Semi-supervised Linear Discriminant Clustering

Chien-Liang Liu, *Member, IEEE,* Wen-Hoar Hsaio, Chia-Hoang Lee, and Fu-Sheng Gou

***Abstract*—This paper devises a semi-supervised learning method called semi-supervised linear discriminant clustering (Semi-LDC). The proposed algorithm considers clustering and dimensionality reduction simultaneously by connecting *K*-means and linear discriminant analysis (LDA). The goal is to find a feature space where the *K*-means can perform well in the new space. To exploit the information brought by unlabeled examples, this paper proposes to use soft labels to denote the labels of unlabeled examples. The Semi-LDC uses the proposed algorithm, called constrained-PLSA, to estimate the soft labels of unlabeled examples. We use soft LDA with hard labels of labeled examples and soft labels of unlabeled examples to find a projection matrix. The clustering is then performed in the new feature space. We conduct experiments on three data sets. The experimental results indicate that the proposed method can generally outperform other semi-supervised methods. We further discuss and analyze the influence of soft labels on classification performance by conducting experiments with different percentages of labeled examples. The finding shows that using soft labels can improve performance particularly when the number of available labeled examples is insufficient to train a robust and accurate model. Additionally, the proposed method can be viewed as a framework, since different soft label estimation methods can be used in the proposed method according to application requirements.**

***Index Terms*—Clustering, linear discriminant analysis, semi-supervised learning, soft label, text mining.**

I. Introduction

CLUSTERING is one of the most frequently encountered tasks of machine learning. The goal of clustering is to automatically assign objects into groups so that objects from the same cluster are more similar to each other than objects from different clusters. It is also one of the most widely used techniques for exploratory data analysis, since it can capture the natural structure of the data. Additionally, clustering is an unsupervised learning approach, so it does not require labeled data during the course of clustering. The K-means is a typical clustering algorithm, which aims at the minimization of the average squared distance between the objects and the cluster centers. The K-means generally uses Euclidean distance as the distance metric, explaining why it can have a good performance on the data set with compact super-sphere distributions, but tends to fail in the data organized in more complex and unknown shapes [1]. In many applications such as document classification and pattern recognition, each object generally comprises thousands of features. One of the problems with high-dimensional data sets is that not all the measured variables are important for understanding the underlying phenomena of interest. As a result, K-means does not generally perform well on high-dimensional data sets. This paper proposes to use dimensionality reduction technique to improve K-means clustering performance. The goal is to find an appropriate feature space where the K-means can perform well in the new space. We propose a method called semi-supervised linear discriminant clustering (Semi-LDC), which connects K-means and linear discriminant analysis (LDA), to consider clustering and dimensionality reduction simultaneously. The K-means is an unsupervised learning method, while LDA is a supervised dimensionality reduction method. The goal of LDA is to find a vector which can separate two or more classes of objects, but LDA requires sufficient labeled examples to obtain the projection vector. Labeling is a time-consuming process, since it is typically done manually. Conversely, unlabeled data is relatively easy to collect, explaining why the proposed method uses a few labeled examples and enormous unlabeled examples in finding a feature space where K-means can function well. The proposed method can be applied to the situations where only a few labeled examples are available. One typical example is to classify news documents into different categories with a few labeled news documents. Semi-supervised learning, learning from a combination of both labeled and unlabeled data, has become a topic of significant recent interest [2]–[7]. This paper proposes to use soft labels to denote the labels of unlabeled examples due to the uncertainty on the estimation. Compared with hard labels of labeled examples, soft labels allow each object to belong to all of the clusters with membership degrees or probabilities. In text analysis, Hofmann [8] proposed probabilistic latent semantic analysis (PLSA) for factor analysis of binary and count data, such as text data collected by counting terms occurring in documents with the bag-of-words representation or images represented through feature counts. Notably, two PLSA models are the aspect model and statistical clustering model [8], [9]. One important characteristic of PLSA is that it defines a proper generative data model and views the topic as a latent variable, since topic is not directly observed but is rather inferred from observed data. In PLSA aspect model,

Manuscript received November 13, 2012; revised July 26, 2013; accepted July 30, 2013. Date of publication August 27, 2013; date of current version June 12, 2014. This work was supported in part by the National Science Council under Grant NSC-102-2221-E-009-102 and in part by the Ministry of Economic Affairs, Taiwan, through project No. C301AR2910, conducted by the Industrial Technology Research Institute. This paper was recommended by Associate Editor J. Zhang.

