JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 34, 289-305 (2018)
DOI: 10.6688/JISE.2018.34.1.17

Locality Sensitive *K*-means Clustering

CHIEN-LIANG LIU[1], WEN-HOAR HSAIO[2] AND TAO-HSING CHANG[3]
[1]*Department of Industrial Engineering and Management*
[2]*Department of Computer Science*
National Chiao Tung University
Hsinchu, 300 Taiwan
[3]*Department of Computer Science and Information Engineering*
National Kaohsiung University of Applied Sciences
Kaohsiung, 807 Taiwan

This study considers clustering and dimensionality reduction simultaneously to devise an unsupervised clustering algorithm called locality sensitive *K*-means (LS-K-means). The goal is to find a linear transformation to project data points into a lower dimensional space, so that clustering can perform well in the new space. We design a novel objective function for LS-Kmeans to achieve the goal, and further show that the proposed method can be reformulated as a matrix trace minimization with constraints problem. The original optimization problem becomes a generalized eigenvalue problem when relaxing the optimization problem of LS-Kmeans by allowing the indicator entries to take arbitrary values in $\mathbb{R}$. This paper also shows that the continuous solutions for the transformed cluster membership indicator vectors of LS-Kmeans are located in the subspace spanned by the first $K-1$ eigenvectors. In the experiments, we use two synthetic datasets to show that the proposed method can cluster non-linearly separable data points. Besides, the experimental results of eight real datasets indicate that the proposed algorithm can generally outperform other alternatives.

Keywords: unsupervised learning, clustering, locality sensitive clustering, dimensionality reduction, linear transformation

1. INTRODUCTION

Clustering is a fundamental data mining process that has been extensively studied across varied disciplines over several decades. The goal of clustering is to identify latent information in the underlying data, so that objects from the same cluster are more similar to each other than objects from different clusters. With an increasing number of applications that deal with very large high dimensional datasets, clustering has emerged as a very important research area in many disciplines [21-24, 34, 36, 37], including, but not limited to, computer vision, document analysis and Bioinformatics. For example, images usually contain billions pixels with color information, and text documents are associated with hundreds of thousands of vocabularies [13]. Studies about DNA microarray technology in Bioinformatics typically produce large-scale data that contain measures on thousands of genes under hundreds of conditions [7].

Various clustering algorithms have been devised, including *K*-means, Fuzzy *c*-means (FCM) [4], hierarchical clustering, and spectral clustering [26, 31]. The *K*-means and FCM are two of the most popular and efficient clustering algorithms, aiming at the minimization of the average squared distance between the objects and the cluster centers. The *K*-means algorithm assigns each data point to a single cluster; while FCM, an extension of *K*-means, allows each data point to be a member of multiple clusters with a

Received April 8, 2016; revised July 22, 2016; accepted August 3, 2016.
Communicated by Tzung-Pei Hong.

membership value. The flexibility of FCM yields better result for overlapped dataset and comparatively better than *K*-means. The algorithms for *K*-means and FCM are similar, since they both use iterative refinement technique until convergences. Given initial prototypes or centers of the clusters, the algorithm proceeds by alternating between two steps: (1) update membership values u_{ij}, which denotes the degree of data point $\mathbf{x}_i$ belonging to cluster c_j; (2) compute the new centroid or prototype for each cluster.

The *K*-means and FCM generally use Euclidean distance as the distance metric, explaining why they can have good performances on the datasets with compact supersphere distributions, but tend to fail in the data organized in more complex and unknown shapes [35]. In many applications such as document classification and pattern recognition, each object generally comprises thousands of features. One of the problems with high-dimensional datasets is that not all the measured variables are important for understanding the underlying phenomena of interest. As a result, *K*-means and FCM fail to generally perform well on high-dimensional datasets. One approach to simplification is to assume that the data of interest lies on an embedded linear subspace or non-linear manifold within the higher-dimensional space. The above assumption leads one to consider dimensionality reduction that allows one to represent the data in a lower dimensional space.

This study devises an unsupervised clustering algorithm called locality sensitive *K*-means (LS-Kmeans), which considers clustering criterion and retains local geometrical structure in projecting the data points to a lower dimensional space. Compared to previous work, this study considers clustering and dimensionality reduction simultaneously. To retain local geometrical structure of the data, this study uses graph to represent all data points, and encodes neighborhood information of the data nodes by using a Gaussian weighting function to represent the similarity between data points. We formalize the proposed algorithm as finding a liner transformation, which considers clustering criterion and preserves local neighborhood information, such that the clustering can perform well in the new space.

The main contributions of this study include: (1) this study devises an unsupervised clustering algorithm called LS-Kmeans; (2) this study further shows that the objective function of LS-Kmeans can be reformulated as a matrix trace minimization with constraints problem. The result leads the optimization of the proposed objective function to become a generalized eigenvalue problem; (3) this study shows that the continuous solutions for the transformed cluster membership indicator vectors of LS-Kmeans are located in the subspace spanned by the first $K-1$ eigenvector; (4) this study uses two synthetic datasets and eight real datasets to conduct experiments. The experimental results of synthetic datasets show that the proposed algorithm can separate non-linearly separable clusters. The experimental results of real datasets indicate that the proposed algorithm can generally outperform other alternatives.

The rest of this study is organized as follows. In Section 2, related surveys are presented. In Section 3, the locality sensitive *K*-means algorithm is introduced. In Section 4, several experiments are introduced. In Section 6, the conclusion is presented.

2. RELATED WORK

High-dimensional datasets present many mathematical challenges to machine learn-

ing tasks. First, the curse of dimensionality problem may arise when dealing with high-dimensional datasets. The volume of the space increases and the available data becomes sparse when the dimensionality increases. Therefore, enormous data examples are required for machine learning algorithms to learn models. Second, the concept of distance becomes less precise as the number of dimensions grows, since the distance between any two points in a given dataset converges [19]. Third, not all the measured features are important for understanding the underlying phenomena of interest, so some irrelevant features may affect machine learning performance. While certain computationally expensive novel methods [5] can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modelling of the data.

