On Information-Maximization Clustering: Tuning Parameter Selection and Analytic Solution

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Makoto Yamada YAMADA@SG.CS.TITECH.AC.JP
Manabu Kimura KIMURA@SG.CS.TITECH.AC.JP
Hirotaka Hachiya HACHIYA@SG.CS.TITECH.AC.JP
Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan.

Abstract

Information-maximization clustering learns a probabilistic classifier in an unsupervised manner so that mutual information between feature vectors and cluster assignments is maximized. A notable advantage of this approach is that it only involves continuous optimization of model parameters, which is substantially easier to solve than discrete optimization of cluster assignments. However, existing methods still involve non-convex optimization problems, and therefore finding a good local optimal solution is not straightforward in practice. In this paper, we propose an alternative information-maximization clustering method based on a *squared-loss* variant of mutual information. This novel approach gives a clustering solution *analytically* in a computationally efficient way via kernel eigenvalue decomposition. Furthermore, we provide a practical model selection procedure that allows us to objectively optimize tuning parameters included in the kernel function. Through experiments, we demonstrate the usefulness of the proposed approach.

1. Introduction

The goal of *clustering* is to classify data samples into disjoint groups in an unsupervised manner. *K-means* is a classic but still popular clustering algorithm. However, since k-means only produces linearly separated clusters, its usefulness is rather limited in practice.

To cope with this problem, various non-linear clustering methods have been developed. *Kernel k-means* (Girolami, 2002) performs k-means in a feature space induced by a reproducing kernel function. *Spectral clustering* (Shi & Malik, 2000) first unfolds non-linear data manifolds by a spectral embedding method, and then performs k-means in the embedded space. *Blurring mean-shift* (Fukunaga & Hostetler, 1975) uses a non-parametric kernel density estimator for modeling the data-generating probability density and finds clusters based on the modes of the estimated density. *Discriminative clustering* (Xu et al., 2005; Bach & Harchaoui, 2008) learns a discriminative classifier for separating clusters, where class labels are also treated as parameters to be optimized. *Dependence-maximization clustering* (Song et al., 2007; Faivishevsky & Goldberger, 2010) determines cluster assignments so that their dependence on input data is maximized.

These non-linear clustering techniques would be capable of handling highly complex real-world data. However, they suffer from lack of objective model selection strategies[1]. More specifically, the above non-linear clustering methods contain tuning parameters such as the width of Gaussian functions and the number of nearest neighbors in kernel functions or similarity measures, and these tuning parameter values need to be heuristically determined in an unsupervised manner. The problem of learning similarities/kernels was addressed in earlier works, but they considered supervised setups, i.e., labeled samples are assumed to be given. Zelnik-Manor & Perona (2005) provided a useful unsupervised heuristic to determine the similarity in a data-dependent way. However, it still requires the number of nearest neighbors to be determined man-

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

[1] 'Model selection' in this paper refers to the choice of tuning parameters in kernel functions or similarity measures, not the choice of the number of clusters.

ually (although the magic number '7' was shown to work well in their experiments).

Another line of clustering framework called *information-maximization clustering* (Agakov & Barber, 2006; Gomes et al., 2010) exhibited the state-of-the-art performance. In this information-maximization approach, probabilistic classifiers such as a kernelized Gaussian classifier (Agakov & Barber, 2006) and a kernel logistic regression classifier (Gomes et al., 2010) are learned so that *mutual information* (MI) between feature vectors and cluster assignments is maximized in an unsupervised manner. A notable advantage of this approach is that classifier training is formulated as continuous optimization problems, which are substantially simpler than discrete optimization of cluster assignments. Indeed, classifier training can be carried out in computationally efficient manners by a gradient method (Agakov & Barber, 2006) or a quasi-Newton method (Gomes et al., 2010). Furthermore, Agakov & Barber (2006) provided a model selection strategy based on the common information-maximization principle. Thus, kernel parameters can be systematically optimized in an unsupervised way.

However, in the above MI-based clustering approach, the optimization problems are non-convex, and finding a good local optimal solution is not straightforward in practice. The goal of this paper is to overcome this problem by providing a novel information-maximization clustering method. More specifically, we propose to employ a variant of MI called *squared-loss MI* (SMI), and develop a new clustering algorithm whose solution can be computed analytically in a computationally efficient way via eigenvalue decomposition. Furthermore, for kernel parameter optimization, we propose to use a non-parametric SMI estimator called *least-squares MI* (LSMI) (Suzuki et al., 2009), which was proved to achieve the optimal convergence rate with analytic-form solutions. Through experiments on various real-world datasets such as images, natural languages, accelerometric sensors, and speech, we demonstrate the usefulness of the proposed clustering method.

2. Information-Maximization Clustering with Squared-Loss Mutual Information

In this section, we describe our novel clustering algorithm.