C.-L. Liu is with the Computational Intelligence Technology Center, Industrial Technology Research Institute, Hsinchu 31040, Taiwan (e-mail: jacky168@gmail.com).

W.-H. Hsaio and C.-H. Lee are with the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan (e-mail: mr.papa@msa.hinet.net; chl@cs.nctu.edu.tw).

F.-S. Gou is with the Chung Shan Institute of Science and Technology, Taoyuan, Taiwan (e-mail: goby.cs99g@nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2013.2278466

the latent variable z is introduced for each observation $(\mathbf{x}_i, w_j)$ over a finite set $\mathcal{Z} = \{z_1, \ldots, z_K\}$, where the pair $(\mathbf{x}_i, w_j)$ represents the occurrence of a term w_j in a document $\mathbf{x}_i$. In the clustering model for documents, PLSA clustering model assumes that each document belongs to exactly one cluster and it is only the finiteness of the number of observations per document that induces uncertainty about a document's cluster membership. This paper uses the proposed constrained-PLSA algorithm, which is an extension of PLSA clustering model with a few labeled examples, to obtain soft labels of unlabeled examples. The constrained-PLSA, an expectation maximization (EM) algorithm [10], uses the available labeled examples as constraints to bias the clustering toward a good region of the search space. The output of constrained-PLSA algorithm is a document-topic matrix, whose entry indicates the probability of a document belonging to a specific cluster. The proposed method uses the soft labels to find a projection matrix by using soft LDA, and then clusters the data points in the new feature space. The main contribution of this paper is that this paper devises a semi-supervised algorithm called semi-supervised linear discriminant clustering. Compared with traditional methods, the proposed method considers clustering and dimensionality reduction simultaneously to devise the algorithm. Although this paper focuses on document clustering problem, the proposed method connects K-means and LDA seamlessly, making it feasible to extend the proposed method to the other problem domains. Additionally, this paper also shows that the proposed method can be mapped to a high-dimensional feature space by means of kernel trick. We conduct experiments on three data sets, and experimental results indicate that the proposed method generally outperforms other semi-supervised learning methods. The rest of this paper is organized as follows. Section II presents related surveys of semi-supervised learning and high-dimensional data. Section III then introduces constrained-PLSA algorithm and Semi-LDC algorithm. Next, Section IV summarizes the results of several experiments. Conclusions are drawn in Section V.

II. Related Work

A. Semi-supervised Learning

Semi-supervised learning methods can be further classified into semi-supervised classification and semi-supervised clustering methods. Semi-supervised classification employs labeled data along with unlabeled data to construct a more accurate classifier; while semi-supervised clustering uses a few labeled data to bias the clustering of unlabeled data. Various semi-supervised algorithms have been proposed, including co-training [2], [5], semi-supervised naive Bayes [3], transductive support vector machines (TSVM) [11], graph-based approaches [12], [13], and clustering-based approaches [14]–[16].

Transductive support vector machines (TSVM) [11], which is an extension of standard support vector machines with unlabeled data, is a typical semi-supervised classification method. Unlike standard SVM, unlabeled examples are included in TSVM model training. Although unsupervised learning approaches do not need labeled data during the course of clustering process, proper seeding biases clustering toward a good region of the search space [15]. Wagstaff *et al.* [14] devised a semi-supervised variant of K-means called COP-KMeans to use constraints in order to represent background knowledge. Two constraints are must-link (i.e., two instances must be together in the same cluster) and cannot-link (i.e., two instances must be in different clusters). These constraints are used during clustering to generate a partition that satisfies all given constraints. Basu *et al.* [15] introduced two semi-supervised variants of K-means clustering that use initial labeled data for seeding, and the experimental results indicated that their proposed method outperforms COP-KMeans. Practically, a good initial labeled seeds provide guidance for semi-supervised clustering methods to obtain reliable clustering results. Nie *et al.* [17] proposed an actively self-training clustering method, in which the samples are actively selected as training set to minimize an estimated Bayes error, and then explore semi-supervised learning to perform clustering.

