Maximum Within-Cluster Association

Yongjin Lee, Seungjin Choi *

Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea

Abstract

This paper addresses a new method and aspect of information-theoretic clustering where we exploits the minimum entropy principle and the quadratic distance measure between probability densities. We present a new minimum entropy objective function which leads to the maximization of *within-cluster association*. A simple implementation using the gradient ascent method is given. In addition, we show that the minimum entropy principle leads to the objective function of the k-means clustering, and the maximum within-cluster association is closed related to the spectral clustering which is an eigen-decomposition-based method. This information-theoretic view of spectral clustering leads us to use the kernel density estimation method in constructing an affinity matrix.

Key words: Clustering, Information-theoretic learning, Minimum entropy, Spectral clustering

1 Introduction

Clustering is a procedure which partitions a set of unlabelled data into natural groups. When clustering is carried out successfully, data points in the same group are expected to be similar each other but dissimilar from data samples in different groups. A natural question arises, "what is a good measure of similarity or dissimilarity between data points?". Depending on similarity or dissimilarity measure, a variety of clustering algorithms with different characteristics, have been developed (Jain et al., 1999). For example, k-means clustering can be interpreted as a method for minimizing the sum of pairwise

* Corresponding author. Tel.: +82-54-279-2259; Fax: +82-54-279-2299
Email: seungjin@postech.ac.kr (S. Choi)

intra-cluster Euclidean distances (Roth et al., 2003). This implicitly implies that k-means clustering method makes the assumption of Gaussian distribution for data and exploits only second-order statistics. It means that we cannot extract all information available from the set of data when the probability density of the data is not Gaussian. This might be a restrictive assumption.

Information-theoretic methods are attractive and powerful but they involve the probability density estimation, which might be cumbersome in the viewpoint of computational complexity. Approaches to density estimation are categorized into three different methods, including parametric, semi-parametric, and non-parametric methods. A parametric method assumes a specific parameterized functional form of probability density. It is computationally less expensive but less flexible. Semi-parametric methods (for example, mixture of Gaussians) are more flexible but the estimation is not trivial. The Parzen window method is one of widely-used non-parametric density estimation methods. It is the most flexible but computationally very expensive when entropy or divergence calculation is involved. It was suggested in (Principe et al., 2000) that Renyi's quadratic entropy and quadratic distance measures between densities simplified entropy or divergence calculation in the framework of Parzen window-based density estimation. Along this line, a variety of unsupervised learning algorithms (Principe et al., 2000) and a feature transformation (Torkkola and Campbell, 2000) were developed.

In this paper, we address an information-theoretic clustering which mainly exploits the minimum entropy principle studied in (Roberts et al., 2000, 2001) where the clustering is carried out by minimizing the overlap between densities of clusters. In the minimum entropy data partitioning (Roberts et al., 2000, 2001), Kullback-Liebler (KL) divergence was used to measure the overlap between cluster densities and the minimization was performed through grouping mixtures of Gaussian. Density estimation through mixture of Gaussians is sensitive to initial conditions and the number of mixture components must be carefully decided, which is a difficult problem. In contrast, we employ the Renyi's quadratic entropy and the quadratic distance measure (Principe et al., 2000) with the Parzen window density estimation, in order to avoid the original difficulties. We show that the minimum entropy principle with the Renyi's quadratic entropy, leads to an objective function of the k-means clustering method. We also show that minimizing the overlap between cluster densities with the quadratic distance measure, leads to the maximization of *within-cluster association*. In addition, we further show that the maximum within-cluster association is closely related to a spectral clustering method where the clustering is carried out through the eigen-decomposition of a properly chosen matrix (affinity matrix). This information-theoretic view of spectral clustering leads us to use the kernel density estimation method in constructing an affinity matrix, which might be considered as a main advantage resulting from our information-theoretic view.

2 Minimum Entropy Data Partitioning

We begin by briefly reviewing the method of minimum entropy (or maximum certainty) data partitioning (Roberts et al., 2000, 2001) since this idea is a starting point for our method. In the maximum certainty data partitioning, one constructs candidate partition models for data sets in such a way that the overlap between partitions is minimized.

Let us consider a partitioning of the data into a set of Kclusters. The probability density function of a single datum $\boldsymbol{x}$, conditioned on a set of K partitions, is given by

$$p(\boldsymbol{x}) = \sum_{i=1}^{K} p(\boldsymbol{x}|i)p(i), \tag{1}$$

where $p(i)$ is the prior probability for the ith partition.

The overlap between the unconditional density $p(\boldsymbol{x})$ and the contribution to this density function of the ith partition, $p(\boldsymbol{x}|i)$, is measured by Kullback-Liebler (KL) divergence between these two distributions:

$$\mathcal{V}_i = -KL\left[p(\boldsymbol{x}|i)||p(\boldsymbol{x})\right], \tag{2}$$

which is upper-bounded by 0 (since KL divergence is always nonnegative). When the ith class is well-separated from all others, $\mathcal{V}_i$ is minimized.

The total overlap over a set of K partitions, $\mathcal{V}$, is defined by

$$\begin{aligned}
\mathcal{V} &\stackrel{\triangle}{=} \sum_{i=1}^{K} p(i)\mathcal{V}_i \\
&= -\sum_{i=1}^{K} p(i) KL[p(\boldsymbol{x}|i)||p(\boldsymbol{x})] \\
&= -\sum_{i=1}^{K} p(i) \int p(\boldsymbol{x}|i) \log\left(\frac{p(\boldsymbol{x}|i)}{p(\boldsymbol{x})}\right) d\boldsymbol{x}.
\end{aligned} \tag{3}$$

