \tag{3}$$

predicts the probability of class c within Voronoi region V_j to be equal to ψ_{jc}. Here $c(\mathbf{x})$ is the auxiliary data value paired to $\mathbf{x}$. Asymptotically for large data sets the log-likelihood approaches

$$L_\infty = -\sum_j \int_{V_j} D_{KL}(p(c|\mathbf{x}), \boldsymbol{\psi}_j) p(\mathbf{x}) d\mathbf{x} + \text{const.} \ , \tag{4}$$

where D_{KL} is the Kullback-Leibler divergence between the prototype distribution $\{\psi_{jc}\}_c$ and the pointwise conditional distribution of auxiliary data $p(c|\mathbf{x})$. In this sense, maximizing the likelihood of the model also makes the distributional clusters internally homogeneous. Notice that this resembles the standard k-means clustering; only difference is that the distortion within the clusters is measured by D_{KL} between distributions, instead of by the Euclidean distance.

The asymptotic likelihood L_∞ has another interpretation besides the homogeneity connection (4). It is also equal, up to a constant, to the mutual information between the auxiliary data and the Voronoi partitioning [1]. Discriminative clustering therefore has an interpretation as finding a partitioning $\{V_j\}_j$ for the primary data space $\mathbb{X}$ such that the statistical dependency between the discrete Voronoi representation and the auxiliary data is maximized.

In practical applications where the amount of data is finite, the likelihood (3) consequently corresponds to empirical mutual information which, due to random fluctuations in the observed joint frequencies of (C, V), is known to be a bad measure of statistical dependency for small data sets. (In other words, the maximum likelihood estimates for $\{\psi_{jc}\}$ are simply normalized class-wise frequencies of data within Voronoi partitions, and clearly for small data sets the frequencies are poor estimates of actual proportions.) Maximizing the likelihood (3) or, equivalently, the empirical mutual information, would then probably favor "fake" statistical dependency due to random fluctuations of the observed joint frequencies of (C, V).

This viewpoint to the likelihood (3), together with the fact that we are actually interested in the parameters $\{\mathbf{m}_j\}$ that define the clusters and less interested in the predictions $\psi_{j,c(\mathbf{x})}$ of $p(c|\mathbf{x})$, motivates a Bayesian approach called MAP DC [2]. It takes the sampling uncertainty better into account by integrating over the parameters $\{\boldsymbol{\psi}_j\}$ in which the sampling variation is primarily manifested.

The derivation of MAP DC proceeds by first extending the likelihood (3) into a posterior probability of the model; a uniform prior for $\{\mathbf{m}_j\}$ is assumed, and a Dirichlet prior for $\{\boldsymbol{\psi}_j\}$. The latter parameters are then marginalized out, leaving the cost function

$$p(\{\mathbf{m}\}|D) \propto \prod_j \frac{\prod_i \Gamma(n_i^0 + n_{ji})}{\Gamma(N^0 + N_j)} \tag{5}$$

to be maximized only with respect to the remaining parameters $\mathbf{m}_j$. Above, data is denoted by D, n_{ji} is the number of data samples of class $c = i$ in cluster j, and $N_j = \sum_i n_{ji}$. The $n_i^0 = 1$ are Dirichlet prior parameters common to all j, and $N^0 = \sum_i n_i^0$.

The objective function (5) can be maximized with gradient methods by introducing "soft counts" $y_j(\mathbf{x})$, with $\sum y_j(\mathbf{x}) = 1$ of the data point $\mathbf{x}$ in the cluster j. Then $n_{ji} = \sum_{c(\mathbf{x})=i} y_j(\mathbf{x})$. In [2]

$$y_j(\mathbf{x}) = \frac{\exp\{-d(\mathbf{x}, \mathbf{m}_j)^2/2\sigma^2\}}{\sum_k \exp\{-d(\mathbf{x}, \mathbf{m}_k)^2/2\sigma^2\}} \tag{6}$$

has been used, where σ is a softness parameter and $d(\mathbf{x}, \mathbf{m}_j)$ are Euclidean distances. This is the currently preferred version of DC for practical applications.