Many semi-supervised learning methods use optimization with constraints technique to view labeled examples as constraints. For instance, many clustering algorithms aim at the minimization of the cost function, which involves distortion measure between the objects and the cluster representatives. Besides the original objective function, semi-supervised learning can formalize labeled examples as regularization terms. This technique has been widely used by many researchers [18]–[20]. For instance, Bouchachia and Pedrycz [21] developed a semi-supervised learning algorithm, which extends the objective function of fuzzy c-means (FCM) [22] to encode labeled data as an additional regularization term. Miyamoto *et al.* [20] used the same technique in fuzzy semi-supervised learning.

B. High-dimensional Data

High-dimensional data presents many mathematical challenges to machine learning tasks. One of the problems with high-dimensional data is that not all the measured variables are important for understanding the underlying phenomena of interest [23]. The assessments on concepts of distance or nearest neighbor are deteriorated in high-dimensional data due to the curse of dimensionality problem. Outlier detection in the high dimensional space is a typical application, since its goal is to identify the objects that considerably deviate from the general distribution of the data. Thus, Kriegel *et al.* [24] proposed a novel approach named angle-based outlier detection and some variants assessing the variance in the angles between the difference vectors of a point to the other points, since angles are more stable than distances in high-dimensional space. To further reduce time complexity, Pham and Pagh [25] used random projection technique to propose a near-linear time algorithm to approximate the variance of angles for each data object. Besides outlier detection, distance metrics are also crucial to many learning algorithms. For instance, kNN classifier has to identify the set of labeled examples that are closest to a given test example in the feature space, which involves the estimation of a distance metric. Therefore, metric learning is another popular approach to process the high-dimensional data. Previous works [26]–[28]

TABLE I
NOTATION

Notation	Meaning
N	The number of documents
M	The number of features
K	The number of clusters
N_c	The number of documents in the cth cluster ($c = 1, \dots, K$)
$\mathbf{x}_i$	The ith document ($i = 1, \dots, N$)
$\mathbf{x}_i^{(c)}$	The ith document of cluster c ($i = 1, \dots, N_c$ and $c = 1, \dots, K$)
$\mathbf{m}_c$	The mean of the cth cluster ($c = 1, \dots, K$)
$\mathbf{m}_0$	The total mean

have shown that appropriately designed distance metrics can significantly improve performance compared to the standard Euclidean distance. For instance, Goldberger *et al.* [26] proposed neighborhood component analysis (NCA) algorithm for learning a Mahalanobis distance metric to be used in the *k*NN classifier by maximizing the leave-one-out cross validation. Weinberger and Saul [27] proposed a distance metric learning algorithm to learn a Mahalanobis distance metric for *k*NN classification from labeled examples. The metric is trained with the goal that the *k*-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin.

Dimensionality reduction, which tries to find a lower dimensional representation of the data according to some criterion, is essentially to learn a distance metric without label information [29]. Dimensionality reduction approaches assume that the data of interest lies on an embedded linear subspace or nonlinear manifold within the higher-dimensional space, and they are commonly used techniques for visualization [30] and feature extraction. Many dimensionality reduction algorithms have been developed to accomplish these tasks. Random projections, projecting original high-dimensional data onto a lower-dimensional subspace using a random matrix whose columns have unit lengths, have recently emerged as a powerful method for dimensionality reduction [31]. Besides randomized algorithms, principal component analysis (PCA), 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, including Isomap [32], locally linear embedding (LLE) [33], and Laplacian eigenmaps [34] have been proposed.

Among these methods, PCA and LDA are two typical linear dimensionality reduction methods, but they use different criteria to reduce dimensionality. PCA is an unsupervised learning method, and the goal is to perform dimensionality reduction while preserving as much of the variance in the high-dimensional space as possible. Conversely, LDA is a supervised learning method, and the goal is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. In other words, LDA aims at finding a feature representation by which the within-class distance is minimized and the between-class distance is maximized. The two criteria lead to LDA objective function, which can be further transformed into a generalized eigenvalue problem.