It follows from Bayes' theorem that Eq. (3) can be rewritten as

$$
\begin{aligned}
\mathcal{V} &= -\sum_{i=1}^{K} \int p(i|\boldsymbol{x}) p(\boldsymbol{x}) \log \left(\frac{p(i|\boldsymbol{x})}{p(i)} \right) d\boldsymbol{x} \\
&= -\int p(\boldsymbol{x}) \left(\sum_{i=1}^{K} p(i|\boldsymbol{x}) \log(p(i|\boldsymbol{x}) \right) d\boldsymbol{x} + \sum_{i=1}^{K} p(i) \log p(i) \\
&= \underbrace{\left[\int p(\boldsymbol{x}) H(i|\boldsymbol{x}) d\boldsymbol{x} \right]}_{\text{expected posterior entropy}} + \underbrace{[-H(i)]}_{\text{negative prior entropy}} . \qquad (4)
\end{aligned}
$$

The total overlap measure $\mathcal{V}$ consists of the expected (Shannon's) entropy of the class posteriors and the negative entropy of the priors. Therefore minimizing $\mathcal{V}$ is equivalent to minimizing the expected entropy of the partitions given a set of observed variables. An ideal data partitioning separates the data such that the overlap between partitions is minimal. The expected entropy of the partitions reaches its minimum when for each datum, some partition posteriors are close to unity, while all the others are close to zero (Roberts et al., 2000, 2001).

Alternatively, we can rewrite the total overlap measure $\mathcal{V}$ in Eq. (3) as

$$
\begin{aligned}
\mathcal{V} &= -\sum_{i=1}^{K} p(i) \int p(\boldsymbol{x}|i) \left[\log p(\boldsymbol{x}|i) - \log p(\boldsymbol{x}) \right] d\boldsymbol{x} \\
&= -\sum_{i=1}^{K} \int p(\boldsymbol{x}|i) \log p(\boldsymbol{x}|i) d\boldsymbol{x} + \int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \\
&= -\left[H(\boldsymbol{x}) - \sum_{i=1}^{K} p(i) H(\boldsymbol{x}|i) \right] . \qquad (5)
\end{aligned}
$$

Minimizing the total overlap measure is equivalent to minimizing the expected entropy of the class-conditional density.

3 Revisit of k-Means

In this section, we briefly review the Renyi's quadratic entropy and show that an objective function of k-means can be approximately derived in the framework of the minimum entropy principle and the Renyi's quadratic entropy (Lee and Choi, 2004).

For a continuous random variable $\boldsymbol{x} \in \mathbb{R}^d$ whose realization is given by $\{\boldsymbol{x}_n\}_{n=1}^N$, where N is the number of data points, the probability density of $\boldsymbol{x}$ estimated by the Parzen window using a Gaussian kernel is given by

$$p(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} G(\boldsymbol{x}; \boldsymbol{x}_n, \sigma^2), \tag{6}$$

where

$$G(\boldsymbol{x}; \boldsymbol{x}_n, \sigma^2) = \frac{1}{(2\pi\sigma)^{d/2}} \exp\left\{-\frac{||\boldsymbol{x} - \boldsymbol{x}_n||^2}{2\sigma^2}\right\}. \tag{7}$$

The Renyi's entropy of order α is defined as

$$H_{R_\alpha} = \frac{1}{1-\alpha} \log \int p^\alpha(\boldsymbol{x}) d\boldsymbol{x}. \tag{8}$$

The Shannon's entropy is a limiting case of Renyi's entropy as $\alpha \to 1$. For $\alpha = 2$, Renyi's entropy (8) is called *Renyi's quadratic entropy*, H_{R_2}, that has the form

$$H_{R_2} = -\log \int p^2(\boldsymbol{x}) d\boldsymbol{x}, \tag{9}$$

where the scaling factor $\frac{1}{2}$ is neglected.

Note that the convolution of two Gaussian is again Gaussian, i.e.,

$$\int G\left(\boldsymbol{x}; \boldsymbol{x}_n, \sigma^2\right) G\left(\boldsymbol{x}; \boldsymbol{x}_m, \sigma^2\right) d\boldsymbol{x} = G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right). \tag{10}$$

It follows from this relation that the Renyi's quadratic entropy with the Parzen window density estimation, leads to

$$\int p^2(\boldsymbol{x}) d\boldsymbol{x} = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} G(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2). \tag{11}$$

Thus the Renyi's quadratic entropy can be easily computed as a sum of local interactions as defined by the kernel, over all pairs of samples (Principe et al., 2000; Torkkola and Campbell, 2000).

The k-means clustering partitions the data in such a way that the sum of intra-cluster distances to the prototype vectors is minimized. The objective function of k-means for the case of K clusters, is given by

$$\mathcal{J}_{km} = \sum_{i=1}^{K} \sum_{n=1}^{N} z_{in} \|\boldsymbol{x}_n - \boldsymbol{\mu}_i\|^2, \tag{12}$$

where z_{in} is an indicating variable defined as

$$z_{in} = \begin{cases} 1 \text{ if } \boldsymbol{x}_n \in C_i \\ 0 \text{ otherwise} \end{cases},$$

where C_i denotes the ith cluster and

$$\begin{aligned} \boldsymbol{\mu}_i &= \frac{1}{N_i} \sum_{n=1}^{N} z_{in} \boldsymbol{x}_n, \\ N_i &= \sum_{n=1}^{N} z_{in}. \end{aligned}$$

This is equivalent to minimizing the sum of pairwise intra-cluster distances (Roth et al., 2003) that is defined by

$$\mathcal{J}_{km} = \frac{1}{2} \sum_{i=1}^{K} \frac{1}{N_i} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} \|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2. \tag{13}$$

Now we show that the objective function (13) can be approximately derived from the minimum entropy rule by employing the Renyi's quadratic entropy and the Parzen window method. Using indicator variables $\{z_{in}\}$, we write the ith class conditional density as

$$p(\boldsymbol{x}|i) = \frac{1}{N_i} \sum_{n=1}^{N_i} z_{in} G(\boldsymbol{x}; \boldsymbol{x}_n, \sigma^2). \tag{14}$$