B. Asymptotic Connection of DC to K-means in Fisher Metrics

In [3] it was suggested that asymptotically, under suitable regularity conditions including continuity of $p(c|\mathbf{x})$ with respect to $\mathbf{x}$, and with many local Voronoi regions, the cost function (4) becomes

$$E_{KL} \approx \sum_j \int_{V_j} (\mathbf{x} - \tilde{\mathbf{m}}_j)^T \mathbf{J}_F(\tilde{\mathbf{m}}_j)(\mathbf{x} - \tilde{\mathbf{m}}_j) p(\mathbf{x}) d\mathbf{x}, \tag{7}$$

with $\mathbf{J}_F(\tilde{\mathbf{m}}_j)$ from (2), and $\tilde{\mathbf{m}}_j$ being a point of the Voronoi region j where $\psi_{j,c} = p(c|\tilde{\mathbf{m}}_j)$. In other words, asymptotically DC performs k-means with a Fisherian distortion but with Euclidean Voronoi regions. In the original DC the Voronoi regions are defined in the Euclidean metric, whereas the asymptotic result suggests that the distortion would better be minimized with Fisherian Voronoi regions [3]. It is therefore tempting to try to find approximate Fisherian Voronoi regions even with finite data and a smaller number of clusters.

III. Algorithms

In this section we present two new algorithms utilizing an approximate Fisher metric of the data space: k-means in Fisher metric and MAP DC in Fisher metric.

A. K-means in Approximated Fisher Metrics

The first approach is motivated by the asymptotic connection (7). It differs from DC algorithms by performing explicit, simple k-means in a modified metric instead of explicitly optimizing the distributional homogeneity of the clusters. Homogeneity will appear because distances along directions of fast change of $p(c|\mathbf{x})$ in the $\mathbb{X}$-space are long in the new metric.

Most straightforwardly, we would estimate $p(c|\mathbf{x})$ and then use its gradients with respect to $\mathbf{x}$ to construct $\mathbf{J}_F(\mathbf{x})$, in the same way as in [4]. Here we introduce an improved, although computationally more intensive, approach where the kernel shapes of a Parzen estimator are iteratively updated. An initial estimator with kernels spherical in the Euclidean metric is used to find an approximation of the Fisher metric. It is plugged in to the Parzen estimator to get kernels spherical in the approximated Fisher metric, etc. In our implementation, Gaussian kernels are used for their analytical simplicity.

For computational reasons, local approximations of the Fisher metric are evaluated at each data point instead of at the model vectors. To balance inaccuracies in the local approximation and to regularize possible singularities in the metrics, we use

$$\mathbf{J}_R(\mathbf{x}) = (1 - \lambda)\mathbf{J}_F(\mathbf{x}) + \lambda \mathbf{I} \tag{8}$$

instead of $\mathbf{J}_F(\mathbf{x})$, with $\lambda \in [0, 1]$ chosen individually for each data point by

$$\lambda(\mathbf{x}) = \arg\max_\lambda \log \psi_{j,c(\mathbf{x})}, \tag{9}$$

which maximizes the contribution of an individual sample to the overall cost function (3).

When k-means is computed in the approximated Fisher metric, each data point will have its own metric matrix $\mathbf{J}_R(\mathbf{x})$. The matrix defines distances in the neighbourhood of the point, included in the k-means distortion (not shown for brevity). Assuming the metric changes slowly in the data space, it turns out that the update is identical to ordinary k-means (see [4] for a rigorous derivation in an analogous case, the Self-Organizing Map). In other words, the change of the metric makes a difference only in choosing the Voronoi regions for the data points, not in the update itself.

Note that the optimization of the λ's, the sample-wise regularization coefficients in (9), is carried out with fixed cluster centers and auxiliary data distributions $\psi_{j,c(\mathbf{x})}$. Therefore we need to iterate the locations of the cluster centers and the amount of regularization, resulting in the following algorithm:

1) Estimate $\mathbf{J}_F(\mathbf{x})$ with Parzen kernels using either Euclidean metric (in the first iteration) or $\mathbf{J}_R(\mathbf{x})$ (from the previous iteration).
2) Iterate until convergence of the likelihood (3):
   - For each sample, optimize λ by (9).
   - Compute distances in the regularized metric $\mathbf{J}_R(\mathbf{x})$. Find closest cluster centers, update centroids by the k-means update rule.
3) Go back to 1).