III. SEMI-SUPERVISED LINEAR DISCRIMINANT CLUSTERING

A. Notation

The notations that will be used in the following sections are listed in Table I. Each document $\mathbf{x}_i$ is represented as a feature vector, whose length is M. There are N documents in the collection, and the goal is to partition the document collection into K clusters, each of which comprises N_c, $1 \leq c \leq K$, documents. The cluster center or mean for each cluster c is $\mathbf{m}_c$, and the center for all of the documents is $\mathbf{m}_0$. Additionally, we use a matrix $\mathbf{X}$ to denote all document vectors as shown in (1)

$$\mathbf{X}^T = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]. \tag{1}$$

B. Linear Discriminant Clustering

The K-means is a typical clustering algorithm, which aims at the minimization of the average squared distance between the objects and the cluster centers. Equation (2) shows the objective function of K-means. The K-means generally uses Euclidean distance as the distance metric, explaining why it can have a good performance on the data with compact supersphere distributions, but tends to fail in the data organized in more complex and unknown shapes [1]. However, the analysis on high-dimensional data sets becomes a topic of significant recent interest due to the advances in data collection and storage capabilities during the past decades.

The K-means objective function listed in (2) is performed in the original input space. This paper proposes to use dimensionality reduction technique to find an appropriate feature space, so that the clustering can perform well in the new feature space. We start the derivation from projecting data points onto a line by using a projection vector $\mathbf{a}$, and then the original objective function of K-means can be represented as the form listed in (3), where $\mathbf{S}_w$ is $\frac{1}{N}\sum_{c=1}^{K}\sum_{i=1}^{N_c}(\mathbf{x}_i^{(c)} - \mathbf{m}_c)(\mathbf{x}_i^{(c)} - \mathbf{m}_c)^T$. The goal becomes to find a projection vector $\mathbf{a}$ to minimize the objective function listed in (3). Moreover, Ding and He [35] have shown that the minimization of K-means objective function implicitly implies that the average between-class distances should be maximized. On the other hand, the $\mathbf{S}_w$ presented in (3) is equal to the within-class scatter matrix of LDA. As a result, the derivation of (3) and the between-class criterion can connect K-means and LDA. Essentially, K-means and LDA have different objectives. The K-means is a clustering algorithm; while LDA is generally used for dimensionality reduction, aiming at finding a linear combination of features which characterizes or separates two or more classes of objects. The above derivation shows that using a projection vector on K-means objective function and considering between-class criterion in the new feature space can connect K-means and LDA seamlessly. We call the above processes as linear discriminant clustering (LDC), since it considers dimensionality reduction and clustering simultaneously

$$\begin{aligned} J &= \frac{1}{N}\sum_{c=1}^{K}\sum_{i=1}^{N_c}||\mathbf{x}_i^{(c)} - \mathbf{m}_c||^2 \\ &= \frac{1}{N}\sum_{c=1}^{K}\sum_{i=1}^{N_c}(\mathbf{x}_i^{(c)} - \mathbf{m}_c)^T(\mathbf{x}_i^{(c)} - \mathbf{m}_c) \end{aligned} \tag{2}$$

$$
\begin{aligned}
J(\mathbf{a}) &= \frac{1}{N}\sum_{c=1}^{K}\sum_{i=1}^{N_c}(\mathbf{a}^T\mathbf{x}_i^{(c)} - \mathbf{a}^T\mathbf{m}_c)^T(\mathbf{a}^T\mathbf{x}_i^{(c)} - \mathbf{a}^T\mathbf{m}_c) \\
&= \mathbf{a}^T\frac{1}{N}\sum_{c=1}^{K}\sum_{i=1}^{N_c}(\mathbf{x}_i^{(c)} - \mathbf{m}_c)(\mathbf{x}_i^{(c)} - \mathbf{m}_c)^T\mathbf{a} \\
&= \mathbf{a}^T\mathbf{S}_w\mathbf{a}. \qquad (3)
\end{aligned}
$$

As an unsupervised learning method, the K-means does not require labeled examples during the course of clustering. Conversely, LDA is a supervised dimensionality reduction method, meaning that sufficient labeled examples are required in finding a projection vector which can separate two or more classes of objects. Labeling must typically be done manually and it is a time-consuming process obviously. In general, unlabeled data is relatively easy to collect. Although unsupervised dimensionality reduction methods do not require labeled examples, they do not generally consider classification or clustering criteria when performing dimensionality reduction. As a result, this paper proposes an algorithm called semi-supervised linear discriminant clustering (Semi-LDC) to use a few labeled examples to find an appropriate feature space where K-means can function properly. Essentially, using insufficient labeled examples cannot determine a reliable discriminative projection vector, explaining why we use soft labels to denote the labels of unlabeled examples. Compared with hard labels of labeled examples, soft labels allow each object to belong to all of the clusters with different membership degrees or probabilities. This paper uses the proposed constrained-PLSA method to obtain soft labels of unlabeled examples.