Many dimensionality reduction algorithms have been developed to accomplish these tasks. Principal component analysis (PCA), linear discriminant analysis (LDA) and multidimensional scaling (MDS) are methods that provide a sequence of best linear approximations to a given high-dimensional observation. In order to resolve the problem of dimensionality reduction in nonlinear cases, many recent techniques have been devised in the last decade, including Isomap [32], locally linear embedding (LLE) [30], Laplacian eigenmaps [3], and locality preserving projections (LPP) [16]. These methods have been shown to be effective in discovering the geometrical structure of the underlying manifold. Among these methods, LPP possesses several useful properties [16]. First, LPP is linear, making it fast and suitable for practical applications. Second, LPP focuses on preserving locality information, making it to be of particular use in several domains, including dimensionality reduction [41], text retrieval [6], brain-computer interface [38], speech recognition [33], multimedia retrieval [15] and pattern recognition [17, 40]. Third, the linear transformation obtained from available training data can be applied to any new data point to locate it in the reduced representation space.

Practically, the purposes of cluster analysis and dimensionality reduction are different. Cluster analysis assigns the data points into clusters so that the data points in the same cluster are more similar to each other than to those in other clusters; while dimensionality reduction techniques try to find a lower dimensional representation of the data according to some criterion. However, unsupervised dimensionality reduction is closely related to unsupervised clustering. Ding and He [12] showed that principal components of PCA are continuous (relaxed) solution of the cluster membership indicators in *K*-means clustering. Honda *et al.* [18] further proposed a robust clustering algorithm by using a noise-rejection mechanism based on the noise-clustering approach. The responsibility weight of each sample for the *K*-means process is estimated by considering the noise degree of the sample, and cluster indicators are calculated in a fuzzy principal component analysis (PCA) guided manner, where fuzzy PCA-guided robust *K*-means is performed by considering responsibility weights of samples. Additionally, Dhillon and Modha [11] devised a spherical *K*-means clustering algorithm to perform concept decomposition, and their finding empirically showed that the approximation errors of the concept decompositions are close to truncated singular value decompositions [14] (SVD), which is a popular and well studied matrix approximation scheme. The linear algebra SVD operation is the key component of latent semantic indexing (LSI) [9], which reduces dimensionality of document-term matrix to further discover latent relationships between correlated terms and documents. Recently, Kumar and Srinivas [20] further show-

ed that concept decomposition based on FCM clustering provides better approximation than that based on spherical *K*-means clustering.

The previous work closely related to our proposed clustering model is *p*-Kmeans [39], which formalizes the *K*-means clustering problem as a matrix trace maximization problem. However, several differences exist between the two approaches. First, *p*-Kmeans only focuses on clustering criterion; while LS-Kmeans considers clustering and dimensionality reduction simultaneously. Second, *p*-Kmeans becomes a matrix trace maximization problem after the derivation; while LS-Kmeans is a matrix trace minimization problem. Third, *p*-Kmeans only considers clustering criterion in the optimization problem; while LS-Kmeans uses locality preserving and clustering as the criteria in the optimization problem. This study further conducts experiments to compare *p*-Kmeans with the proposed algorithm. Another research related to the proposed method is the approach called integrated KL clustering (IKL) [1], combining *K*-means clustering on data attributes and normalized cut spectral clustering on pairwise relations. Wang *et al.* [1] related IKL with linear discriminant analysis (LDA) to relax and formalize IKL as an optimization problem. The relaxed problem involves the computation of pseudo inverse of normalized Laplacian matrix, and it is generally a computational intensive task. We compare IKL with LS-Kmeans in the experiments.

3. LOCALITY SENSITIVE *K*-MEANS

3.1 Notation

The notations that are used in the following sections are described here. Given a set of data points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}$ in $\mathbb{R}^n$, the goal is to partition the data points into K clusters, $S_1, \ldots, S_k, \ldots, S_K$, each of which comprises $N_k (1 \le k \le K)$ data points. This study uses a matrix $\mathbf{X}$ to denote all data points as shown in Eq. (1). We define a diagonal matrix $\mathbf{N}$ containing the information for each cluster size as shown in Eq. (2), in which the diagonal entries are $1/N_1, 1/N_2, \ldots, 1/N_K$. This study uses $\boldsymbol{\mu}_k \in \mathbb{R}^n$ $(1 \le k \le K)$ to denote the mean of the kth cluster as shown in Eq. (3). The clustering task can be formalized as finding an assignment matrix $\mathbf{S}$ with dimension $m \times K$, as well as a set of vectors $\{\boldsymbol{\mu}_k\}$, such that a specific clustering criterion is achieved.

In matrix representation, the trace of a matrix $\mathbf{A}$ is denoted as $\mathbf{tr(A)}$, and I_K represents a $K \times K$ identity matrix. The Frobenius norm of matrix $\mathbf{A}$ is represented as $\|\mathbf{A}\|_F$, which is equal to $\sqrt{\mathbf{tr}(\mathbf{A}\mathbf{A}^T)}$. This study uses $\mathbf{e}$ to denote a vector with all entries one. Besides data matrix, we introduce a mean matrix $\mathbf{C}$ as shown in Eq. (4), in which $\mathbf{c}^{(i)}$ represents the cluster center of $\mathbf{x}^{(i)}$'s cluster. For instance, if $\mathbf{x}^{(i)}$ belongs to cluster k, then $\mathbf{c}^{(i)}$ is $\boldsymbol{\mu}_k$.

$$\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}) \in \mathbb{R}^{n \times m} \tag{1}$$

$$\boldsymbol{N} = \begin{pmatrix} 1/N_1 & 0 & \ldots & 0 \\ 0 & 1/N_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/N_K \end{pmatrix} \tag{2}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{\mathbf{x}^{(i)} \in S_k} \mathbf{x}^{(i)} \tag{3}$$

$$\mathbf{C} = (\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(m)}) \in \mathbb{R}^{n \times m} \tag{4}$$

3.2 Matrix Form of *K*-means Clustering

The *K*-means is a typical clustering algorithm, aiming at the minimization of the average squared distance between the data points and the cluster centers. We start the derivation from *K*-means objective function as shown in Eq. (5). It can be further represented as a matrix form as shown in Eq. (6) in terms of the mean matrix listed in Eq. (4).