In order to find the cluster centroid $\mathbf{m}_j$ closest to new (test) data points, a local metric estimate ($\mathbf{J}_F(\mathbf{x})$) is computed separately at each new data point with the current Parzen kernels. In our implementation the amount of regularization $\lambda(\mathbf{x})$ was set to the mean of the λ's for the training data, that is, the same for all new points. (A more sophisticated method might produce better results.)

B. Marginalized DC in Approximated Fisher Metrics

The second, alternative new algorithm of this paper, explicitly optimizes the marginalized discriminative clustering criterion (5) but for (softened) Voronoi regions of the (approximated) Fisher metric.

Instead of Euclidean distance, we now compute distance $d(\mathbf{x}, \mathbf{m}_j)$ in (6) by the learning metric (1). In practice, for the same reasons as in Sect. III-A, we use the regularized version $\mathbf{J}_R(\mathbf{x})$ defined in (8).

The conditional densities behind the metric, needed in (2), are estimated using Parzen kernels; again updated in an iterative manner. In this implementation, however, the construction of the metric estimate $\mathbf{J}_R(\mathbf{x})$ differs from the one in Sect. III-A, since now it is not practical to select the amount of regularization for each data point individually. With the conjugate gradient algorithm used to optimize (5), it would be computationally extremely demanding to adjust the regularization coefficients λ of the metric separately for each sample and for each cluster center position tried out during line searches. Hence we resort here to a much simpler method and just select a common, overall amount of regularization for all data points by running the algorithm with different candidate values. Summarized, the algorithm iterates the following two steps:

1) Compute $\mathbf{J}_F(\mathbf{x})$ with Parzen kernels of the Euclidean metric (first iteration) or of Fisher metric ($\mathbf{J}_R(\mathbf{x})$ from previous iteration).
2) Find the amount of regularization. For different values of $\lambda \in [0..1]$, optimize the cluster centers in the regularized metric with the MAP DC algorithm using the conjugate gradient method. Select the metric $\mathbf{J}_R(\mathbf{x})$ giving the largest value of the objective function.

The cluster assignments for new data were found by first forming $\mathbf{J}_F(\mathbf{x})$ at the test data points with the current Parzen kernels. Then the optimal amount of regularization was added (found at step 2 above) to get the regularized metric, in which the value of the objective function (3) was evaluated with hard cluster assignments.

C. Difference Between the Methods

The main difference between the two introduced algorithms is in how $p(c|\mathbf{x})$ influences the result. In the algorithm introduced in Sect. III-A it only affects the (explicit) estimate of the metric, whereas in the algorithm in Sect. III-B the homogeneity of the clusters in terms of $p(c|\mathbf{x})$ is explicitly optimized, without absolutely relying on the goodness of a local estimate of the metric as in the previous approach. In our modified version, the data affects the partitioning in two ways, primarily through the explicitly optimized DC cost function, and secondly through the density function estimate that generates the metric used to define the Fisherian Voronoi regions.

IV. Experiments

Experiments were run with the Wisconsin breast cancer data set from the UCI machine learning repository [9]. The 569 samples consisted of 30 attributes, measured from malignant and benign tumors.

We chose the ordinary k-means as the baseline reference method. Other methods in the comparison were an implementation of non-iterating Fisher k-means according to [4], and the original MAP DC [2] (in Table IV denoted as "orig. Fisher k-means" and "MAP DC", respectively). The optimal parameter values (Parzen kernel width, softening parameter of MAP DC) were chosen in preliminary tests, and the cluster centroids were initialized to randomly sampled data points. The number of prototypes, k, was 5 and the number of Parzen kernel iterations was 8 for iterative Fisher k-means and 3 for Fisher MAP DC.

For iterative Fisher k-means, the optimization of the λ values in the criterion (9) was carried out with the Fibonacci search. If the Fisher metric did not improve the result, ordinary Euclidean metric was preferred.

Since the intrinsic dimensionality of the approximative Fisher metric was smaller than original data dimensionality (the rank of the Fisher metric is essentially the number of

TABLE I

EXPERIMENTAL RESULTS WITH 20-FOLD CROSS-VALIDATION. MEAN OF (3) FOR TEST DATA AND 95% CONFIDENCE INTERVALS (**: SIGNIFICANT DIFFERENCE AGAINST THE THREE OTHER METHODS).