Intuitively, the indicator variable can be considered as the posterior over the class variable, i.e., $z_{in} = p\,(i|\boldsymbol{x}_n)$. Neglect an irrelevant term $H(\boldsymbol{x})$ in (5), then the objective function (5) becomes

$$
\begin{aligned}
\mathcal{V} &= \sum_{i=1}^{K} p(i) H(\boldsymbol{x}|i) \\
&= -\sum_{i=1}^{K} p(i) \log \left[\int p^2(\boldsymbol{x}|i) d\boldsymbol{x} \right] \\
&= -\sum_{i=1}^{K} \frac{N_i}{N} \log \left[\frac{1}{N_i^2} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} G(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2) \right].
\end{aligned}
\tag{15}
$$

Minimizing (15) is equivalent to maximizing $\mathcal{L}$ which is given by

$$
\mathcal{L} = \sum_{i=1}^{K} \frac{N_i}{N} \log \left[\frac{1}{N_i^2} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} G(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2) \right],
\tag{16}
$$

which is lower-bounded by

$$
\begin{aligned}
\mathcal{L} &\geq \sum_{i=1}^{K} \frac{N_i}{N} \frac{1}{N_i^2} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} \log \left[G(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2) \right] \\
&= \mathcal{L}_l,
\end{aligned}
\tag{17}
$$

where the Jensen's inequality was used.

Then the lower-bound $\mathcal{L}_l$ is given by

$$
\begin{aligned}
\mathcal{L}_l = &-\frac{1}{2\sigma^2} \sum_{i=1}^{K} \frac{1}{N_i} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} \|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2 \\
&-\sum_{i=1}^{K} \frac{1}{N_i} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} \log(2\pi\sigma)^{d/2}.
\end{aligned}
\tag{18}
$$

Hence the maximization of this lower-bound $\mathcal{L}_l$ is equivalent to the minimization of the sum of pairwise intra-cluster distances in (13).

4 Minimum Entropy and Spectral Clustering

In this section we present the *maximum within-cluster association* that is derived using the quadratic distance measure instead of the KL divergence in the framework of the minimum entropy data partitioning. Then we show that

the maximum within-cluster association is closely related to the *average association* which belongs to a class of spectral clustering methods where the clustering is based on the eigen-decomposition of an affinity matrix.

4.1 Quadratic Distance Measure

Principe *et al.* proposed a quadratic distance measure between probability densities which resembles the Euclidean distance between two vectors (Principe et al., 2000). A main motivation of the quadratic distance measure lies in its simple form for the case where the Parzen window density estimation with Gaussian kernel is involved.

The squared Euclidean distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$\|\boldsymbol{x}-\boldsymbol{y}\|^2 = (\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y}) = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\boldsymbol{x}^T\boldsymbol{y}, \tag{19}$$

which is always non-negative. In a similar manner, the quadratic distance between two probability densities, $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ is defined by

$$D[f\|g] = \int f^2(\boldsymbol{x})\,d\boldsymbol{x} + \int g^2(\boldsymbol{x})\,d\boldsymbol{x} - 2\int f(\boldsymbol{x})\,g(\boldsymbol{x})\,d\boldsymbol{x} \geq 0. \tag{20}$$

The quadratic distance measure is always non-negative and it becomes zero only when $f(\boldsymbol{x}) = g(\boldsymbol{x})$.

When density functions are replaced by their associated Parzen window estimates, the quadratic distance measure is simplified as

$$\begin{aligned} D[f\|g] &= \int f^2(\boldsymbol{x})\,d\boldsymbol{x} + \int g^2(\boldsymbol{x})\,d\boldsymbol{x} - 2\int f(\boldsymbol{x})\,g(\boldsymbol{x})\,d\boldsymbol{x} \\ &= \frac{1}{N_f^2}\sum_{n=1}^{N_f}\sum_{m=1}^{N_f} G\left(\boldsymbol{x}_n^f;\boldsymbol{x}_m^f,2\sigma^2\right) + \frac{1}{N_g^2}\sum_{n=1}^{N_g}\sum_{m=1}^{N_g} G\left(\boldsymbol{x}_n^g;\boldsymbol{x}_m^g,2\sigma^2\right) \\ &\quad -2\frac{1}{N_f}\cdot\frac{1}{N_g}\sum_{n=1}^{N_f}\sum_{m=1}^{N_g} G\left(\boldsymbol{x}_n^f;\boldsymbol{x}_m^g,2\sigma^2\right), \end{aligned} \tag{21}$$

where

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{N_f}\sum_{n=1}^{N_f} G\left(\boldsymbol{x};\boldsymbol{x}_n^f,\sigma^2\right), \\ g(\boldsymbol{x}) &= \frac{1}{N_g}\sum_{n=1}^{N_g} G\left(\boldsymbol{x};\boldsymbol{x}_n^g,\sigma^2\right), \end{aligned} \tag{22}$$

and $\{\boldsymbol{x}_n^f\}$ and $\{\boldsymbol{x}_n^g\}$ are the set of data points (the number of data points are denoted by N_f and N_g) that were used to evaluate f and g, respectively.

4.2 Maximum Within-Cluster Association

In Eq. (2), the overlap between the unconditional density $p(\boldsymbol{x})$ and the contribution to this density function of the ith partition, $p(\boldsymbol{x}|i)$, was measured by KL divergence between them in Roberts et al. (2000, 2001). Now we compute this overlap using the quadratic distance between densities estimated by the Parzen window method.