Without loss of generality, we assume that the data points within the same cluster are arranged together. Then, the binary indicator matrix can be represented as the form listed in Eq. (7), indicating that $\mathbf{x}^{(1)}$, …, $\mathbf{x}^{(N_1)}$ belong to the first cluster, and so on. Then, we can use a matrix form to represent the mean matrix in terms of binary indicator matrix. First, according to the definition and linear algebra operations, the mean matrix can be decomposed into the multiplication of two matrices as shown in Eq. (8). Next, the first matrix in Eq. (8) can be further decomposed into the multiplication of matrix $\mathbf{X}$ and matrix $\mathbf{S}$; while the second matrix can be decomposed into the multiplication of matrix $\mathbf{N}$ and matrix $\mathbf{S}^T$. Eq. (9) presents another matrix representation of mean matrix.

$$J = \sum_{k=1}^{K} \sum_{x_i \in S_k} \| \mathbf{x}_i - \boldsymbol{\mu}_k \|^* \tag{5}$$

$$= \| \mathbf{X} - \mathbf{C} \|_F^2 . \tag{6}$$

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{7}$$

$$\mathbf{C} = \left(\sum_{\mathbf{x}^{(i)} \in S_1} \mathbf{x}^{(i)}, \ldots, \sum_{\mathbf{x}^{(i)} \in S_K} \mathbf{x}^{(i)} \right) \cdot \begin{pmatrix} 1/N_1 & \cdots & 1/N_1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/N_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 1/N_K \end{pmatrix} \tag{8}$$

$$= \mathbf{XSNS}^T \tag{9}$$

The *K*-means generally uses Euclidean distance as the distance metric, explaining why it can have a good performance on the data with compact super-sphere distributions, but tends to fail in the data organized in more complex and unknown shapes [35]. However, the analysis on high-dimensional datasets becomes a topic of significant recent interest due to the advances in data collection and storage capabilities during the past decades. This study proposes to use dimensionality reduction technique and consider clustering criterion to improve clustering performance.

3.3 Dimensionality Reduction with Locality Preserving

Given a set of data points, the goal of dimensionality reduction is to find a lower dimensional representation of the data points according to some criterion. Let $\mathbf{Z} = (\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)})$ be such a map, which projects a data point to a lower dimensional space according to different criteria. For instance, the goal of PCA is to perform dimensionality reduction while preserving as much of the variance in the high-dimensional space as possible. This study uses the geometric structures as a criterion to reduce dimensionality such that the distance relationship between data points is retained during the course of dimensionality reduction. The criterion of dimensionality reduction used in this paper can be represented as $J(a)$ presented in Eq. (10), where $\mathbf{W}$ is a similarity matrix between data points. Then, we use a similarity graph to denote the relationship between the data points. The graph is constructed with *K*-nearest neighbor scheme and Gaussian weighting function listed in Eq. (11), where σ is a constant controlling width of the graph. It is apparent that the weight between two points is between 0 and 1.

$$J(\mathbf{a}) = \sum_{i,j} (\mathbf{z}^{(i)} - \mathbf{z}^{(j)})^2 \mathbf{W}_{ij}, \text{ where } \mathbf{a}^T \mathbf{x}^{(i)} \text{ and } \mathbf{z}^{(j)} = \mathbf{a}^T \mathbf{x}^{(j)} \tag{10}$$

$$\mathbf{W}_{ij} = \exp\left(-\frac{\| \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \|^2}{2\sigma^2} \right), \text{where } 1 \le i, j \le m \tag{11}$$

Then, the minimization of Eq. (10) is a reasonable criterion for choosing the mapping [3, 16], in which a is a projection vector, and $\mathbf{W}_{ij}$ denotes the similarity between neighboring points $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$. The Eq. (10) attempts to ensure that if points $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ are connected with high weight, their correspondent mapping points $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$ are close in the projected space. Additionally, the Eq. (10) can be further represented as a matrix form as shown in Eq. (12), in which $\mathbf{D}$ is a diagonal matrix with diagonal entries $D_{ii} = \sum_{j=1} \mathbf{W}_{ij}$, and $\mathbf{L}$ is called graph Laplacian [8].

$$\begin{aligned} J(\mathbf{a}) &= \sum_{i,j} (\mathbf{a}^T \mathbf{x}^{(i)} - \mathbf{a}^T \mathbf{x}^{(j)})^2 \mathbf{W}_{ij} \\ &= \sum_{i,j} (\mathbf{a}^T \mathbf{x}^{(i)} - \mathbf{a}^T \mathbf{x}^{(j)})(\mathbf{a}^T \mathbf{x}^{(i)} - \mathbf{a}^T \mathbf{x}^{(j)})^T \mathbf{W}_{ij} \\ &= \mathbf{a}^T \left[\sum_{i,j} (\mathbf{x}^{(i)} - \mathbf{x}^{(j)})(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})^T \mathbf{W}_{ij} \right] \mathbf{a} \\ &= 2\mathbf{a}^T \left[\sum_{i,j} (\mathbf{x}^{(i)} \mathbf{x}^{(i)^T} \mathbf{W}_{ij} - \sum_{i,j} (\mathbf{x}^{(i)} \mathbf{x}^{(j)^T} \mathbf{W}_{ij} \right] \mathbf{a} \end{aligned}$$

$$
\begin{aligned}
&= 2\mathbf{a}^T \left[\sum_i (\mathbf{x}^{(i)}\mathbf{x}^{(i)^T} \sum_j \mathbf{W}_{ij} - \sum_{i,j} (\mathbf{x}^{(i)}\mathbf{x}^{(j)^T} \mathbf{W}_{ij} \right] \mathbf{a} \\
&= 2\mathbf{a}^T \left[\sum_i (\mathbf{x}^{(i)}\mathbf{x}^{(i)^T} \mathbf{D}_{ij} - \sum_{i,j} (\mathbf{x}^{(i)}\mathbf{x}^{(j)^T} \mathbf{W}_{ij} \right] \mathbf{a} \\
&= 2\mathbf{a}^T \left[\mathbf{X}\mathbf{D}\mathbf{X}^T - \mathbf{X}\mathbf{W}\mathbf{X}^T \right] \mathbf{a} \\
&= 2\mathbf{a}^T \mathbf{X}\mathbf{L}\mathbf{X}^T \mathbf{a}, \text{ where } \mathbf{L} = \mathbf{D} - \mathbf{W}
\end{aligned} \tag{12}
$$

The minimization of Eq. (12) with appropriate constraints can be transformed into a generalized eigenvalue problem, in which projection vectors $\mathbf{a}^{(i)}$ (i = 1, . . . , K) are the correspondent eigenvectors for the minimization. The projection vectors are collected as a matrix $\mathbf{A}$ as shown in Eq. (13).

$$\mathbf{A} = (\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(K)}) \in \mathbb{R}^{n \times m} \tag{13}$$

