Minimum Entropy, k-Means, Spectral Clustering

Yongjin Lee
Biometrics Technology Research Team
ETRI
161 Kajong-dong, Yusong-gu
Daejon 305-350, Korea
Email: solarone@etri.re.kr

Seungjin Choi
Department of Computer Science
POSTECH
San 31 Hyoja-dong, Nam-gu
Pohang 790-784, Korea
Email: seungjin@postech.ac.kr

***Abstract*— This paper addresses an information-theoretic aspect of k-means and spectral clustering. First, we revisit the k-means clustering and show that its objective function is approximately derived from the minimum entropy principle when the Renyi's quadratic entropy is used. Then we present a maximum within-clustering association that is derived using a quadratic distance measure in the framework of minimum entropy principle, which is very similar to a class of spectral clustering algorithms that is based on the eigen-decomposition method.**

I. 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 the 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. For example, k-means clustering can be interpreted as a method for minimizing the sum of pair-wise intra-cluster distances [4]. This implies that k-means clustering uses the Euclidean distance as a dissimilarity measure. This also leads to the fact that k-means assumes Gaussian distribution for data, hence, exploits only second order statistics. It means that we cannot extract all information available from the given data when the probability density of the data is not Gaussian. This might be a restrictive assumption.

An information-theoretic method is an attractive and powerful approach but it involves probability density estimation, which is cumbersome in the viewpoint of computational complexity. Density estimation is categorized into three different classes: parametric, semi-parametric, and non-parametric methods. 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 [1] that Renyi's quadratic entropy and quadratic distance measures between densities simplified entropy or divergence calculation in the framework of the Parzen window-based density estimation. Along this line, a variety of unsupervised learning algorithms [1] and a feature transformation method [6] were developed.

In this paper, we address an information-theoretic clustering which mainly exploits the minimum entropy principle studied in [2], [3] where the clustering is carried out by minimizing the overlap between densities of clusters. In minimum entropy data partitioning [2], [3], the 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 to [2], [3], we employ the Renyi's quadratic entropy and the quadratic distance measure [1] with the Parzen 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 k-means. We also show that minimizing the overlap between cluster densities with a quadratic distance measure leads to the maximization of *within-cluster association*. This maximum within-cluster association is closely related with a spectral clustering which is an eigen-decomposition-based method. This gives some link between information-theoretic clustering and spectral clustering.

II. Minimum Entropy Data Partitioning

We begin by briefly reviewing the method of minimum entropy (or maximum certainty) data partitioning [2], [3] since this idea is a starting point for our method. In 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

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

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

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

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 [2], [3].

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

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

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

III. 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.

A. Renyi's Quadratic Entropy

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}, which 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 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 [1], [6].

B. An Objective Function

The k-means clustering partitions the data in such a way that the sum of intra-cluster distances to the cluster mean vector 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 [4] 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., $z_{in} = p\left(i|\boldsymbol{x}_n\right)$. 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

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

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).

IV. Minimum Entropy And Spectral Clustering

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

A. Quadratic Distance Measure

Principe *et al.* proposed a quadratic distance measure between probability densities which resembles the Euclidean distance between two vectors [1]. A main motivation of the quadratic distance measure lies in its simple form for the case where the Parzen window-based density estimation with Gaussian kernel is involved.

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

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

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

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

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

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

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

where

$$f(\boldsymbol{x}) = \frac{1}{N_f}\sum_{n=1}^{N_f} G\left(\boldsymbol{x};\boldsymbol{x}_n^f,\sigma^2\right), \quad (22)$$

$$g(\boldsymbol{x}) = \frac{1}{N_g}\sum_{n=1}^{N_g} G\left(\boldsymbol{x};\boldsymbol{x}_n^g,\sigma^2\right), \quad (23)$$

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.

B. 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 [2], [3]. Now we compute this overlap using the quadratic distance between densities estimated by the Parzen window method.

Like the previous section, ith class conditional density is written as

$$p(\boldsymbol{x}|i) = \frac{1}{N_i}\sum_{n=1}^{N_i} z_{in} G(\boldsymbol{x};\boldsymbol{x}_n,\sigma^2). \quad (24)$$

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

$$\begin{aligned}
\mathcal{V}_i &= -D\left[p(\boldsymbol{x}|i)||p(\boldsymbol{x})\right] \\
&= -\int p^2(\boldsymbol{x}|i)\,d\boldsymbol{x} - \int p^2(\boldsymbol{x})\,d\boldsymbol{x} \\
&\quad +2\int p(\boldsymbol{x}|i)\,p(\boldsymbol{x})\,d\boldsymbol{x} \\
&= -\frac{1}{N_i^2}\sum_{n=1}^{N}\sum_{m=1}^{N} z_{in}z_{im} G\left(\boldsymbol{x}_n;\boldsymbol{x}_m,2\sigma^2\right) \\
&\quad -\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} G\left(\boldsymbol{x}_n;\boldsymbol{x}_m,2\sigma^2\right) \\
&\quad +\frac{2}{NN_i}\sum_{n=1}^{N}\sum_{m=1}^{N} z_{in} G\left(\boldsymbol{x}_n;\boldsymbol{x}_m,2\sigma^2\right) \\
&= -\frac{1}{N_i^2}\boldsymbol{z}_i^T\boldsymbol{G}\boldsymbol{z}_i - \frac{1}{N^2}\mathbf{1}^T\boldsymbol{G}\mathbf{1} + \frac{2}{NN_i}\boldsymbol{z}_i^T\boldsymbol{G}\mathbf{1}, \quad (25)
\end{aligned}$$