As in the previous section, the ith class conditional density is written as

$$p(\boldsymbol{x}|i) = \frac{1}{N_i} \sum_{n=1}^{N_i} z_{in} G(\boldsymbol{x}; \boldsymbol{x}_n, \sigma^2). \tag{23}$$

Incorporate the density estimated by the Parzen window method into the quadratic distance measure, then Eq.(2) becomes

$$\begin{aligned}
\mathcal{V}_i =& -D\left[p(\boldsymbol{x}|i)||p(\boldsymbol{x})\right] \\
=& -\int p^2\left(\boldsymbol{x}|i\right) d\boldsymbol{x} - \int p^2\left(\boldsymbol{x}\right) d\boldsymbol{x} + 2\int p\left(\boldsymbol{x}|i\right) p\left(\boldsymbol{x}\right) d\boldsymbol{x} \\
=& -\frac{1}{N_i^2} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right) - \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right) \\
&+ \frac{2}{NN_i} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right) \\
=& -\frac{1}{N_i^2} \boldsymbol{z}_i^T \boldsymbol{G} \boldsymbol{z}_i - \frac{1}{N^2} \mathbf{1}^T \boldsymbol{G} \mathbf{1} + \frac{2}{NN_i} \boldsymbol{z}_i^T \boldsymbol{G} \mathbf{1},
\end{aligned} \tag{24}$$

where $\boldsymbol{G} \in \mathbb{R}^{N \times N}$ is a kernel matrix whose (n, m)-element is $[\boldsymbol{G}]_{nm} = G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right)$ and $\boldsymbol{z}_i \in \mathbb{R}^N$ $(i = 1, \ldots, K)$ are the indicator variable vectors, i.e., $[\boldsymbol{z}_i]_n = z_{in}$.

Then the total overlap can be written as

$$\begin{aligned}
\mathcal{V} =& \sum_{i=1}^{K} p\left(i\right) \mathcal{V}_i \\
=& -\sum_{i=1}^{K} p\left(i\right) D\left[p\left(\boldsymbol{x}|i\right) \| p\left(\boldsymbol{x}\right)\right] \\
=& \frac{1}{N} \left[\frac{\mathbf{1}^T \boldsymbol{G} \mathbf{1}}{N} - \sum_{i=1}^{K} \frac{\boldsymbol{z}_i^T \boldsymbol{G} \boldsymbol{z}_i}{N_i} \right],
\end{aligned} \tag{25}$$

where $p(i) = \frac{N_i}{N}$.

This is reminiscent of Eq. (5). The first term in Eq. (25) is constant and the second term can be considered as a within-cluster association. Therefore, the minimization of overlap between partitions leads to the maximization of within-cluster association $\mathcal{L}_{wca}$ which is defined as

$$\begin{aligned}\mathcal{L}_{wca} &= \sum_{i=1}^{K} \frac{1}{N_i} \sum_{n=1}^{N} \sum_{m=1}^{N} z_{in} z_{im} G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right) \\ &= \sum_{i=1}^{K} \frac{\boldsymbol{z}_i^T \boldsymbol{G} \boldsymbol{z}_i}{N_i}. \end{aligned} \tag{26}$$

Remark: It may be interesting to compare this with k-means clustering which uses only second order statistics and implicitly assumes Gaussian distribution for cluster densities. It follows from (13) that k-means clustering partitions the data in such a way that the dissimilarity within cluster with Euclidean distance measure is minimized. On the other hand, within-cluster association criterion in Eq. (26) looks for partitions which maximizes the similarity within cluster with Gaussian kernel $G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right)$ being used as a similarity measure. Replacing $G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right)$ by $\log G\left(\boldsymbol{x}_n; \boldsymbol{x}_m, 2\sigma^2\right)$, the maximization of within-cluster association leads to the minimization of pairwise intra-cluster distances in Eq. (13).

4.3 Algorithm

The indicator variables, $\{z_{in}\}$, are adjusted in such a way that the within-cluster association (26) is maximized. We derive a simple gradient ascent algorithm which iteratively finds a maximum of (26). Since the indicator variables are bounded in the interval [0,1], we parameterize them using a softmax function

$$z_{in} = \frac{\exp\left[\theta_{in}\right]}{\sum_{c=1}^{K} \exp\left[\theta_{cn}\right]}. \tag{27}$$

The gradient of Eq. (26) with respect to θ_{in} is given by

$$\frac{\partial \mathcal{L}_{wca}}{\partial \theta_{in}} = \sum_{j=1}^{K} \frac{\partial \mathcal{L}_j}{\partial z_{jn}} \cdot \frac{\partial z_{jn}}{\partial \theta_{in}}, \tag{28}$$

where

$$
\begin{aligned}
\mathcal{L}_j &= \frac{\boldsymbol{z}_j^T \boldsymbol{G} \boldsymbol{z}_j}{N_j}, \\
\frac{\partial \mathcal{L}_j}{z_{jn}} &= \frac{2}{N_j} \sum_{m=1}^{N} z_{jm} G\left(\boldsymbol{x}_n, \boldsymbol{x}_m, 2\sigma^2\right) - \frac{1}{N_j} \mathcal{L}_j, \\
\frac{\partial z_{jn}}{\theta_{in}} &= z_{jn}\delta_{ij} - z_{in} z_{jn},
\end{aligned} \tag{29}
$$

where δ_{ij} is the Kronecker delta equal to 1 if $i = j$, otherwise it is zero.

Therefore, the updating rule for θ_{in} is given by

$$
\begin{aligned}
\theta_{in}^{(k+1)} &= \theta_{in}^{(k)} + \eta \frac{1}{2} \frac{\partial \mathcal{L}_{wca}}{\partial \theta_{in}} \\
&= \theta_{in}^{(k)} + \eta \sum_{j=1}^{K} \frac{1}{N_j} \left(\sum_{m=1}^{N} z_{jm} G\left(\boldsymbol{x}_n, \boldsymbol{x}_m, 2\sigma^2\right) - \mathcal{L}_j \right) (z_{jn}\delta_{ij} - z_{in} z_{jn}),
\end{aligned} \tag{30}
$$

where $\eta > 0$ is the learning rate for $i = 1, \ldots, K$ and $n = 1, \ldots, N$.

5 From Maximum Within-Cluster Association to Spectral Clustering

Here we consider a special case in our maximum within-cluster association. In the case of two clusters, we show that the maximization of (26) leads to one of well-known spectral clustering criterion, *average association* in (Shi and Malik, 2000). In other words, in such a case, the indicator variables in Eq. (26) can be easily computed by the eigne-decomposition method. This, in fact, provides an information-theoretic view to spectral clustering.