3.4 Locality Sensitive *K*-means Algorithm

This study considers clustering and dimensionality reduction simultaneously to devise a novel unsupervised clustering algorithm, called locality sensitive *K*-means (LS-Kmeans), to cluster data points in the reduced feature space such that clustering performance can be improved. Eq. (14) shows the objective function, in which λ is a constant controlling the weight of the regularization term. The most prominent property of the proposed approach is the complete preservation of both clustering and local geometrical structure in the data. The other methods, such as LDA, can only preserve the global discriminant structure, while the local geometrical structure is ignored.

$$\min_a \|\mathbf{X} - \mathbf{C}\|_F^2 + \lambda \mathbf{tr}(\mathbf{A}^T \mathbf{X}\mathbf{L}\mathbf{X}^T \mathbf{A}) \tag{14}$$

This study considers clustering and dimensionality reduction simultaneously, inspiring us to project the data points within the same cluster to the same point on the new feature space. Thus, the projection of the data points, $\mathbf{A}^T\mathbf{X}$, can be approximated by $\mathbf{N}^{1/2}\mathbf{S}^T$. Using the above mapping mechanism, the points in the first cluster are mapped to $(\frac{1}{\sqrt{N_1}}, 0, ..., 0)^T$; while the points in the second cluster are mapped to $(0, \frac{1}{\sqrt{N_1}}, ..., 0)^T$, and so on.

It is apparent that the above mapping is an optimal reduction from cluster's point of view, since the points within the same cluster are grouped together, and the points from different clusters are far apart. Using the definition of Frobenius norm and the matrix form of mean matrix as shown in Eq. (9), Eq. (14) can be further transformed into matrix trace representation form as shown in Eq. (15), where we introduce a matrix $\mathbf{H}^T$ to denote $\mathbf{N}^{1/2}\mathbf{S}^T$ and $\mathbf{A}^T\mathbf{X}$ simultaneously.

$$
\begin{aligned}
&\|\mathbf{X} - \mathbf{C}\|_F^1 + \lambda \mathbf{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}) \\
&= \mathbf{tr}((\mathbf{X} - \mathbf{C})(\mathbf{X} - \mathbf{C})^T) + \lambda \mathbf{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}) \\
&= \mathbf{tr}((\mathbf{X}\mathbf{X} - \mathbf{X}\mathbf{S}\mathbf{N}\mathbf{S}^T)(\mathbf{X} - \mathbf{X}\mathbf{S}\mathbf{N}\mathbf{S}^T)^T) + \lambda \mathbf{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}) \\
&= \mathbf{tr}(\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}\mathbf{N}\mathbf{S}^T\mathbf{X}^T) + \lambda \mathbf{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}) \\
&= \mathbf{tr}(\mathbf{X}\mathbf{X}^T) - \mathbf{tr}(\mathbf{N}^{1/2}\mathbf{S}^T\mathbf{X}^T - \mathbf{X}\mathbf{S}\mathbf{N}^T)^T) + \lambda \mathbf{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A})
\end{aligned}
$$

$$= \mathbf{tr}(\mathbf{X}\mathbf{X}^T) - \mathbf{tr}(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}) + \lambda\mathbf{tr}(\mathbf{H}^T\mathbf{L}\mathbf{H}) \tag{15}$$

The entries of discrete indicator matrix $\mathbf{H}$ have only one sign, but a continuous solution with positive and negative entries will be much closer to its discrete form [12]. This work uses the scheme proposed by Ding and He [12] to estimate continuous solutions to the discrete cluster membership indicators for clustering. We perform a linear transformation on $\mathbf{H}$ to produce an $m \times K$ matrix $\mathbf{Q}_K$ as shown in Eq. (16):

$$\boldsymbol{Q}_K = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K) = \mathbf{HT} \tag{16}$$

where $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_K)$ is a $K \times K$ orthonormal matrix, and the last column of $\mathbf{T}$ is

$$\mathbf{t}_K + \left(\sqrt{\frac{N_1}{N}}, \sqrt{\frac{N_2}{N}}, \ldots, \sqrt{\frac{N_K}{N}}\right)^T. \tag{17}$$

The above transformation gives rise to two properties of $\mathbf{Q}_K$. First, $\mathbf{Q}_K$ is an $m \times K$ orthonormal matrix. Second, $\mathbf{q}_K$, the last column of $\mathbf{Q}_K$, is equal to $\sqrt{\frac{1}{N}}\mathbf{e}$. Then, the H matrix in objective function of LS-Kmeans can be replaced by $\mathbf{Q}_K\mathbf{T}^T$. With some algebra operations, the objective function can be represented as the form listed in Eq. (18), in which $\mathbf{tr}(\mathbf{X}\mathbf{X}^T)$, $\mathbf{e}^T\mathbf{X}^T\mathbf{X}\mathbf{e}/N$ and $\mathbf{e}^T\mathbf{L}\mathbf{e}/N$ are constants.

$$\begin{aligned} J &= \mathbf{tr}(\mathbf{X}\mathbf{X}^T) - \mathbf{tr}(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}) + \lambda\mathbf{tr}(\mathbf{H}^T\mathbf{L}\mathbf{H}) \\ &= \mathbf{tr}(\mathbf{X}\mathbf{X}^T) - \mathbf{tr}(\mathbf{T}\mathbf{Q}_K^T\mathbf{X}^T\mathbf{X}\mathbf{Q}_K\mathbf{T}^T) + \lambda\mathbf{tr}(\mathbf{T}\mathbf{Q}_K^T\mathbf{L}\mathbf{Q}_K\mathbf{T}^T) \\ &= \mathbf{tr}(\mathbf{X}\mathbf{X}^T) - \mathbf{tr}(\mathbf{q}_K^T\mathbf{X}^T\mathbf{X}\mathbf{q}_K) - \mathbf{tr}(\mathbf{Q}_{K\text{-}1}^T\mathbf{X}^T\mathbf{X}\mathbf{Q}_{K\text{-}1}) + \lambda(\mathbf{tr}(\mathbf{q}_K^T\mathbf{L}\mathbf{q}_K) + \mathbf{tr}(\mathbf{Q}_{K\text{-}1}^T\mathbf{L}\mathbf{Q}_{K\text{-}1})) \\ &= \mathbf{tr}(\mathbf{X}\mathbf{X}^T) - \mathbf{e}^T\mathbf{X}^T\mathbf{X}\mathbf{e}/N + \mathbf{tr}(\mathbf{Q}_{K\text{-}1}^T\mathbf{X}^T\mathbf{X}\mathbf{Q}_{K\text{-}1}) + \lambda(\mathbf{e}^T\mathbf{L}\mathbf{e}/N + \mathbf{tr}(\mathbf{Q}_{K\text{-}1}^T\mathbf{L}\mathbf{Q}_{K\text{-}1})) \end{aligned} \tag{18}$$