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

Then the total overlap can be written as

$$\begin{aligned}
\mathcal{V} &= \sum_{i=1}^{K} p(i)\,\mathcal{V}_i \\
&= -\sum_{i=1}^{K} p(i)\,D\left[p(\boldsymbol{x}|i)\,\|p(\boldsymbol{x})\right] \\
&= \frac{1}{N}\left[\frac{\mathbf{1}^T\boldsymbol{G}\mathbf{1}}{N} - \sum_{i=1}^{K}\frac{\boldsymbol{z}_i^T\boldsymbol{G}\boldsymbol{z}_i}{N_i}\right], \quad (26)
\end{aligned}$$

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

This is reminiscent of Eq. (5). The first term in Eq. (26) 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

$$\begin{aligned}
\mathcal{L}_{wca} &= \sum_{i=1}^{K}\frac{1}{N_i}\sum_{n=1}^{N}\sum_{m=1}^{N} z_{in}z_{im} G\left(\boldsymbol{x}_n;\boldsymbol{x}_m,2\sigma^2\right) \\
&= \sum_{i=1}^{K}\frac{\boldsymbol{z}_i^T\boldsymbol{G}\boldsymbol{z}_i}{N_i}. \quad (27)
\end{aligned}$$

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. (27) 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).

C. 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, the maximization of (27) leads to one of well-known spectral clustering criterion, *average association* in [5]. In other words, in such a case, the indicator variables in Eq. (27) 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

$$p(\boldsymbol{x}|1) = \frac{1}{N_1}\sum_{n=1}^{N} z_n G\left(\boldsymbol{x};\boldsymbol{x}_n,\sigma^2\right), \quad (28)$$

$$p(\boldsymbol{x}|2) = \frac{1}{N_2}\sum_{n=1}^{N} (1-z_n) G\left(\boldsymbol{x};\boldsymbol{x}_n,\sigma^2\right), \quad (29)$$

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. (27) can be written as

$$\mathcal{L}_{wca} = \frac{\boldsymbol{z}^T\boldsymbol{G}\boldsymbol{z}}{\boldsymbol{z}^T\boldsymbol{z}} + \frac{(\mathbf{1}-\boldsymbol{z})^T\boldsymbol{G}(\mathbf{1}-\boldsymbol{z})}{(\mathbf{1}-\boldsymbol{z})^T(\mathbf{1}-\boldsymbol{z})}, \quad (30)$$

where $\mathbf{1}\in\mathbb{R}^N$ by $[\mathbf{1}]_n = 1$. 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. (30) can be rewritten as

$$4\mathcal{L}_{wca} = \frac{(\mathbf{1}+\boldsymbol{y})^T \boldsymbol{G} (\mathbf{1}+\boldsymbol{y})}{\frac{1}{4}(\mathbf{1}+\boldsymbol{y})^T (\mathbf{1}+\boldsymbol{y})} + \frac{(\mathbf{1}-\boldsymbol{y})^T \boldsymbol{G} (\mathbf{1}-\boldsymbol{y})}{\frac{1}{4}(\mathbf{1}-\boldsymbol{y})^T (\mathbf{1}-\boldsymbol{y})}. \quad (31)$$

Adopting a similar method that used in [5], the maximum within-cluster association reduces to the following simple optimization problem:

$$\max_{\boldsymbol{t}^T\mathbf{1}=0} \frac{\boldsymbol{t}^T \boldsymbol{G} \boldsymbol{t}}{\boldsymbol{t}^T \boldsymbol{t}}, \quad (32)$$

where $\boldsymbol{t} = (\mathbf{1}+\boldsymbol{y}) - \frac{N_1}{N_2}(\mathbf{1}-\boldsymbol{y})$ 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.

V. A Numerical Experiment

We present a simple numerical example of a set of ring data which consists of two clusters (inner ring and outer ring). Data samples were 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 $N = 400$. The width of Gaussian kernel in the Parzen density estimation was set as $\sigma^2 = 0.2$. 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. 1, clearly exhibits discrimination. The average value over the elements in the largest eigenvector of $\boldsymbol{G}$ is used as a threshold value. Successful clustering result in shown in Fig. 2. Two clusters share the same mean vector. Hence, k-means clustering method fails to correctly partition the data (see Fig. 2) because it is based on the Euclidean distance between the data point and the mean vector.