In the case of two clusters, the class-conditional densities estimated by Parzen window method, can be written as

$$
p\left(\boldsymbol{x}|1\right) = \frac{1}{N_1} \sum_{n=1}^{N} z_n G\left(\boldsymbol{x}, \boldsymbol{x}_n, \sigma^2\right), \tag{31}
$$

$$
p\left(\boldsymbol{x}|2\right) = \frac{1}{N_2} \sum_{n=1}^{N} \left(1 - z_n\right) G\left(\boldsymbol{x}, \boldsymbol{x}_n, \sigma^2\right), \tag{32}
$$

where $\{z_n\}$ are indicator variables defined by

$$z_n = \begin{cases} 1 \text{ if } \boldsymbol{x}_n \in C_1 \\ 0 \text{ otherwise} \end{cases},$$

and $N_1 = \sum_{n=1}^{N} z_n, N_2 = \sum_{n=1}^{N} (1 - z_n)$.

Then our maximum within-cluster association criterion in Eq. (26) can be written as

$$\mathcal{L}_{wca} = \frac{\boldsymbol{z}^T \boldsymbol{G} \boldsymbol{z}}{\boldsymbol{z}^T \boldsymbol{z}} + \frac{(\boldsymbol{1} - \boldsymbol{z})^T \boldsymbol{G} (\boldsymbol{1} - \boldsymbol{z})}{(\boldsymbol{1} - \boldsymbol{z})^T (\boldsymbol{1} - \boldsymbol{z})}, \tag{33}$$

where $\boldsymbol{1} \in \mathbb{R}^N$ by $[\boldsymbol{1}]_n = 1$. Introducing another indicator variables, $\{y_n\}$, defined by

$$y_n = \begin{cases} +1 \text{ if } \boldsymbol{x}_n \in C_1 \\ -1 \text{ otherwise} \end{cases}.$$

With these indicator variables, Eq. (33) can be rewritten as

$$4\mathcal{L}_{wca} = \frac{(\boldsymbol{1} + \boldsymbol{y})^T \boldsymbol{G} (\boldsymbol{1} + \boldsymbol{y})}{\frac{1}{4} (\boldsymbol{1} + \boldsymbol{y})^T (\boldsymbol{1} + \boldsymbol{y})} + \frac{(\boldsymbol{1} - \boldsymbol{y})^T \boldsymbol{G} (\boldsymbol{1} - \boldsymbol{y})}{\frac{1}{4} (\boldsymbol{1} - \boldsymbol{y})^T (\boldsymbol{1} - \boldsymbol{y})}. \tag{34}$$

Adopting a similar method that used in (Shi and Malik, 2000), the maximum within-cluster association reduces to the following simple optimization problem:

$$\arg\max_{\boldsymbol{t}^T \boldsymbol{1} = 0} \frac{\boldsymbol{t}^T \boldsymbol{G} \boldsymbol{t}}{\boldsymbol{t}^T \boldsymbol{t}} \tag{35}$$

where $\boldsymbol{t} = (\boldsymbol{1} + \boldsymbol{y}) - \frac{N_1}{N_2} (\boldsymbol{1} - \boldsymbol{y})$ and $[\boldsymbol{y}]_n = y_n$, which is a N-dimensional vector. See Appendix for detailed derivation. Indicator variables are estimated through the eigenvector associated with the largest eigenvalue of the matrix $\boldsymbol{G}$, as in spectral clustering methods.

6 Numerical Experiments

The information-theoretic view of the spectral clustering, leads us to use the kernel density estimation methods in constructing an affinity matrix or in

determining the kernel size for the maximum within-cluster association. In fact, this clarifies a guideline of how an appropriate affinity matrix should be constructed for successful clustering. After a brief overview of methods of determining a kernel size, we present two numerical experimental results, in order to compare two objective functions for k-means and the maximum within-cluster association.

6.1 Determination of a Kernel Size

One of drawback of spectral clustering is that there is no theoretical guideline to choose a kernel size to construct an affinity matrix. In the previous section, we derived the similarity function between data points from the Parzen window density estimation. This gives a clue to solve the problem in the view point of density estimation. A good discussion on the kernel size in the Parzen Window method can be found in (Duda et al., 2001) and a variety of practical methods can be found in (Turlach, 1993). Among those methods, the simplest method is to set σ as

$$\sigma = s1.06N^{-0.2}, \tag{36}$$

or

$$\sigma = 1.06 \min\left(s, \frac{R}{1.34}\right) N^{-0.2}, \tag{37}$$

where $s^2 = \frac{1}{d}\sum_i S_{ii}$, S_{ii} are the diagonal elements of the sample covariance matrix, and R is interquartile range. The latter one is known as more robust to outlier in density estimation. In (Jenssen et al., 2004), which also discussed spectral clustering in terms of information theory, Eq.(36) is used to choose the kernel size. Another easy method is to use cross validation technique. We can select the kernel size which maximize the log-likelihood for the validation set. To reflect highly nonlinear structure of data, manifold Parzen window can be used (Vincent and Bengio, 2003). Manifold Parzen window choose the kernel size at each data point using the local covariance. In this paper, we use Eq.(36) in experiments for its simplicity.

6.2 Experiment 1: Two Clusters

In the first experiment, we consider the ring data (see Fig. 1) where samples are drawn from two generator distributions: (1) an isotropic Gaussian distribution for inner cluster; (2) a uniform ring distribution for outer cluster. A total of

200 data points were drawn from each distribution, which gives $N = 400$. We construct a Gaussian kernel matrix $\boldsymbol{G}$ (which is also known as affinity matrix in spectral clustering) and compute the first eigenvector associated with the largest eigenvalue of $\boldsymbol{G}$. The largest eigenvector of $\boldsymbol{G}$ as shown in Fig. 2, clearly exhibits discrimination. The average value over the elements in the largest eigenvector of $\boldsymbol{G}$ is used in thresholding out in clustering. Successful clustering result in shown in Fig. 1. On the other hand, two clusters share the same mean vector. Hence, k-means clustering method fails to correctly partition the data (see Fig. 3) because it is based on the Euclidean distance between the data point and the mean vector.