Therefore, the optimization can be further represented as the minimization of Eq. (19). Besides the matrix trace minimization, $\mathbf{Q}_{K\text{-}1}^T\mathbf{L}\mathbf{Q}_{K\text{-}1} = \mathbf{I}_{K\text{-}1}$ is a constraint of the optimization. Eq. (19) becomes an optimization with constraint problem. However, it is a discrete optimization problem, so it is a NP hard problem. This study relaxes the problem to allow the solution to take arbitrary values in $\mathbb{R}$. The optimization problem becomes a generalized eigenvalue problem. Algorithm 1 shows LS-Kmeans algorithm. First, we use Gaussian similarity function to construct weighted adjacency matrix as shown in Line 2. Then, Line 3 and 4 construct the required matrices, $\mathbf{D}$ and $\mathbf{L}$. Based on the above derivation, the optimization is to solve a generalized eigenvalue problem for $\lambda\mathbf{L}-\mathbf{X}^T\mathbf{X}$ as shown in Line 5. Line 6 shows that the first $K-1$ eigenvectors are used to construct a matrix $\mathbf{Q}$. Finally, we use embedding technique to project the data points to a space with $K-1$ dimensions and cluster the data points in the new space with K-means. The above embedding and clustering processes are listed in Lines 7 and 8.

$$\min_{\mathbf{Q}_{K-1}^T\mathbf{Q}_{K-1}=\mathbf{I}_{K-1}} \mathbf{tr}(\mathbf{Q}_{K-1}^T(\lambda\mathbf{L} - \mathbf{X}^T\mathbf{X})\mathbf{Q}_{K-1}) \tag{19}$$

4. EXPERIMENTS

This study uses two synthetic datasets and eight real datasets to conduct experiments. In K-means and FCM, an initial set of prototypes should be given in advance. We use

random approach to determine the initial prototypes. Each evaluation runs ten times. We present the experimental results by using the average with two standard deviations.

This work uses eight datasets in the experiments, including "Arcene", "Planning Relax", "Balance Scale", "Pima Indians", "MAGIC Gamma Telescope(Magic04)", "Connectionist Bench(Sonar)", "Lymphography", and "Musk" datasets. They can be accessed from UCI machine learning repository.

Algorithm 1: Locality Sensitive *K*-means Algorithm

Input: An $n \times m$ document-term matrix $\mathbf{X}$, regularization term parameter λ and the number of clusters K

Output: An $m \times K$ clustering assignment matrix $\mathbf{S}$.

begin
Construct a similarity graph with K-nearest neighbor scheme, and the Gaussian similarity function $s(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp(-\|\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\|^2/(2\sigma^2))$. Let $\mathbf{W}$ be its weighted adjacency matrix.
Construct a diagonal matrix $\mathbf{D}$, where $\mathbf{D}_{ii} = \sum_{j=1}\mathbf{W}_{ij}$.
Construct the Laplacian matrix $\mathbf{L}$, where $\mathbf{L} = \mathbf{D}-\mathbf{W}$.
Compute the first $K-1$ eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_{K-1}$ of $\lambda\mathbf{L}-\mathbf{X}^T\mathbf{X}$, where their corresponding eigenvalues are sorted in ascending order.
Let $\mathbf{Q} \in \mathbb{R}^{m \times (K-1)}$ be the matrix with $\mathbf{q}_1, \ldots, \mathbf{q}_{K-1}$ as columns.
For $i = 1, \ldots, m$, let be the vector corresponding to the ith row of $\mathbf{Q}$.
Cluster the points $\{\mathbf{z}^{(i)}\}_{i=1}^{m} \in \mathbb{R}^{K-1}$ with the K-means algorithm to obtain the clustering assignment matrix $\mathbf{S}$.
return $\mathbf{S}$
end

4.1 Evaluation Measurements

This work evaluates clustering quality using pairwise F_1 cluster evaluation measure [25]. The F_1 cluster evaluation measure considers both precision and recall, where precision and recall here are computed over pairs of documents of which the two cluster assignments either agree or disagree. Four evaluation metrics are necessary for the computation.

- True Positives (TP)
  The clustering algorithm placed the two articles in a pair into the same cluster, and data corpus has them in the same class.
- False Positives (FP)
  The clustering algorithm placed the two articles in a pair into the same cluster, but data corpus has them in differing classes.
- True Negatives (TN)
  The clustering algorithm placed the two articles in a pair into differing clusters, and data corpus has them in differing classes.
- False Negatives (FN)
  The clustering algorithm placed the two articles in a pair into differing clusters, but data corpus has them in the same class.

Similar to traditional information retrieval definition, Eq. (20) shows the formulas of precision, recall and F_1 metric. The F_1 metric is the harmonic mean of precision and recall, so it weights recall and precision equally. A good algorithm should maximize both precision and recall simultaneously, explaining why F_1 metric has been widely used in information retrieval community [27-29]. Consequently, this work uses F_1 to present experimental results.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ F_1 &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{20}$$

4.2 Comparison Methods

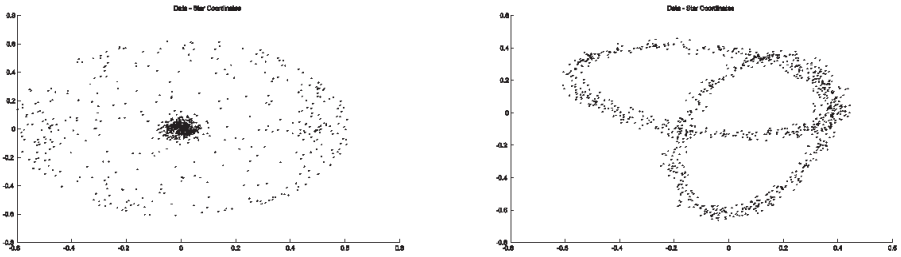
This study compares the proposed algorithm with several unsupervised clustering algorithms, including FCM, *p*-Kmeans, two spectral clustering algorithms, and integrated KL algorithm. In the experiments, the regularization parameter λ of the proposed method is determined by cross validation technique. All of the methods in the experiments are implemented with Matlab.