2.1. Formulation of Information-Maximization Clustering

Suppose we are given d-dimensional i.i.d. feature vectors of size n,

$$\{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n,$$

which are assumed to be drawn independently from a distribution with density $p^*(\boldsymbol{x})$. The goal of clustering is to give cluster assignments,

$$\{y_i \mid y_i \in \{1, \ldots, c\}\}_{i=1}^n,$$

to the feature vectors $\{\boldsymbol{x}_i\}_{i=1}^n$, where c denotes the number of classes. Throughout this paper, we assume that c is known.

In order to solve the clustering problem, we take the *information-maximization* approach (Agakov & Barber, 2006; Gomes et al., 2010). That is, we regard clustering as an unsupervised classification problem, and learn the class-posterior probability $p^*(y|\boldsymbol{x})$ so that 'information' between feature vector $\boldsymbol{x}$ and class label y is maximized.

The *dependence-maximization* approach (Song et al., 2007; Faivishevsky & Goldberger, 2010) is related to, but substantially different from the above information-maximization approach. In the dependence-maximization approach, cluster assignments $\{y_i\}_{i=1}^n$ are directly determined so that their dependence on feature vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ is maximized. Thus, the dependence-maximization approach intrinsically involves combinatorial optimization with respect to $\{y_i\}_{i=1}^n$. On the other hand, the information-maximization approach involves continuous optimization with respect to the parameter $\boldsymbol{\alpha}$ included in a class-posterior model $p(y|\boldsymbol{x};\boldsymbol{\alpha})$. This continuous optimization of $\boldsymbol{\alpha}$ is substantially easier to solve than discrete optimization of $\{y_i\}_{i=1}^n$.

Another advantage of the information-maximization approach is that it naturally allows out-of-sample clustering based on the discriminative model $p(y|\boldsymbol{x};\boldsymbol{\alpha})$, i.e., a cluster assignment for a new feature vector can be obtained based on the learned discriminative model.

2.2. Squared-Loss Mutual Information

As an information measure, we adopt *squared-loss mutual information* (SMI). SMI between feature vector $\boldsymbol{x}$ and class label y is defined by

$$\mathrm{SMI} := \frac{1}{2} \int \sum_{y=1}^{c} p^*(\boldsymbol{x}) p^*(y) \left(\frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x}) p^*(y)} - 1 \right)^2 \mathrm{d}\boldsymbol{x}, \tag{1}$$

where $p^*(\boldsymbol{x}, y)$ denotes the joint density of $\boldsymbol{x}$ and y, and $p^*(y)$ is the marginal probability of y. SMI is the *Pearson divergence* (Pearson, 1900) from $p^*(\boldsymbol{x}, y)$ to $p^*(\boldsymbol{x})p^*(y)$, while the ordinary MI (Cover & Thomas, 2006) is the *Kullback-Leibler divergence* (Kullback & Leibler, 1951) from $p^*(\boldsymbol{x}, y)$ to $p^*(\boldsymbol{x})p^*(y)$:

$$\mathrm{MI} := \int \sum_{y=1}^{c} p^*(\boldsymbol{x}, y) \log \frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x})p^*(y)} \mathrm{d}\boldsymbol{x}. \quad (2)$$

The Pearson divergence and the Kullback-Leibler divergence both belong to the class of *Ali-Silvey-Csiszár divergences* (which is also known as f-divergences, see (Ali & Silvey, 1966; Csiszár, 1967)), and thus they share similar properties. For example, SMI is non-negative and takes zero if and only if $\boldsymbol{x}$ and y are statistically independent, as the ordinary MI.

In the existing information-maximization clustering methods (Agakov & Barber, 2006; Gomes et al., 2010), MI is used as the information measure. On the other hand, in this paper, we adopt SMI because it allows us to develop a clustering algorithm whose solution can be computed analytically in a computationally efficient way via eigenvalue decomposition, as described below.

2.3. Clustering by SMI Maximization

Here, we give a computationally-efficient clustering algorithm based on SMI (1).

We can express SMI as

$$\mathrm{SMI} = \frac{1}{2} \int \sum_{y=1}^{c} p^*(\boldsymbol{x}, y) \frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x})p^*(y)} \mathrm{d}\boldsymbol{x} - \frac{1}{2} \quad (3)$$

$$= \frac{1}{2} \int \sum_{y=1}^{c} p^*(y|\boldsymbol{x}) p^*(\boldsymbol{x}) \frac{p^*(y|\boldsymbol{x})}{p^*(y)} \mathrm{d}\boldsymbol{x} - \frac{1}{2}. \quad (4)$$

Suppose that the class-prior probability $p^*(y)$ is set to be uniform: $p^*(y) = 1/c$. Then Eq.(4) is expressed as

$$\frac{c}{2} \int \sum_{y=1}^{c} p^*(y|\boldsymbol{x}) p^*(\boldsymbol{x}) p^*(y|\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \frac{1}{2}. \quad (5)$$