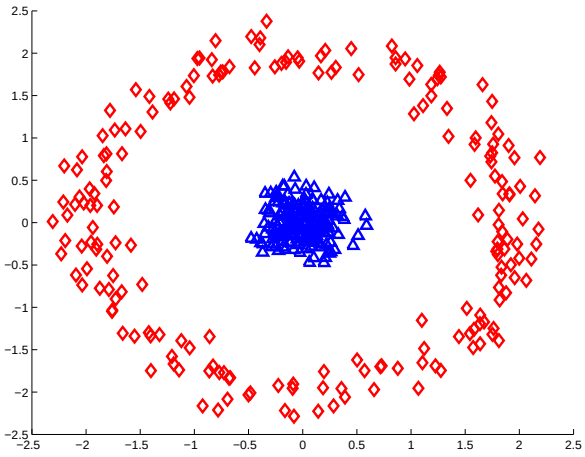

Fig. 1. Clustering result by our method in Experiment 1.

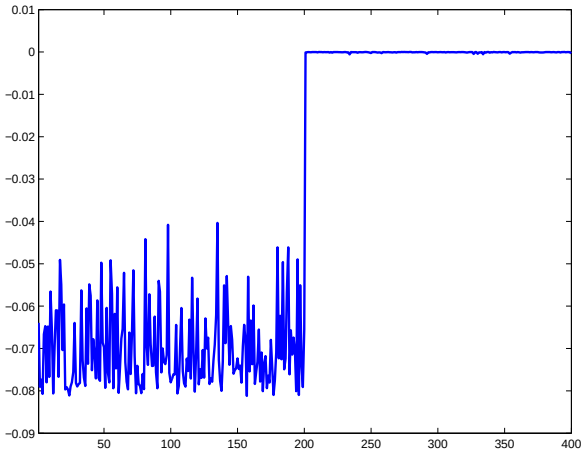

Fig. 2. The largest eigenvector of $\boldsymbol{G}$ in Experiment 1.

6.3 Experiment 2: Three Clusters

We applied the gradient-based algorithm in Eq. (30) in a ellipse-shaped ring data with three clusters, where samples were drawn from three generator distributions: (1) isotropic Gaussian distributions with different mean for two

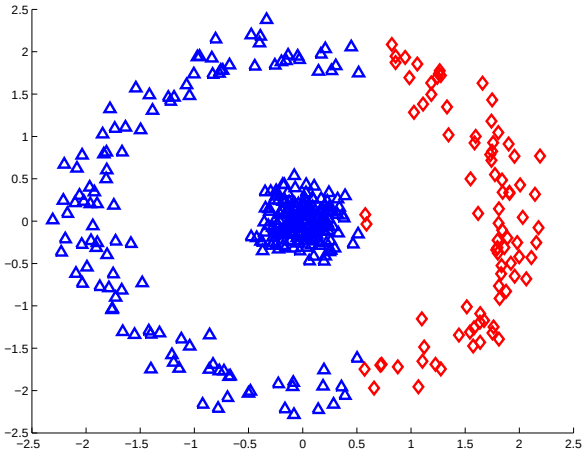

Fig. 3. Clustering result by k-means in Experiment 1.

inner clusters; (2) a uniform ellipse-shaped ring for outer cluster (see Fig. 4). Posterior probabilities over class variables (which correspond to estimated indicated variables) are shown in Fig. 5. A correct clustering result using our own algorithm (30) is shown in Fig. 4, whereas k-means clustering has difficulty in grouping the data successfully.

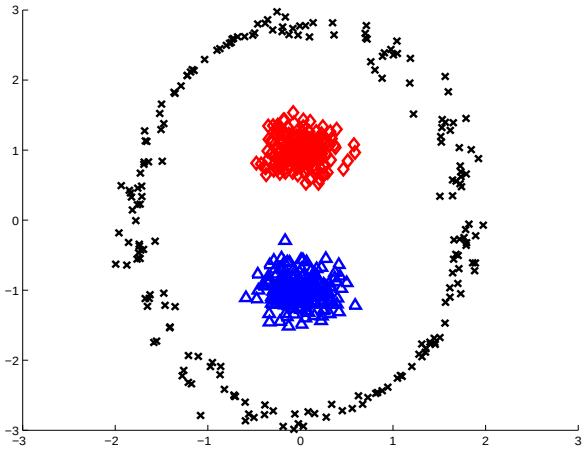

Fig. 4. Clustering result by our method in Experiment 2.

7 Discussion

We started from the idea of the minimum entropy data partitioning (Roberts et al., 2000, 2001) where the goal of clustering is viewed as the minimization of the overlap between cluster densities. The KL divergence in the overlap measure was used in (Roberts et al., 2000, 2001), which required a heavy computation for density estimation. Following this minimum entropy principle, we employed the Renyi's quadratic entropy and the quadratic distance measure with the Parzen window density estimation, which was introduced in

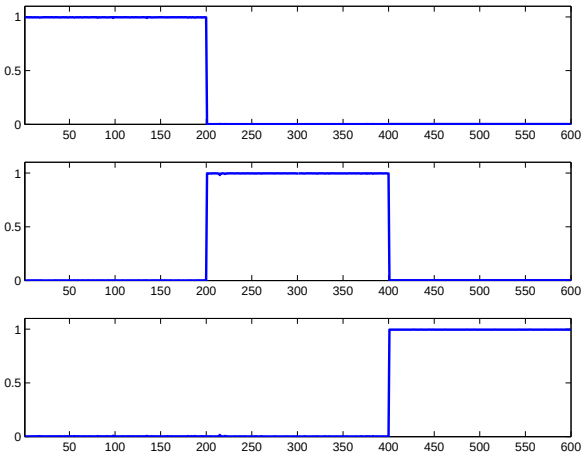

Fig. 5. Posterior probabilities over class variables (corresponding to estimated indicator variables in Experiment 2.