- FCM
  The FCM is a soft version of *K*-means, allowing one piece of the object to belong to two or more clusters. Each object has a membership degree to indicate the degree belonging to each cluster. FCM is frequently used in pattern recognition.
- *p*-Kmeans
  Zha *et al.* [39] devised a spectral relaxation technique for *K*-means clustering. They showed that a relaxed version of the trace maximization problem possesses global optimal solutions, which can be obtained by computing a partial eigen-decomposition of the Gram matrix.
- Normalized Spectral Clustering
  A major drawback to *K*-means is that it fails to separate clusters that are non-linearly separable in input space [10]. Spectral clustering algorithm [26, 31], which treats clustering task as a graph cut problem, is one of the algorithms that can separate non-linearly separable clusters. The experiments use two normalized spectral clustering algorithms. The first one is devised by Shi and Malik [31]; while the other one is proposed by Ng *et al.* [26].
- Integrated KL (IKL)
  Wang *et al.* [1] proposed a clustering method called IKL, which considers attributes information about data objects and various pairwise relations between data objects. The objective function of IKL is closely related to LDA formulation. Wang *et al.* [1] further relaxed the problem and showed that the optimization solution is composed by the largest K eigenvectors of the matrix $\hat{\mathbf{L}}^{+}\mathbf{X}^{T}\mathbf{X}$, where $\hat{\mathbf{L}}$ is the normalized Laplacian matrix and $\hat{\mathbf{L}}^{+}$ denotes the pseudo inverse of $\hat{\mathbf{L}}$.

4.3 Unsupervised Clustering Experiments

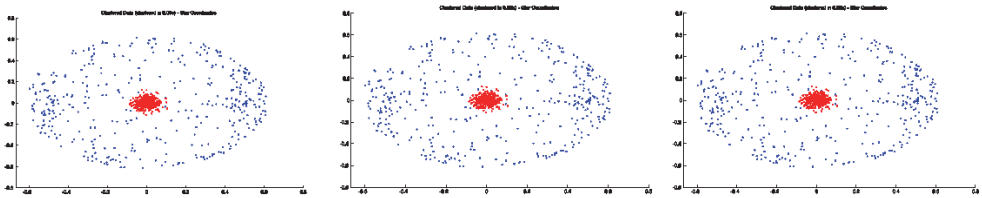
First, this study uses two synthetic datasets to conduct experiments. To further assess the clustering capability of the proposed algorithm, the experiments focus on nonlinearly separable clustering datasets. Figs. 1 (a) and (b) list the scatter plots of the two datasets. The shape of Fig. 1 (a) is like an atom, in which a nucleus is located in the center. The data points in Fig. 1 (b) present a shape of two chains. The two synthetic datasets both comprise two clusters. In the experiments, LS-Kmeans and spectral clustering algorithms need to construct data graph by using Gaussian weighting function and *k*NN scheme, in which the parameters σ and k should be given in advance. The experiments set the σ as 1 and k as 5.

Fig. 2 lists the experimental results on atom dataset with the clustering algorithms mentioned above. The data points with the same color are assigned to the same cluster by the clustering algorithms. Ideally, the data points around the nucleus in Fig. 1 (a) should be assigned to the same cluster, since they are close to each other. As for the two-chains dataset, Fig. 3 lists the experimental results. The best clustering result is to assign the data points in the same chain to the same cluster.

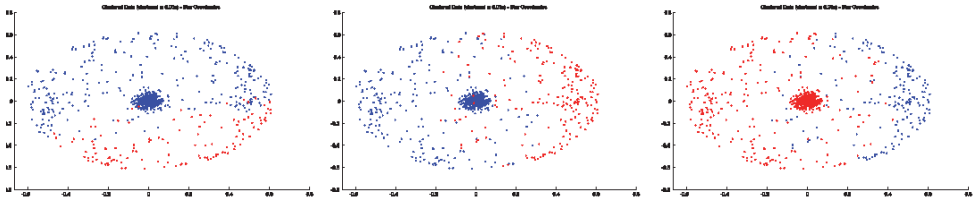


(a) Atom dataset. (b) Two chains dataset.

Fig. 1. Synthetic datasets.



(a) Locality sensitive *K*-means. (b) Spectral clustering (NJW). (c) Spectral clustering (SM).



(d) *K*-means. (e) FCM. (f) PKmeans.

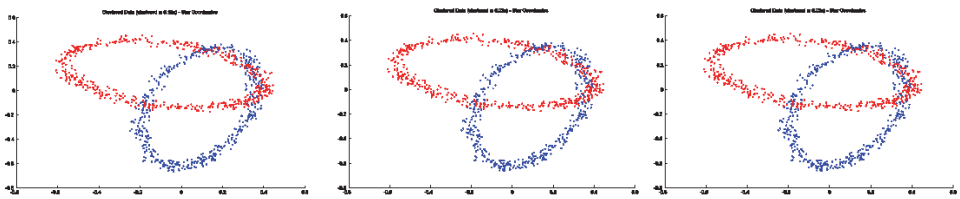
Fig. 2. Clustering results on atom dataset.

Besides, the experiments use eight datasets obtained from UCI machine learning data repository. All the methods in the experiments need seeds to initialize the clustering, and the experiments determine the seeds randomly. Therefore, each evaluation is repeated ten times and the average of the results become the experimental result. Table I indicates the experimental results, each of which is the mean plus or minus two standard deviations. The experiments set σ as 1 and use five nearest neighbors in similarity graph construction.

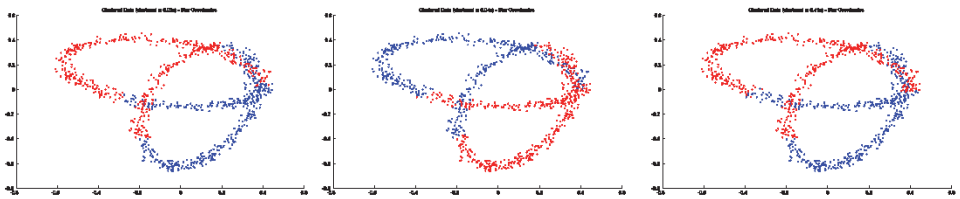
4.4 Discussion

The experiments use two synthetic datasets and eight real datasets to evaluate the proposed method and compare with several clustering methods. As compared with traditional clustering methods, the proposed method considers clustering and dimensionality reduction simultaneously. This study focuses on retaining local geometric structure of the data points in a lower dimensional space, which is an important feature to cluster non-linearly separable data points.