Algorithm	log-likelihood	
k-means	-7.93 $\pm$ 1.23	
orig. Fisher k-means	-7.89 $\pm$ 1.25	
orig. MAP DC	-7.49 $\pm$ 2.27	
Fisher MAP DC	-6.54 $\pm$ 2.29	**
Fisher k-means	-6.28 $\pm$ 0.77	**

c-categories), we added some Euclidean metric to the approximations of Fisher metric in each data point and normalized the matrices to have unit determinant before plugging them to the Parzen estimator.

Results were measured in 20-fold cross-validation. Since k-means is likely to get trapped in local minima (and even more so in our experiments where, for computational reasons, only one randomly chosen starting point for the prototype vectors was used in each cross-validation run), the results cannot be expected to follow a Gaussian distribution. Hence the statistical significance of the differences between the attained values of the objective function (3) for different methods was measured with the Wilcoxon signed-rank test for paired data.

In these experiments, neither the original Fisher k-means (p=0.5) nor the original MAP DC (p=0.057) turned out to be significantly better than ordinary k-means. However, the versions that iteratively updated the approximations of the Fisher Voronoi regions did perform significantly better than k-means (p-values for Fisher k-means and Fisher MAP DC were 0.003 and 0.008, respectively) and their original non-iterative implementations (p=0.005 and p=0.006; respectively). The difference between the new Fisher k-means and Fisher MAP DC was not significant (p=0.28) nor the difference between the old versions of the algorithms (p=0.062). The means of the objective function (3) values for the test set, and their 95% confidence intervals (assuming a Gaussian distribution of the results) are shown in Table IV.

V. DISCUSSION

We have examined different ways to compute clusters in approximated Fisher metrics. A method for shaping a Parzen density estimator towards a theoretically optimal form was introduced, and it proved successful in the empirical test.

Actually, with a better initialization of the cluster centroids directly from the conventional MAP DC, the iterative methods introduced here can only improve performance of discriminative clustering, since Euclidean metric is preferred if Fisher metric does not increase the objective function value (Naturally one must remember to be careful not to overfit the data.)

Since we are doing k-means (albeit in a different metric), one might expect that the usual pitfalls of k-means hold also here, that is, the algorithms may be susceptible to get trapped in local minima. The effect of this problem on performance needs to be measured empirically with a larger number of data sets.

The methods using approximative Fisher metric were computationally extremely heavy. For larger data sets, optimizations and approximations are needed.

ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland, grants 52123 and 50061.

REFERENCES

[1] J. Sinkkonen and S. Kaski, "Clustering based on conditional distributions in an auxiliary space," *Neural Computation*, vol. 14, pp. 217–239, 2002.

[2] J. Sinkkonen, S. Kaski, and J. Nikkilä, "Discriminative clustering: Optimal contingency tables by learning metrics," in *Proceedings of the ECML'02, 13th European Conference on Machine Learning*, T. Elomaa, H. Mannila, and H. Toivonen, Eds. Berlin: Springer, 2002, pp. 418–430.

[3] S. Kaski and J. Sinkkonen, "Principle of learning metrics for data analysis," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, Special Issue on Data Mining and Biomedical Applications of Neural Networks*, accepted for publication.

[4] S. Kaski, J. Sinkkonen, and J. Peltonen, "Bankruptcy analysis with self-organizing maps in learning metrics," *IEEE Transactions on Neural Networks*, vol. 12, pp. 936–947, 2001.

[5] S. Amari and H. Nagaoka, *Methods of Information Geometry*, ser. Translations of Mathematical Monographs. American Mathematical Society and Oxford University Press, 2000.

[6] C. Domeniconi, J. Peng, and D. Gunopulos, "Locally adaptive metric nearest-neighbor classificaton," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1281–1285, 2002.

[7] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neigbor classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 607–616, 1996.

[8] S. Amari and S. Wu, "An information-geometrical method for improving the performance of support vector machines," in *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*. London: IEE, 1999, pp. 85–90.

[9] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 1998.