Let us approximate the class-posterior probability $p^*(y|\boldsymbol{x})$ by the following kernel model:

$$p(y|\boldsymbol{x}; \boldsymbol{\alpha}) := \sum_{i=1}^{n} \alpha_{y,i} K(\boldsymbol{x}, \boldsymbol{x}_i), \quad (6)$$

where $K(\boldsymbol{x}, \boldsymbol{x}')$ denotes a kernel function with a kernel parameter t. In the experiments, we will use a sparse variant of the *local-scaling kernel* (Zelnik-Manor & Perona, 2005):

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} \exp\left(-\dfrac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma_i \sigma_j}\right) \\ \quad \text{if } \boldsymbol{x}_i \in \mathcal{N}_t(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in \mathcal{N}_t(\boldsymbol{x}_i), \\ 0 \qquad\qquad\qquad\qquad \text{otherwise,} \end{cases} \quad (7)$$

where $\mathcal{N}_t(\boldsymbol{x})$ denotes the set of t nearest neighbors for $\boldsymbol{x}$ (t is the kernel parameter), σ_i is a local scaling factor defined as $\sigma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i^{(t)}\|$, and $\boldsymbol{x}_i^{(t)}$ is the t-th nearest neighbor of $\boldsymbol{x}_i$.

Further approximating the expectation with respect to $p^*(\boldsymbol{x})$ included in Eq.(5) by the empirical average of samples $\{\boldsymbol{x}_i\}_{i=1}^n$, we arrive at the following SMI approximator:

$$\widehat{\mathrm{SMI}} := \frac{c}{2n} \sum_{y=1}^{c} \boldsymbol{\alpha}_y^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_y - \frac{1}{2}, \quad (8)$$

where $^\top$ denotes the transpose, $\boldsymbol{\alpha}_y := (\alpha_{y,1}, \ldots, \alpha_{y,n})^\top$, and $K_{i,j} := K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

For each cluster y, we maximize $\boldsymbol{\alpha}_y^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_y$ under[2] $\|\boldsymbol{\alpha}_y\| = 1$. Since this is the *Rayleigh quotient*, the maximizer is given by the normalized principal eigenvector of $\boldsymbol{K}$ (Horn & Johnson, 1985). To avoid all the solutions $\{\boldsymbol{\alpha}_y\}_{y=1}^c$ to be reduced to the same principal eigenvector, we impose their mutual orthogonality: $\boldsymbol{\alpha}_y^\top \boldsymbol{\alpha}_{y'} = 0$ for $y \neq y'$. Then the solutions are given by the normalized eigenvectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_c$ associated with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ of $\boldsymbol{K}$. Since the sign of $\boldsymbol{\phi}_y$ is arbitrary, we set the sign as

$$\widetilde{\boldsymbol{\phi}}_y = \boldsymbol{\phi}_y \times \mathrm{sign}(\boldsymbol{\phi}_y^\top \mathbf{1}_n),$$

where $\mathrm{sign}(\cdot)$ denotes the sign of a scalar and $\mathbf{1}_n$ denotes the n-dimensional vector with all ones.

On the other hand, since

$$p^*(y) = \int p^*(y|\boldsymbol{x}) p^*(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{i=1}^{n} p(y|\boldsymbol{x}_i; \boldsymbol{\alpha}) = \boldsymbol{\alpha}_y^\top \boldsymbol{K} \mathbf{1}_n,$$

and the class-prior probability $p^*(y)$ was set to be uniform, we have the following normalization condition:

$$\boldsymbol{\alpha}_y^\top \boldsymbol{K} \mathbf{1}_n = 1/c.$$

Furthermore, probability estimates should be non-negative, which can be achieved by rounding up negative outputs to zero. Taking these issues into account,

[2] Note that this unit-norm constraint is not essential since the obtained solution is renormalized later.

cluster assignments $\{y_i\}_{i=1}^n$ for $\{\boldsymbol{x}_i\}_{i=1}^n$ are determined as

$$y_i = \operatorname*{argmax}_y \frac{[\max(\mathbf{0}_n, \widetilde{\boldsymbol{\phi}}_y)]_i}{\max(\mathbf{0}_n, \widetilde{\boldsymbol{\phi}}_y)^\top \mathbf{1}_n},$$

where the max operation for vectors is applied in the element-wise manner and $[\cdot]_i$ denotes the i-th element of a vector. Note that we used $\boldsymbol{K}\widetilde{\boldsymbol{\phi}}_y = \lambda_y \widetilde{\boldsymbol{\phi}}_y$ in the above derivation.

We call the above method *SMI-based clustering* (SMIC).

2.4. Kernel Parameter Choice by SMI Maximization

Since the above clustering approach was developed in the framework of SMI maximization, it would be natural to determine the kernel parameters so that SMI is maximized. A direct approach is to use the above SMI estimator $\widehat{\mathrm{SMI}}$ also for kernel parameter choice. However, this direct approach is not favorable because $\widehat{\mathrm{SMI}}$ is an unsupervised SMI estimator (i.e., SMI is estimated only from unlabeled samples $\{\boldsymbol{x}_i\}_{i=1}^n$). In the model selection stage, however, we have already obtained labeled samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and thus supervised estimation of SMI is possible. For supervised SMI estimation, a non-parametric SMI estimator called *least-squares mutual information* (LSMI) (Suzuki et al., 2009) was shown to achieve the optimal convergence rate. For this reason, we propose to use LSMI for model selection, instead of $\widehat{\mathrm{SMI}}$ (8).