As shown in Figs. 2 and 3, LS-Kmeans and two spectral clustering algorithms can cluster non-linearly separable datasets very well. The LS-Kmeans and spectral clustering algorithms consider geometric structure of the data points, so they tend to assign the data points which are close to each other to the same cluster. Conversely, the other clustering algorithms fail to cluster non-linearly separable data points. The two classical clustering algorithms, *K*-means and FCM, are typically used with Euclidean distance in which case centroids become component wise mean of cluster points. One of the benefits using Euclidean distance is low computational complexity, so *K*-means and FCM are suitable candidates to try on large datasets. However, they often fail when clusters are non-linearly separable or when the data is cluttered with outliers [2]. The experimental results of the syntactic datasets conform to the previous research findings.



(a) Locality sensitive *K*-means. (b) Spectral clustering (NJW). (c) Spectral clustering (SM).



(d) *K*-means. (e) FCM. (f) PKmeans.

Fig. 3. Clustering results on two chains dataset.

Table 1. Experimental results of clustering.

	LS-Kmeans	FCM	SC (NJW)	SC (SM)	*p*-Kmeans	Integrated KL
Arcene	**0.624 ± 0.027**	0.536 ± 0.000	0.536 ± 0.000	0.547 ± 0.026	0.540 ± 0.019	0.535 ± 0.085
Planning Relax	**0.701 ± 0.030**	0.539 ± 0.001	0.554 ± 0.000	0.561 ± 0.000	0.538 ± 0.001	0.543 ± 0.008
Balance Scale	**0.587 ± 0.020**	0.543 ± 0.073	0.578 ± 0.003	0.578 ± 0.001	0.470 ± 0.038	0.505 ± 0.104
Sonar	0.650 ± 0.032	0.502 ± 0.000	**0.652 ± 0.000**	0.652 ± 0.000	0.547 ± 0.081	0.544 ± 0.059
Lymphography	**0.618 ± 0.023**	0.405 ± 0.010	0.413 ± 0.087	0.429 ± 0.083	0.421 ± 0.002	0.386 ± 0.024
Magic04	**0.693 ± 0.000**	0.544 ± 0.000	0.548 ± 0.000	0.554 ± 0.000	0.691 ± 0.000	0.543 ± 0.009
Pima Indians	**0.700 ± 0.008**	0.522 ± 0.000	0.686 ± 0.004	0.693 ± 0.000	0.524 ± 0.002	0.572 ± 0.050
Musk	**0.644 ± 0.000**	0.511 ± 0.004	0.536 ± 0.000	0.540 ± 0.000	0.518 ± 0.000	0.531 ± 0.071

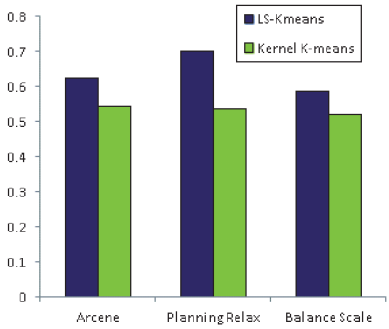


Fig. 4. Performance comparison between LS-Kmeans and Kernel *K*-means.

Besides two synthetic datasets, this study further conducts experiments on eight real datasets, including various domains. As shown in Table 1, the proposed method generally outperforms the other methods. As compared with the two spectral clustering algorithms, the proposed method considers retaining local geometric information and clustering criterion in the objective function, explaining why the proposed method functions properly in the real datasets. Therefore, the experimental results indicate that the proposed algorithm can benefit from locality sensitivity and clustering criteria to cluster non-linearly separable datasets and real datasets.

The proposed method projects the data points into a new space by using dimensionality reduction technique, and then clusters the data points in the new space. Besides dimensionality reduction, projecting the data points into a high-dimensional space and then performing machine learning tasks in the high-dimensional space is another commonly used technique in machine learning. Essentially, it is very difficult to find the explicit mapping function to project the data points into a high-dimensional feature space. Kernel trick offers a way to use dot products between the vectors of data points rather than the explicit mapping. Many machine learning algorithms such as support vector machines (SVM) and kernel *K*-means use kernel trick to perform non-linear classification and clustering, respectively. This study further compares the proposed method with kernel *K*-means. Each evaluation runs ten times, and the average is used as the experimental result. Fig. 4 presents the experimental results, indicating that the proposed method outperforms kernel *K*-means in the experiments. Compared with kernel *K*-means, the proposed method finds a lower dimensional space rather than projecting into a high-dimen-

sional space, and the experimental results indicate that the proposed approach functions properly in the new lower space. One of the reasons is that the proposed method considers local geometric information and clustering criterion to perform linear transformation, so clustering performance can benefit from the dimensionality reduction.

This study proposes a novel objective function to consider clustering and dimensionality reduction simultaneously. The proposed method can be further formalized as a generalized eigenvalue problem, and it can be solved efficiently. Moreover, this study shows that the continuous solutions for the transformed cluster membership indicator vectors of LS-Kmeans are located in the subspace spanned by the first $K - 1$ eigenvector. The experimental results indicate that the proposed method can function properly in non-linearly separable datasets and real datasets.

5. CONCLUSION

This study devises an unsupervised clustering algorithm called LS-Kmeans, which considers clustering and dimensionality reduction simultaneously. Central to LS-Kmeans is considering clustering and locality sensitivity in the objective function, which we argue captures important clustering patterns. Classical clustering algorithms such as *K*-means and FCM fail to cluster nonlinearly separable data points, and the proposed algorithm designs a novel objective function to tackle the problem. We use two synthetic datasets to demonstrate that the proposed algorithms can benefit from locality sensitive information to cluster non-linearly separable data points. Preserving local geometric information has shown to be useful in many application domains, including pattern recognition, information retrieval and multimedia retrieval. This study retains local geometric information while clustering, and applies the proposed algorithm to eight real datasets. The experimental results indicate that the proposed algorithm generally outperforms the other unsupervised clustering algorithms. The future work is to consider dimensionality reduction and clustering simultaneously to devise a semi-supervised learning algorithm with a few labeled examples, so that the clustering performance can be further improved.

ACKNOWLEDGMENTS

This work was supported in part by Ministry of Science and Technology, Taiwan, under Grant No. MOST 105-2221-E-009-092, and partially supported by Project F367-B12110 and conducted at Industrial Technology Research Institute under the sponsorship of the Ministry of Economic Affairs, Taiwan. We are grateful to the National Center for High-Performance Computing for Computer Time and Facilities.

REFERENCES

1. F. Wang, C. H.-Q. Ding, and T. Li, "Integrated *kl* (*k*-means – laplacian) clustering: A new clustering approach by combining attribute data and pairwise relations," in *Proceedings of the 9th SIAM International Conference on Data Mining*, 2009, pp. 38-48.