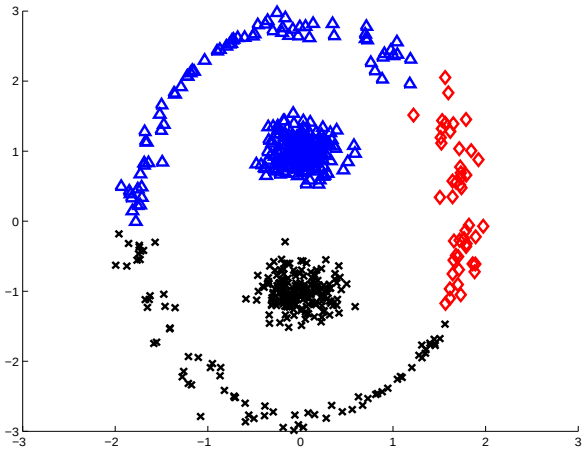

Fig. 6. Clustering result by k-means.

Principe et al. (2000). Adopting the Renyi's quadratic entropy leaded to the objective function of Kk-meas, and the quadratic distance measure leaded to the *maximum within-cluster association*. In a special case (two clusters), we showed that our *maximum within-cluster association* was closely related to one of well-known spectral clustering methods which are based on the eigen-decomposition. This information-theoretic view of spectral clustering justified that the construction of an appropriate affinity matrix could be done through the kernel density estimation.

A Detailed Derivation of Eq. (35)

Let $k = \frac{N_1}{N}$, then Eq. (34) becomes

$$
\begin{aligned}
4\mathcal{L}_{wca} &= \frac{(\mathbf{1}+\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}+\boldsymbol{y})}{k\mathbf{1}^T\mathbf{1}} + \frac{(\mathbf{1}-\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}-\boldsymbol{y})}{(1-k)\mathbf{1}^T\mathbf{1}} \\
&= \frac{(1-k)(\mathbf{1}+\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}+\boldsymbol{y})}{k(1-k)\mathbf{1}^T\mathbf{1}} + \frac{k(\mathbf{1}-\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}-\boldsymbol{y})}{k(1-k)\mathbf{1}^T\mathbf{1}} \\
&= \frac{(1-k)(\mathbf{1}^T\boldsymbol{G}\mathbf{1} + 2\mathbf{1}^T\boldsymbol{G}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{G}\boldsymbol{y})}{k(k-1)\mathbf{1}^T\mathbf{1}} + \frac{k(\mathbf{1}^T\boldsymbol{G}\mathbf{1} - 2\mathbf{1}^T\boldsymbol{G}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{G}\boldsymbol{y})}{k(k-1)\mathbf{1}^T\mathbf{1}} \\
&= \frac{\boldsymbol{y}^T\boldsymbol{G}\boldsymbol{y} + \mathbf{1}^T\boldsymbol{G}\mathbf{1}}{k(1-k)\mathbf{1}^T\mathbf{1}} + \frac{2(1-2k)\mathbf{1}^T\boldsymbol{G}\boldsymbol{y}}{k(1-k)\mathbf{1}^T\mathbf{1}}.
\end{aligned}
$$

Define $\alpha(\boldsymbol{y}) = \boldsymbol{y}^T\boldsymbol{G}\boldsymbol{y}$, $\beta(\boldsymbol{y}) = \mathbf{1}^T\boldsymbol{G}\boldsymbol{y}$, $\gamma = \mathbf{1}^T\boldsymbol{G}\mathbf{1}$, and $N = \mathbf{1}^T\mathbf{1}$. With these definitions, we can further expand the above equation as

$$
\begin{aligned}
4\mathcal{L}_{wca} &= \frac{\alpha(\boldsymbol{y}) + \gamma + 2(1-2k)\beta(\boldsymbol{y})}{k(1-k)N} \\
&= \frac{(\alpha(\boldsymbol{y}) + \gamma) + 2(1-2k)\beta(\boldsymbol{y})}{k(1-k)N} - \frac{2(\alpha(\boldsymbol{y}) + \gamma)}{N} + \frac{2\alpha(\boldsymbol{y})}{N} + \frac{2\gamma}{N}.
\end{aligned}
$$

Dropping the last constant term, $\frac{2\gamma}{N}$, leads to

$$
\begin{aligned}
4\mathcal{L}_{wca} &= \frac{(\alpha(\boldsymbol{y}) + \gamma) + 2(1-2k)\beta(\boldsymbol{y})}{k(1-k)N} - \frac{2(\alpha(\boldsymbol{y}) + \gamma)}{N} + \frac{2\alpha(\boldsymbol{y})}{N} \\
&= \frac{(1-2k+2k^2)(\alpha(\boldsymbol{y}) + \gamma) + 2(1-2k)\beta(\boldsymbol{y})}{k(1-k)N} + \frac{2\alpha(\boldsymbol{y})}{N} \\
&= \frac{\frac{(1-2k+2k^2)}{(1-k)^2}(\alpha(\boldsymbol{y}) + \gamma) + \frac{2(1-2k)}{(1-k)^2}\beta(\boldsymbol{y})}{\frac{k}{(1-k)}N} + \frac{2\alpha(\boldsymbol{y})}{N}.
\end{aligned}
$$

Letting $b = \frac{k}{1-k}$, it becomes

$$
4\mathcal{L}_{wca} = \frac{(1+b^2)(\alpha(\boldsymbol{y}) + \gamma) + 2(1-b^2)\beta(\boldsymbol{y})}{bN} + \frac{2b\alpha(\boldsymbol{y})}{bN}.
$$