LSMI is an estimator of SMI based on paired samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. The key idea of LSMI is to learn the following *density-ratio function*,

$$r^*(\boldsymbol{x}, y) := \frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x})p^*(y)}, \tag{9}$$

without going through density estimation of $p^*(\boldsymbol{x}, y)$, $p^*(\boldsymbol{x})$, and $p^*(y)$. More specifically, let us employ the following density-ratio model:

$$r(\boldsymbol{x}, y; \boldsymbol{\theta}) := \sum_{\ell: y_\ell = y} \theta_\ell L(\boldsymbol{x}, \boldsymbol{x}_\ell), \tag{10}$$

where $L(\boldsymbol{x}, \boldsymbol{x}')$ is a kernel function with kernel parameter γ. In the experiments, we will use the Gaussian kernel:

$$L(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\gamma^2}\right). \tag{11}$$

The parameter $\boldsymbol{\theta}$ in the above density-ratio model is learned so that the following squared error is minimized:

$$\frac{1}{2}\int \sum_{y=1}^{c} \Big(r(\boldsymbol{x}, y; \boldsymbol{\theta}) - r^*(\boldsymbol{x}, y)\Big)^2 p^*(\boldsymbol{x}) p^*(y) \mathrm{d}\boldsymbol{x}. \tag{12}$$

Among n cluster assignments $\{y_i\}_{i=1}^n$, let n_y be the number of samples in cluster y. Let $\boldsymbol{\theta}_y$ be the parameter vector corresponding to the kernel bases $\{L(\boldsymbol{x}, \boldsymbol{x}_\ell)\}_{\ell: y_\ell = y}$, i.e., $\boldsymbol{\theta}_y$ is the n_y-dimensional subvector of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ consisting of indices $\{\ell \mid y_\ell = y\}$. Then an empirical and regularized version of the optimization problem (12) is given for each y as follows:

$$\min_{\boldsymbol{\theta}_y} \left[\frac{1}{2}\boldsymbol{\theta}_y^\top \widehat{\boldsymbol{H}}^{(y)} \boldsymbol{\theta}_y - \boldsymbol{\theta}_y^\top \widehat{\boldsymbol{h}}^{(y)} + \delta \boldsymbol{\theta}_y^\top \boldsymbol{\theta}_y\right], \tag{13}$$

where δ (≥ 0) is the regularization parameter. $\widehat{\boldsymbol{H}}^{(y)}$ is the $n_y \times n_y$ matrix and $\widehat{\boldsymbol{h}}^{(y)}$ is the n_y-dimensional vector defined as

$$\widehat{H}^{(y)}_{\ell,\ell'} := \frac{n_y}{n^2} \sum_{i=1}^{n} L(\boldsymbol{x}_i, \boldsymbol{x}_\ell^{(y)}) L(\boldsymbol{x}_i, \boldsymbol{x}_{\ell'}^{(y)}),$$
$$\widehat{h}^{(y)}_{\ell} := \frac{1}{n} \sum_{i: y_i = y} L(\boldsymbol{x}_i, \boldsymbol{x}_\ell^{(y)}),$$

where $\boldsymbol{x}_\ell^{(y)}$ is the ℓ-th sample in class y (which corresponds to $\widehat{\theta}_\ell^{(y)}$).

A notable advantage of LSMI is that the solution $\widehat{\boldsymbol{\theta}}^{(y)}$ can be computed analytically as

$$\widehat{\boldsymbol{\theta}}^{(y)} = (\widehat{\boldsymbol{H}}^{(y)} + \delta \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}^{(y)}.$$

Then a density-ratio estimator is obtained analytically as follows:

$$\widehat{r}(\boldsymbol{x}, y) = \sum_{\ell=1}^{n_y} \widehat{\theta}_\ell^{(y)} L(\boldsymbol{x}, \boldsymbol{x}_\ell^{(y)}).$$

The accuracy of the above least-squares density-ratio estimator depends on the choice of the kernel parameter γ and the regularization parameter δ. They can be systematically optimized based on cross-validation as follows (Suzuki et al., 2009). The samples $\mathcal{Z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are divided into M disjoint subsets $\{\mathcal{Z}_m\}_{m=1}^M$ of approximately the same size. Then a density-ratio estimator $\widehat{r}_m(\boldsymbol{x}, y)$ is obtained using $\mathcal{Z} \backslash \mathcal{Z}_m$ (i.e., all samples without $\mathcal{Z}_m$), and its out-of-sample error (which corresponds to Eq.(12) without irrelevant constant) for the hold-out samples $\mathcal{Z}_m$ is computed as

$$\mathrm{CV}_m := \frac{1}{2|\mathcal{Z}_m|^2} \sum_{\boldsymbol{x}, y \in \mathcal{Z}_m} \widehat{r}_m(\boldsymbol{x}, y)^2 - \frac{1}{|\mathcal{Z}_m|} \sum_{(\boldsymbol{x}, y) \in \mathcal{Z}_m} \widehat{r}_m(\boldsymbol{x}, y).$$

This procedure is repeated for $m = 1, \ldots, M$, and the average of the above hold-out error over all m is computed. Finally, the kernel parameter γ and the regularization parameter δ that minimize the average hold-out error are chosen as the most suitable ones.