C. Soft Label via Constrained-PLSA

Inspired by latent semantic analysis (LSA), Hofmann [8] proposed PLSA for factor analysis of binary and count data. PLSA comprises several important properties. First, it is an unsupervised learning method, so it does not require labeled data. Second, as a generative model, PLSA is based on a mixture decomposition derived from a latent class model, where the latent variable is discrete. Third, the latent variable introduced by PLSA can infer more semantic information from the observations. For instance, PLSA can handle polysemy problem, namely, a word with many possible meanings.

The constrained-PLSA proposed in this paper is an extension of PLSA clustering model with a few labeled examples. The main difference between the constrained-PLSA and the conventional PLSA is that the conventional PLSA is an unsupervised learning method, while the proposed constrained-PLSA is a semi-supervised learning method, which extends the PLSA by using the seeds to direct the clustering to toward a good region of the search space. The constrained-PLSA can estimate maximum likelihood in latent variable models using the EM algorithm [10]. The E-step is to calculate the probability of the latent variables, given the observed variables and the current values of the parameters. Then the posterior distribution is used to compute the expected complete data log likelihood to estimate the new parameter value in the M-step. Meanwhile, convergence is assured since the EM algorithm is guaranteed to increase the likelihood at each iteration. Equation (4) presents the E-step, where $\mathbf{Q}$ represents the posterior probability distribution of latent variable and $n(\mathbf{x}_i, w_j)$ denotes the term frequency, that is, the number of times w_j occurred in $\mathbf{x}_i$. The probability matrix $\mathbf{Q}$ is a $N \times K$ matrix, and each entry $\mathbf{Q}_{ik}$ represents the probability of document $\mathbf{x}_i$ assigned to topic k. The topic-term distribution matrix $\mathbf{\Theta}$ is a $K \times M$ matrix, where each row $\mathbf{\Theta}_k$ represents a topic and the entry value $\mathbf{\Theta}_{kj}$ represents the probability of topic k generating term w_j

$$
\begin{aligned}
\mathbf{Q}_{ik} &= P(z_k|\mathbf{x}_i) \\
&= P(z_k)\exp\Big(\sum_{j=1}^{M} n(\mathbf{x}_i, w_j)\ln \mathbf{\Theta}_{kj}\Big). \qquad (4)
\end{aligned}
$$

Algorithm 1: constrained-PLSA Algorithm

```
Input: A $N \times M$ document-term matrix $\mathbf{X}$, the number
       of topics $K$ and the seeds $S = \{S_1, \ldots, S_K\}$.
       Without loss of generality, $S_k$ represents the
       document seeds for topic $k$ ($k = 1, \ldots, K$).
Output: A $N \times K$ document-topic matrix $\mathbf{Q}$.
begin
    $\mathbf{X}_i \longleftarrow \frac{\mathbf{X}_i}{\sum_j \mathbf{X}_{ij}}$, for $i = 1, \ldots, N$;
    $\mathbf{\Theta}_k \longleftarrow \frac{1}{|S_k|}\sum_{\mathbf{x}_i \in S_k} \mathbf{X}_i$, for $k = 1, \ldots, K$;
    Initialize topic proportion components
    $P(z_1) = \ldots = P(z_K) = \frac{1}{K}$;
    repeat
        E-step: to compute latent variable posterior
        probability $\mathbf{Q}$ according to Equation (4);
        normalize $\mathbf{Q}_i$, for $i = 1, \ldots, N$ ;
        $\mathbf{Q}_{S_k,k} \longleftarrow 1$ and $\mathbf{Q}_{S_k,l} \longleftarrow 0$ where $l \neq k$, for
        $k = 1, \ldots, K$ ;
        Normalize $\mathbf{Q}_{S_k}$, for $k = 1, \ldots, K$ ;
        M-step: to update proportion parameter $P(z_k)$ and
        $\mathbf{\Theta}_k$ according to Equation (5) and Equation (6),
        respectively, for $k = 1, \ldots, K$
    until convergence ;
    return $Q$
end
```