2. N. Asgharbeygi and A. Maleki, "Geodesic *K*-means clustering," in *Proceedings of the 19th IEEE International Conference on Pattern Recognition*, 2008, pp. 1-4.
3. M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Computation*, Vol. 15, 2003, pp. 1373-1396.
4. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, MA, 1981.
5. L. Breiman, "Random forests," *Machine Learning*, Vol. 45, 2001, pp. 5-32.
6. D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, 2005, pp. 1624-1637.
7. K. Chakrabarty and J. Zeng, "Design automation for microfluidics-based biochips," *ACM Journal on Emerging Technologies in Computing Systems*, Vol. 1, 2005, pp. 186-223.
8. F. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.
9. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, Vol. 41, 1990, pp. 391-407.
10. I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel *K*-means: spectral clustering and normalized cuts," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 551-556.
11. I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, Vol. 42, 2001, pp. 143-175.
12. C. Ding and X. He, "*K*-means clustering via principal component analysis," in *Proceedings of the 21st ACM International Conference on Machine Learning*, 2004, pp. 29-37.
13. E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 35, 2013, pp. 2765-2781.
14. G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
15. X. He, "Incremental semi-supervised subspace learning for image retrieval," in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, 2004, pp. 2-8.
16. X. He and P. Niyogi, "Locality preserving projections," S. Thrun, L. Saul, and B. Schölkopf, ed., *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA, 2004.
17. X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, 2005, pp. 328-340.
18. K. Honda, A. Notsu, and H. Ichihashi, "Fuzzy pca-guided robust *K*-means clustering," *Transactions on Fuzzy Systems*, Vol. 18, 2010, pp. 67-79.
19. H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data*, Vol. 3, 2009, pp. 1:1-1:58.
20. C. A. Kumar and S. Srinivas, "A note on weighted fuzzy *K*-means clustering for concept decomposition," *Cybernetics and Systems*, Vol. 41, 2010, pp. 455-467.

21. Y. Lee, "Semantic-based data mashups using hierarchical clustering and pattern analysis methods," *Journal of Information Science and Engineering*, Vol. 30, 2014, pp. 1601-1618.
22. C. Liu, W. Hsaio, C. Lee, and F. Gou, "Semi-supervised linear discriminant clustering," *IEEE Transactions on Cybernetics*, Vol. 44, 2014, pp. 989-1000.
23. C.-L. Liu, W.-H. Hsaio, C.-H. Lee, and C.-H. Chen, "Clustering tagged documents with labeled and unlabeled documents," *Information Processing & Management*, Vol. 49, 2013, pp. 596-606.
24. K. Lu and D. Yang, "HIC: A robust and efficient hyper-image-based clustering for very large datasets," *Journal of Information and Science Engineering*, Vol. 26, 2010, pp. 461-483.
25. C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*, Cambridge University Press, NY, 2008.
26. A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, Vol. 14, 2001, pp. 849-856.
27. C. Otto, D. Wang, and A. K. Jain, "Clustering millions of faces by identity," *CoRR*, abs/1604.00989, 2016.
28. D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina, "Clustering the tagged web," in *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, 2009. pp. 54-63.
29. E. Rangel, W. Liao, A. Agrawal, A. N. Choudhary, and W. Hendrix, "AGORAS: A fast algorithm for estimating medoids in large datasets," in *Proceedings of International Conference on Computational Science*, 2016, pp. 1159-1169.
30. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, Vol. 290, 2000, pp. 2323-2326.
31. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 888-905.
32. J. B. Tenenbaum, V. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, Vol. 290, 2000, pp. 2319-2323.
33. V. S. Tomar and R. C. Rose, "Noise aware manifold learning for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2013, pp. 7087-7091.
34. W. Tsai, Y. Xu, J. Chien, and W. Huang, "Blind clustering of fingerprints for database indexing," *Journal of Information Science Engineering*, Vol. 30, 2014, pp. 195-212.
35. L. Wang, L. Bo, and L. Jiao, "A modified *K*-means clustering with a density-sensitive distance metric," in *Proceedings of the 1st International Conference on Rough Sets and Knowledge Technology*, 2006, pp. 544-551.
36. A. Watve, S. Pramanik, S. Jung, B. Jo, S. Kumar, and S. Sural, "Clustering non-ordered discrete data," *Journal of Information Science and Engineering*, Vol. 30, 2014, pp. 1-23.
37. T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using wordnet and lexical chains," *Expert Systems with Application*, Vol. 42, 2015, pp. 2264-2275.

38. R. Xu, N. Jiang, N. Mrachacz-Kersting, C. Lin, G. A. Prieto, J. C. Moreno, J. L. Pons, K. Dremstrup, and D. Farina, "A closed-loop brain-computer interface triggering an active ankle-foot orthosis for inducing cortical neural plasticity," *IEEE Transactions on Biomedical Engineering*, Vol. 61, 2014, pp. 2092-2101.
39. H. Zha, X. He, C. H. Q. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for *K*-means clustering," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 2001, pp. 1057-1064.
40. J. Zhao, K. Lu, and X. He, "Locality sensitive semi-supervised feature selection," *Neurocomputing*, Vol. 71, 2008, pp. 1842-1849.
41. F. Zhong, J. Zhang, and D. Li, "Discriminant locality preserving projections based on *l*1-norm maximization," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 25, 2014, pp. 2065-2074.

Chien-Liang Liu (劉建良) received the M.S. and Ph.D. degrees in Department of Computer Science from National Chiao Tung University, Taiwan, in 2000 and 2005, respectively. He is currently an Assistant Professor in Department of Industrial Engineering and Management at National Chiao Tung University, Taiwan. His research interests include machine learning, data mining, and big data analytics.

Wen-Hoar Hsaio (蕭文豪) received the M.S. and Ph.D. degrees in Department of Computer Science from National Chiao Tung University, Taiwan, in 1996 and 2015. He is currently an Engineer in National Chung-Shan Institute of Science and Technology, Taiwan. His research interests include information retrieval, data mining, and machine learning.

Tao-Hsing Chang (張道行) received his Ph.D. in Computer Science at National Chiao Tung University in 2007. From 1999 to 2008, he was the Research Fellow with the Research Center for Psychological and Educational Testing at National Taiwan Normal University. He joined National Kaohsiung University of Applied Sciences in 2008 and is currently an Associate Professor in Department of Computer Science and Information Engineering. His research interests are centered around natural language processing, information retrieval, soft computing, and educational technology.