Based on the expression of SMI given by Eq.(3), an SMI estimator called LSMI is given as follows:

$$\text{LSMI} := \frac{1}{2n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i, y_i) - \frac{1}{2}, \tag{14}$$

where $\widehat{r}(\boldsymbol{x}, y)$ is a density-ratio estimator obtained above. Since $\widehat{r}(\boldsymbol{x}, y)$ can be computed analytically, LSMI can also be computed analytically.

We use LSMI for model selection of SMIC. More specifically, we compute LSMI as a function of the kernel parameter t of $K(\boldsymbol{x}, \boldsymbol{x}')$ included in the cluster-posterior model (6), and choose the one that maximizes LSMI.

MATLAB implementation of the proposed clustering method is available from 'http://sugiyama-www.cs.titech.ac.jp/~sugi/software/SMIC'.

3. Existing Methods

In this section, we qualitatively compare the proposed approach with existing methods.

3.1. Spectral Clustering

The basic idea of *spectral clustering* (Shi & Malik, 2000) is to first unfold non-linear data manifolds by a spectral embedding method, and then perform k-means in the embedded space. More specifically, given sample-sample similarity $W_{i,j} \geq 0$, the minimizer of the following criterion with respect to $\{\boldsymbol{\xi}_i\}_{i=1}^n$ is obtained under some normalization constraint:

$$\sum_{i,j}^{n} W_{i,j}\left\|\frac{1}{\sqrt{D_{i,i}}}\boldsymbol{\xi}_i - \frac{1}{\sqrt{D_{j,j}}}\boldsymbol{\xi}_j\right\|^2,$$

where $\boldsymbol{D}$ is the diagonal matrix with i-th diagonal element given by $D_{i,i} := \sum_{j=1}^{n} W_{i,j}$. Consequently, the embedded samples are given by the principal eigenvectors of $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$, followed by normalization. Note that spectral clustering was shown to be equivalent to a weighted variant of kernel k-means with some specific kernel (Dhillon et al., 2004).

The performance of spectral clustering depends heavily on the choice of sample-sample similarity $W_{i,j}$. Zelnik-Manor & Perona (2005) proposed a useful unsupervised heuristic to determine the similarity in a data-dependent manner, called *local scaling*: $W_{i,j} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma_i\sigma_j}\right)$, where σ_i is a local scaling factor defined as $\sigma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i^{(t)}\|$, and $\boldsymbol{x}_i^{(t)}$ is the t-th nearest neighbor of $\boldsymbol{x}_i$. t is the tuning parameter in the local scaling similarity, and $t = 7$ was shown to be useful (Zelnik-Manor & Perona, 2005; Sugiyama, 2007). However, this magic number '7' does not seem to work always well in general.

If $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$ is regarded as a kernel matrix, spectral clustering will be similar to the proposed SMIC method described in Section 2.3. However, SMIC does not require the post k-means processing since the principal components have clear interpretation as parameter estimates of the class-posterior model (6). Furthermore, our proposed approach provides a systematic model selection strategy, which is a notable advantage over spectral clustering.

3.2. Blurring Mean-Shift Clustering

Blurring mean-shift (Fukunaga & Hostetler, 1975) is a non-parametric clustering method based on the *modes* of the data-generating probability density.

In the blurring mean-shift algorithm, a kernel density estimator (Silverman, 1986) is used for modeling the data-generating probability density:

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K\left(\|\boldsymbol{x} - \boldsymbol{x}_i\|^2/\sigma^2\right),$$

where $K(\xi)$ is a kernel function such as a Gaussian kernel $K(\xi) = e^{-\xi/2}$. Taking the derivative of $\widehat{p}(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ and equating the derivative at $\boldsymbol{x} = \boldsymbol{x}_i$ to zero, we obtain the following updating formula for sample $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$:

$$\boldsymbol{x}_i \longleftarrow \frac{\sum_{j=1}^{n} W_{i,j}\boldsymbol{x}_j}{\sum_{j'=1}^{n} W_{i,j'}},$$

where $W_{i,j} := K'\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma^2\right)$ and $K'(\xi)$ is the derivative of $K(\xi)$. Each mode of the density is regarded as a representative of a cluster, and each data point is assigned to the cluster which it converges to.