Then, the Lagrangian function can be obtained based on the expected complete log likelihood function and the probability constraints, $\sum_{k=1}^{K} P(z_k) = 1$ and $\sum_{j=1}^{n} P(w_j|z_k) = 1$. Maximization of Lagrange function with respect to the probability mass functions leads to the following set of stationary equations as shown in (5) and (6), where $\mathbf{Q}_{ik} = P(z_k|\mathbf{x}_i)$. Meanwhile, (5) and (6) are the estimated parameters in M-step

$$
\begin{aligned}
P(z_k) &= \frac{\sum_{i=1}^{N} P(z_k|\mathbf{x}_i)}{\sum_{k'=1}^{K}\sum_{i=1}^{N} P(z_{k'}|\mathbf{x}_i)} \\
&= \frac{\sum_{i=1}^{N} \mathbf{Q}_{ik}}{\sum_{k'=1}^{K}\sum_{i=1}^{N} \mathbf{Q}_{ik'}} \qquad (5)
\end{aligned}
$$

$$\boldsymbol{\Theta}_{kj} = \frac{\sum_{i=1}^{N} P(z_k|\mathbf{x}_i)n(\mathbf{x}_i, w_j)}{\sum_{j=1}^{M}\sum_{i=1}^{N} P(z_k|\mathbf{x}_i)n(\mathbf{x}_i, w_j)} = \frac{\sum_{i=1}^{N} \mathbf{Q}_{ik}n(\mathbf{x}_i, w_j)}{\sum_{j=1}^{M}\sum_{i=1}^{N} \mathbf{Q}_{ik}n(\mathbf{x}_i, w_j)}. \tag{6}$$

Algorithm 1 shows the constrained-PLSA algorithm. The inputs of the constrained-PLSA include document term matrix $\mathbf{X}$, the number of topics K and the document seeds, $S_1, \ldots, S_K$. Without loss of generality, S_k represents the document seeds for topic k. The output of the constrained-PLSA is the document-topic matrix $\mathbf{Q}$ as shown in (7). In the algorithm, the initialization steps are listed in Line 2-4. Then, using E-step and M-step described above to update parameters until convergence

$$\mathbf{Q} = [\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \ldots, \mathbf{Q}^{(K)}] \in \mathbb{R}^{N\times K}. \tag{7}$$

D. Soft LDA

On completion of constrained-PLSA algorithm, we can obtain a document-topic matrix $\mathbf{Q}$, which can be further interpreted as the soft label matrix for the input documents. This paper calls the LDA with soft labels as soft LDA, since the original LDA only uses hard labels to compute projection vector. Nie *et al.* [36] have extended the scatter matrices defined in LDA to the soft label based scatter matrices, inspiring this paper to use the same idea to use soft labels.

The soft labels obtained from $\mathbf{Q}$ do not belong to crisp set any more, since each document can belong to more than one cluster with probability. As a result, the number of data in the cth cluster and total number of data have to be redefined as shown in (8). Similarly, the mean of a cluster should consider soft labels as well. Equation (9) shows the matrix form representation. Additionally, we introduce two diagonal matrices, $\mathbf{D}$ and $\mathbf{B}^{soft}$, where the diagonal entry $\mathbf{D}_{cc}$ is the soft number of data for cth cluster, and the diagonal entry $\mathbf{B}^{soft}_{ii}$ is the sum of membership degrees for ith data. Equation (10) and (11) show the definitions of $\mathbf{D}$ and $\mathbf{B}^{soft}$, respectively

$$\tilde{N}_c = \sum_{i=1}^{N} \mathbf{Q}_{ic}$$
$$\tilde{N} = \sum_{c=1}^{K} \tilde{N}_c \tag{8}$$

$$\begin{aligned}\tilde{\mathbf{m}}_c &= \frac{1}{\tilde{N}_c}\sum_{i=1}^{N}\mathbf{Q}_{ic}\mathbf{x}_i \\ &= \frac{1}{\tilde{N}_c}[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N][\mathbf{Q}_{1c}, \mathbf{Q}_{2c}, \ldots, \mathbf{Q}_{Nc}]^T \\ &= \frac{1}{\tilde{N}_c}\mathbf{X}^T\mathbf{Q}^{(c)}\end{aligned} \tag{9}$$

$$\mathbf{D} = \begin{pmatrix} \frac{1}{\tilde{N}_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\tilde{N}_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\tilde{N}_K} \end{pmatrix} \tag{10}$$

$$\mathbf{B}^{soft} = \begin{pmatrix} \sum_{c=1}^{K}\mathbf{Q}_{1c} & 0 & \cdots & 0 \\ 0 & \sum_{c=1}^{K}\mathbf{Q}_{2c} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{c=1}^{K}\mathbf{Q}_{Nc} \end{pmatrix}. \tag{11}$$