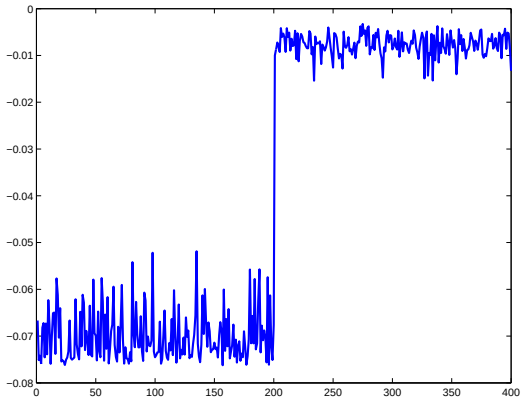


Fig. 1. The largest eigenvector of the kernel matrix $\boldsymbol{G}$ indicates the clear discrimination between two clusters.

VI. Discussion

We started from the idea of minimum entropy data partitioning [2], [3] where the goal of clustering is viewed as the minimization of overlap between cluster densities. The KL divergence in the overlap measure was used in [2], [3], 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-based density estimation, which was introduced in [1]. Adopting the Renyi's quadratic entropy leaded to the objective function of K-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 with one of well-known spectral clustering methods which are based on the eigen-decomposition. In fact this might give an information-theoretic insight to spectral clustering.

Appendix

Let $k = \frac{N_1}{N}$, then Eq. (31) becomes

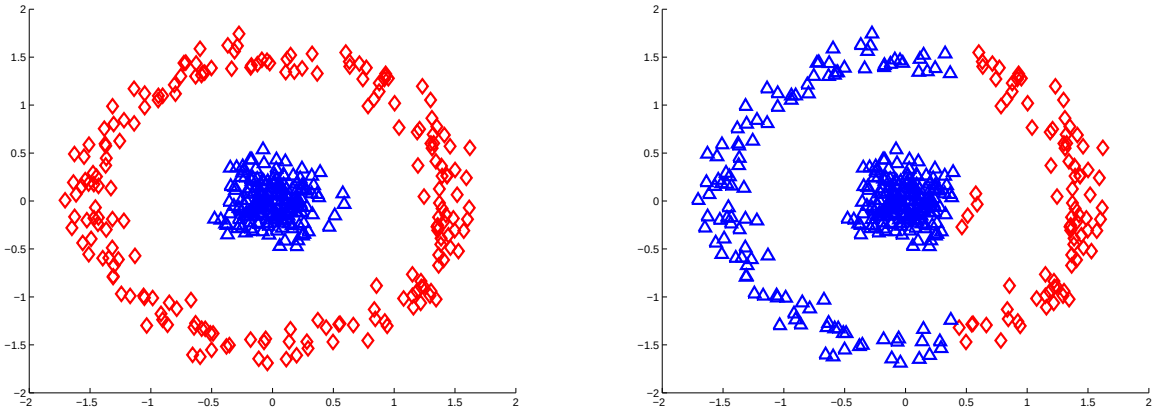
$$\begin{aligned}
4\mathcal{L}_{wca} &= \frac{(\mathbf{1}+\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}+\boldsymbol{y})}{k\mathbf{1}^T\mathbf{1}} + \frac{(\mathbf{1}-\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}-\boldsymbol{y})}{(1-k)\mathbf{1}^T\mathbf{1}} \\
&= \frac{(1-k)(\mathbf{1}+\boldsymbol{y})^T\boldsymbol{G}(\mathbf{1}+\boldsymbol{y})}{k(1-k)\mathbf{1}^T\math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

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

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

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

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



(a) (b)

Fig. 2. A simple example of clustering a ring data: (a) by our minimum entropy spectral clustering; (b) by the standard K-means.

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.,

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

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

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

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.

ACKNOWLEDGMENT

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

REFERENCES

[1] J. C. Principe, D. Xu, and J. W. Fisher III, "Information-theoretic learning," in *Unsupervised Adaptive Filtering: Blind Source Separation*, S. Haykin, Ed. John Wiley & Sons, Inc., 2000.

[2] S. J. Roberts, R. Everson, and I. Rezek, "Maximum certainty data partitioning," *Pattern Recognition*, vol. 33, pp. 833–839, 2000.

[3] S. J. Roberts, C. Holmes, and D. Denison, "Minimum entropy data partitioning using reversible jump Markov chain Monte Carlo," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 909–914, 2001.

[4] V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann, "Optimal cluster preserving embedding of non-metric proximity data," University of Bonn, Tech. Rep. IAI-TR-2002-5, 2002.

[5] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 731–737.

[6] K. Torkkola and W. M. Campbell, "Mutual information in learning feature transformations," in *Proc. Int. Conf. Machine Learning*, 2000, pp. 1015–1022.