Carreira-Perpiñán (2007) showed that the blurring mean-shift algorithm can be interpreted as an *EM algorithm* (Dempster et al., 1977), where $W_{i,j}/(\sum_{j'=1}^{n} W_{i,j'})$ is regarded as the posterior probability of the i-th sample belonging to the j-th cluster. Furthermore, the above update rule can be expressed in a matrix form as $\boldsymbol{X} \longleftarrow \boldsymbol{X}\boldsymbol{P}$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a sample matrix and $\boldsymbol{P} := \boldsymbol{W}\boldsymbol{D}^{-1}$ is a *stochastic matrix* of the random walk in a graph with adjacency $\boldsymbol{W}$ (Chung, 1997). $\boldsymbol{D}$ is defined as

$D_{i,i} := \sum_{j=1}^{n} W_{i,j}$ and $D_{i,j} = 0$ for $i \neq j$. If $\boldsymbol{P}$ is independent of $\boldsymbol{X}$, the above iterative algorithm corresponds to the *power method* (Golub & Loan, 1996) for finding the leading left eigenvector of $\boldsymbol{P}$. Then, this algorithm is highly related to the spectral clustering which computes the principal eigenvectors of $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$ (see Section 3.1). Although $\boldsymbol{P}$ depends on $\boldsymbol{X}$ in reality, Carreira-Perpiñán (2006) insisted that this analysis is still valid since $\boldsymbol{P}$ and $\boldsymbol{X}$ quickly reach a quasi-stable state.

An attractive property of blurring mean-shift is that the number of clusters is automatically determined as the number of modes in the probability density estimate. However, this choice depends on the kernel parameter σ and there is no systematic way to determine σ, which is restrictive compared with the proposed method. Another critical drawback of the blurring mean-shift algorithm is that it eventually converges to a single point (Cheng, 1995), and therefore a sensible stopping criterion is necessary in practice. Although Carreira-Perpiñán (2006) gave a useful heuristic for stopping the iteration, it is not clear whether this heuristic always works well in practice.

4. Experiments

In this section, we experimentally evaluate the performance of the proposed and existing clustering methods.

4.1. Illustration

First, we illustrate the behavior of the proposed method using artificial datasets described in the top row of Figure 1. The dimensionality is $d = 2$ and the sample size is $n = 200$. As a kernel function, we used the sparse local-scaling kernel (7) for SMIC, where the kernel parameter t was chosen from $\{1, \ldots, 10\}$ based on LSMI with the Gaussian kernel (11).

The top graphs in Figure 1 depict the cluster assignments obtained by SMIC, and the bottom graphs in Figure 1 depict the model selection curves obtained by LSMI. The results show that SMIC combined with LSMI works well for these toy datasets.

4.2. Performance Comparison

Next, we systematically compare the performance of the proposed and existing clustering methods using various real-world datasets such as images, natural languages, accelerometric sensors, and speech.

We compared the performance of the following methods, which all do not contain open tuning parameters and therefore experimental results are fair and objective: K-means (KM), spectral clustering with the self-tuning local-scaling similarity (SC) (Zelnik-Manor & Perona, 2005), mean nearest-neighbor clustering (MNN) (Faivishevsky & Goldberger, 2010), MI-based clustering for kernel logistic models (MIC) (Gomes et al., 2010) with model selection by *maximum-likelihood MI* (Suzuki et al., 2008), and the proposed SMIC.

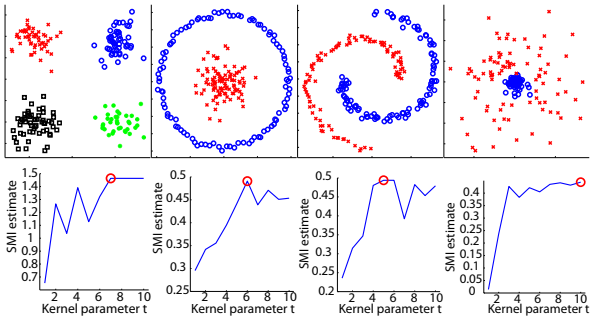

Figure 1. Illustrative examples. Cluster assignments obtained by SMIC (top) and model selection curves obtained by LSMI (bottom).

The clustering performance was evaluated by the *adjusted Rand index* (ARI) (Hubert & Arabie, 1985) between inferred cluster assignments and the ground truth categories. Larger ARI values mean better performance, and ARI takes its maximum value 1 when two sets of cluster assignments are identical. In addition, we also evaluated the computational efficiency of each method by the CPU computation time.

We used various real-world datasets including images, natural languages, accelerometric sensors, and speech: The *USPS* hand-written digit dataset ('digit'), the *Olivetti Face* dataset ('face'), the *20-Newsgroups* dataset ('document'), the *SENSEVAL-2* dataset ('word'), the *ALKAN* dataset ('accelerometry'), and the in-house speech dataset ('speech'). Detailed explanation of the datasets is omitted due to lack of space.

For each dataset, the experiment was repeated 100 times with random choice of samples from a pool. Samples were centralized and their variance was normalized in the dimension-wise manner, before feeding them to clustering algorithms.