Since $\frac{2\gamma}{N}$ is a constant, we can subtract the constant term without affecting the solution, i.e.,

$$
\begin{aligned}
4\mathcal{L}_{wca} &= \frac{(1+b^2)(\alpha(\boldsymbol{y})+\gamma)+2(1-b^2)\beta(\boldsymbol{y})}{bN} + \frac{2b\alpha(\boldsymbol{y})}{bN} - \frac{2b\gamma}{bN} \\
&= \frac{(1+b^2)(\boldsymbol{y}^T\boldsymbol{G}\boldsymbol{y}+\mathbf{1}^T\boldsymbol{G}\mathbf{1})+2(1-b^2)\mathbf{1}^T\boldsymbol{G}\boldsymbol{y}}{b\mathbf{1}^T\mathbf{1}} + \frac{2b\boldsymbol{y}^T\boldsymbol{G}\boldsymbol{y}}{b\mathbf{1}^T\mathbf{1}} - \frac{2b\mathbf{1}^T\boldsymbol{G}\mathbf{1}}{b\mathbf{1}^T\mathbf{1}} \\
&= \frac{(\mathbf{1}+\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}+\boldsymbol{y})+b^2(\mathbf{1}-\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}-\boldsymbol{y})-2b(\mathbf{1}-\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}+\boldsymbol{y})}{b\mathbf{1}^T\mathbf{1}} \\
&= \frac{[(\mathbf{1}+\boldsymbol{y})-b(\mathbf{1}-\boldsymbol{y})]^T\boldsymbol{G}[(\mathbf{1}+\boldsymbol{y})-b(\mathbf{1}-\boldsymbol{y})]}{b\mathbf{1}^T\mathbf{1}}.
\end{aligned}
$$

Set $\boldsymbol{t} = (\mathbf{1}+\boldsymbol{y}) - b(\mathbf{1}-\boldsymbol{y})$, then

$$
\begin{aligned}
\boldsymbol{t}^T\mathbf{1} &= \frac{2(\mathbf{1}+\boldsymbol{y})^T\mathbf{1}}{2} - \frac{2b(\mathbf{1}-\boldsymbol{y})^T\mathbf{1}}{2} \\
&= 2N_1 - 2bN_2 \\
&= 0.
\end{aligned}
$$

Note that $b = \frac{k}{1-k} = \frac{N_1}{N_2}$, and

$$
\begin{aligned}
\boldsymbol{t}^T\boldsymbol{t} &= [(\mathbf{1}+\boldsymbol{y})-b(\mathbf{1}-\boldsymbol{y})]^T[(\mathbf{1}+\boldsymbol{y})-b(\mathbf{1}-\boldsymbol{y})] \\
&= (\mathbf{1}+\boldsymbol{y})^T(\mathbf{1}+\boldsymbol{y}) - 2b(\mathbf{1}+\boldsymbol{y})^T(\mathbf{1}-\boldsymbol{y}) + b^2(\mathbf{1}-\boldsymbol{y})^T(\mathbf{1}-\boldsymbol{y}) \\
&= 4n_1 - 2b(\mathbf{1}^T\mathbf{1} - \mathbf{1}^T\boldsymbol{y} + \boldsymbol{y}^T\mathbf{1} - \boldsymbol{y}^T\boldsymbol{y}) + 4b^2N_2 \\
&= 4bN_2 + 4b^2N_2 \\
&= 4b(N_2 + bN_2) \\
&= 4b(N_2 + N_1) \\
&= 4b\mathbf{1}^T\mathbf{1}.
\end{aligned}
$$

Putting everything together, the maximal within-cluster association results in the following optimization:

$$
\max_{\boldsymbol{t}^T\mathbf{1}=0} \frac{\boldsymbol{t}^T\boldsymbol{G}\boldsymbol{t}}{\boldsymbol{t}^T\boldsymbol{t}}.
$$

Therefore, the eigenvector corresponding to the largest eigenvalue can be thought as values of indicator variables.

B Acknowledgment

This work was supported by Korea Ministry of Science and Technology under Brain Science and Engineering Research Program, by KOSEF 2000-2-20500-009-5, and by Brain Korea 21 in POSTECH.

References

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. John Wiely & Sons.

Jain, A. K., Murty, M. N., Flynn, P. J., 1999. Data clustering: A review. ACM Computing Surveys 31 (3), 264–323.

Jenssen, R., Eltoft, T., Principe, J. C., 2004. Infromation theoretic spectral clustering. In: Proc. Int. Joint Conf. Neural Networks. pp. 111–116.

Lee, Y., Choi, S., 2004. Minimum entropy, k-means, spectral clustering. In: Proc. Int. Joint Conf. Neural Networks. Budapest, Hungary, pp. 117–122.

Principe, J. C., Xu, D., Fisher III, J. W., 2000. Information-theoretic learning. In: Haykin, S. (Ed.), Unsupervised Adaptive Filtering: Blind Source Separation. John Wiley & Sons, Inc., pp. 265–319.

Roberts, S. J., Everson, R., Rezek, I., 2000. Maximum certainty data partitioning. Pattern Recognition 33, 833–839.

Roberts, S. J., Holmes, C., Denison, D., 2001. Minimum entropy data partitioning using reversible jump Markov chain Monte Carlo. IEEE Trans. Pattern Analysis and Machine Intelligence 23 (8), 909–914.

Roth, V., Laub, J., Kawanabe, M., Buhmann, J. M., 2003. Optimal cluster preserving embedding of nonmetric proximity data. IEEE Trans. Pattern Analysis and Machine Intelligence 25 (12), 1540–1551.

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

Torkkola, K., Campbell, W. M., 2000. Mutual information in learning feature transformations. In: Proc. Int. Conf. Machine Learning. pp. 1015–1022.

Turlach, B. A., 1993. Bandwidth selection in kernel density estimation: A review. CORE and Institut de Statistique.

Vincent, P., Bengio, Y., 2003. Manifold Parzen windows. In: Advances in Neural Information Processing Systems. Vol. 15. MIT Press, pp. 825–832.