The experimental results are described in Table 1. For the *digit* dataset, MIC and SMIC outperform KM, SC, and MNN in terms of ARI. The entire computation time of SMIC including model selection is faster than KM, SC, and MIC, and is comparable to MNN which does not include a model selection procedure. For the

Table 1. Experimental results on real-world datasets (with equal cluster size). The average clustering accuracy (and its standard deviation in the bracket) in terms of ARI and the average CPU computation time in second over 100 runs are described. The best method in terms of the average ARI and methods judged to be comparable to the best one by the *t-test* at the significance level 1% are described in boldface. Computation time of MIC and SMIC corresponds to the time for computing a clustering solution after model selection has been carried out. For references, computation time for the entire procedure including model selection is described in the square bracket.

Digit ($d = 256$, $n = 5000$, and $c = 10$)

	KM	SC	MNN	MIC	SMIC
ARI	0.42(0.01)	0.24(0.02)	0.44(0.03)	**0.63(0.08)**	**0.63(0.05)**
Time	835.9	973.3	318.5	84.4[3631.7]	14.4[359.5]

Face ($d = 4096$, $n = 100$, and $c = 10$)

	KM	SC	MNN	MIC	SMIC
ARI	0.60(0.11)	**0.62(0.11)**	0.47(0.10)	**0.64(0.12)**	**0.65(0.11)**
Time	93.3	2.1	1.0	1.4[30.8]	0.0[19.3]

Document ($d = 50$, $n = 700$, and $c = 7$)

	KM	SC	MNN	MIC	SMIC
ARI	0.00(0.00)	0.09(0.02)	0.09(0.02)	0.01(0.02)	**0.19(0.03)**
Time	77.8	9.7	6.4	3.4[530.5]	0.3[115.3]

Word ($d = 50$, $n = 300$, and $c = 3$)

	KM	SC	MNN	MIC	SMIC
ARI	0.04(0.05)	0.02(0.01)	0.02(0.02)	0.04(0.04)	**0.08(0.05)**
Time	6.5	5.9	2.2	1.0[369.6]	0.2[203.9]

Accelerometry ($d = 5$, $n = 300$, and $c = 3$)

	KM	SC	MNN	MIC	SMIC
ARI	0.49(0.04)	0.58(0.14)	**0.71(0.05)**	0.57(0.23)	**0.68(0.12)**
Time	0.4	3.3	1.9	0.8[410.6]	0.2[92.6]

Speech ($d = 50$, $n = 400$, and $c = 2$)

	KM	SC	MNN	MIC	SMIC
ARI	0.00(0.00)	0.00(0.00)	0.04(0.15)	**0.18(0.16)**	**0.21(0.25)**
Time	0.9	4.2	1.8	0.7[413.4]	0.3[179.7]

face dataset, SC, MIC, and SMIC are comparable to each other and are better than KM and MNN in terms of ARI. For the *document* and *word* datasets, SMIC tends to outperform the other methods. For the *accelerometry* dataset, MNN and SMIC work better than the other methods. Finally, for the *speech* dataset, MIC and SMIC work comparably well, and are significantly better than KM, SC, and MNN.

Overall, MIC was shown to work reasonably well, implying that model selectoin by maximum-likelihood MI is practically useful. SMIC was shown to work even better than MIC, with much less computation time. The accuracy improvement of SMIC over MIC was gained by computing the SMIC solution in a closed-form without any heuristic initialization. The computational efficiency of SMIC was brought by the analytic computation of the optimal solution and the class-wise optimization of LSMI (see Section 2.4).

The performance of MNN and SC was rather unstable because of the heuristic averaging of the number of nearest neighbors and the heuristic choice of local scaling. In terms of computation time, they are relatively efficient for small- to medium-sized datasets, but they are expensive for the largest dataset, *digit*. KM was not reliable for the *document* and *speech* datasets because of the restriction that the cluster boundaries are linear. For the *digit*, *face*, and *document* datasets, KM was computationally very expensive since a large number of iterations were needed until convergence to a local optimum solution.

Table 2. Experimental results on real-world datasets under imbalanced setup. ARI values are described in the table. Class-imbalance was realized by setting the sample size of the first class m times larger than other classes. The results for $m = 1$ are the same as the ones reported in Table 1.

Digit ($d = 256$, $n = 5000$, and $c = 10$)

	KM	SC	MNN	MIC	SMIC
$m = 1$	0.42(0.01)	0.24(0.02)	0.44(0.03)	**0.63(0.08)**	**0.63(0.05)**
$m = 2$	0.52(0.01)	0.21(0.02)	0.43(0.04)	0.60(0.05)	**0.63(0.05)**

Document ($d = 50$, $n = 700$, and $c = 7$)

	KM	SC	MNN	MIC	SMIC
$m = 1$	0.00(0.00)	0.09(0.02)	0.09(0.02)	0.01(0.02)	**0.19(0.03)**
$m = 2$	0.01(0.01)	0.10(0.03)	0.10(0.02)	0.01(0.02)	**0.19(0.04)**
$m = 3$	0.01(0.01)	0.10(0.03)	0.09(0.02)	-0.01(0.03)	**0.16(0.05)**
$m = 4$	0.02(0.01)	0.09(0.03)	0.08(0.02)	-0.00(0.04)	**0.14(0.05)**

Word ($d = 50$, $n = 300$, and $c = 3$)

	KM	SC	MNN	MIC	SMIC
$m = 1$	0.04(0.05)	0.02(0.01)	0.02(0.02)	0.04(0.04)	**0.08(0.05)**
$m = 2$	0.00(0.07)	-0.01(0.01)	0.01(0.02)	-0.02(0.05)	**0.03(0.05)**

Accelerometry ($d = 5$, $n = 300$, and $c = 3$)

	KM	SC	MNN	MIC	SMIC
$m = 1$	0.49(0.04)	0.58(0.14)	**0.71(0.05)**	0.57(0.23)	**0.68(0.12)**
$m = 2$	0.48(0.05)	0.54(0.14)	0.58(0.11)	0.49(0.19)	**0.69(0.16)**
$m = 3$	0.49(0.05)	0.47(0.10)	0.42(0.12)	0.42(0.14)	**0.66(0.20)**
$m = 4$	0.49(0.06)	0.38(0.11)	0.31(0.09)	0.40(0.18)	**0.56(0.22)**

Finally, we performed similar experiments under imbalanced setup, where the the sample size of the first class was set to be m times larger than other classes. The results are summarized in Table 2, showing that the performance of all methods tends to be degraded as the degree of imbalance increases. Thus, clustering becomes more challenging if the cluster size is imbalanced. Among the compared methods, the proposed SMIC still worked better than other methods.

Overall, the proposed SMIC combined with LSMI was shown to be a useful alternative to existing clustering approaches.

5. Conclusions

In this paper, we proposed a novel *information-maximization clustering* method, which learns class-posterior probabilities in an unsupervised manner so that the *squared-loss mutual information* (SMI) between feature vectors and cluster assignments is maximized. The proposed algorithm called *SMI-based clustering* (SMIC) allows us to obtain clustering solutions *analytically* by solving a kernel eigenvalue problem. Thus, unlike the previous information-maximization

clustering methods (Agakov & Barber, 2006; Gomes et al., 2010), SMIC does not suffer from the problem of local optima. Furthermore, we proposed to use an optimal non-parametric SMI estimator called *least-squares mutual information* (LSMI) for data-driven parameter optimization. Through experiments, SMIC combined with LSMI was demonstrated to compare favorably with existing clustering methods.

Acknowledgments

We would like to thank Ryan Gomes for providing us his program code of information-maximization clustering. MS was supported by SCAT, AOARD, and the FIRST program. MY and MK were supported by the JST PRESTO program, and HH was supported by the FIRST program.

References

Agakov, F. and Barber, D. Kernelized infomax clustering. *NIPS 18*, pp. 17–24. MIT Press, 2006.

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1):131–142, 1966.

Bach, F. and Harchaoui, Z. DIFFRAC: A discriminative and flexible framework for clustering. *NIPS 20*, pp. 49–56, 2008.

Carreira-Perpiñán, M. Á. Fast nonparametric clustering with Gaussian blurring mean-shift. *ICML*, pp. 153–160, 2006.

Carreira-Perpiñán, M. Á. Gaussian mean shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:767–776, 2007.

Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:790–799, 1995.

Chung, F. R. K. *Spectral Graph Theory*. American Mathematical Society, Providence, 1997.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, Inc., 2nd edition, 2006.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

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

Dhillon, I. S., Guan, Y., and Kulis, B. Kernel k-means, spectral clustering and normalized cuts. *ACM SIGKDD*, pp. 551–556, 2004.

Faivishevsky, L. and Goldberger, J. A nonparametric information theoretic clustering algorithm. *ICML*, pp. 351–358, 2010.

Fukunaga, K. and Hostetler, L. D. The estimation of the gradient of a density function, with application in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

Girolami, M. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13(3): 780–784, 2002.

Golub, G. H. and Loan, C. F. Van. *Matrix Computations*. Johns Hopkins University Press, 1996.

Gomes, R., Krause, A., and Perona, P. Discriminative clustering by regularized information maximization. *NIPS 23*, pp. 766–774. 2010.

Horn, R. A. and Johnson, C. A. *Matrix Analysis*. Cambridge University Press, 1985.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

Song, L., Smola, A., Gretton, A., and Borgwardt, K. A dependence maximization view of clustering. *ICML*, pp. 815–822, 2007.

Sugiyama, M. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. Approximating mutual information by maximum likelihood density ratio estimation. *JMLR Workshop and Conference Proceedings*, 4:5–20, 2008.

Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52, 2009.

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. Maximum margin clustering. *NIPS 17*, pp. 1537–1544. 2005.

Zelnik-Manor, L. and Perona, P. Self-tuning spectral clustering. *NIPS 17*, pp. 1601–1608, 2005.