Next, we can follow the definition of between-class scatter matrix to derive soft label one. The soft between-class scatter matrix $\tilde{\mathbf{S}}_b$ can be represented as a matrix form listed in (12) by using the matrix form representation of $\tilde{\mathbf{m}}_c$, in which the soft total mean $\tilde{\mathbf{m}}_0$ can be eliminated due to zero mean normalization technique. Similarly, the soft within-class scatter matrix can be represented as a matrix form listed in (13)

$$\begin{aligned}\tilde{\mathbf{S}}_b &= \sum_{c=1}^{K}\frac{\tilde{N}_c}{\tilde{N}}(\tilde{\mathbf{m}}_c - \tilde{\mathbf{m}}_0)(\tilde{\mathbf{m}}_c - \tilde{\mathbf{m}}_0)^T \\ &= \sum_{c=1}^{K}\frac{\tilde{N}_c}{\tilde{N}}\tilde{\mathbf{m}}_c\tilde{\mathbf{m}}_c^T \\ &= \frac{1}{\tilde{N}}\sum_{c=1}^{K}\tilde{N}_c\left(\frac{1}{\tilde{N}_c}\mathbf{X}^T\mathbf{Q}^{(c)}\right)\left(\frac{1}{\tilde{N}_c}\mathbf{X}^T\mathbf{Q}^{(c)}\right)^T \\ &= \frac{1}{\tilde{N}}\mathbf{X}^T\left(\sum_{c=1}^{K}\frac{1}{\tilde{N}_c}\mathbf{Q}^{(c)}\mathbf{Q}^{(c)^T}\right)\mathbf{X} \\ &= \frac{1}{\tilde{N}}\mathbf{X}^T\mathbf{Q}\mathbf{D}\mathbf{Q}^T\mathbf{X}\end{aligned} \tag{12}$$

$$\begin{aligned}\tilde{\mathbf{S}}_w &= \frac{1}{\tilde{N}}\sum_{c=1}^{K}\sum_{i=1}^{N}\mathbf{Q}_{ic}(\mathbf{x}_i - \tilde{\mathbf{m}}_c)(\mathbf{x}_i - \tilde{\mathbf{m}}_c)^T \\ &= \frac{1}{\tilde{N}}\sum_{c=1}^{K}(\sum_{i=1}^{N}\mathbf{Q}_{ic}\mathbf{x}_i\mathbf{x}_i^T - \tilde{N}_c\tilde{\mathbf{m}}_c\tilde{\mathbf{m}}_c^T) \\ &= \frac{1}{\tilde{N}}\mathbf{X}^T\left(\mathbf{B}^{soft} - \mathbf{Q}\mathbf{D}\mathbf{Q}^T\right)\mathbf{X}.\end{aligned} \tag{13}$$

Using matrix form to represent scatter matrices has several advantages. First, matrix form provides an elegant way to represent the formula in a compact form. Thus, it is easy to formulate the soft LDA problem as an optimization problem listed in (14). Classical LDA is not applicable for small sample size problems due to the singularity of the within-class scatter matrices involved [37], since the dimension of sample exceeds the number of samples. Regularization techniques can be applied to deal with the singularity problem of LDA by adding a constant $\mu > 0$ to the diagonal elements of $\tilde{\mathbf{S}}_w$ as $\tilde{\mathbf{S}}_w + \mu\mathbf{I}$, where $\mathbf{I}$ is an identity matrix [38]–[40]. The optimization problem can be transformed into a generalized eigenvalue problem. Second, matrix is the basic data element in some programming languages such as MATLAB, and vectorization technique can be used in the programs to speed up the code. Third, the soft LDA can be employed in a nonlinear way by means of the kernel trick, which only requires dot products between the vectors in feature space, and chooses the mapping such that these high-dimensional dot products can be computed within the original space by means of a kernel function

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}} \frac{\mathbf{a}^T \tilde{\mathbf{S}}_b \mathbf{a}}{\mathbf{a}^T \left(\tilde{\mathbf{S}}_w + \mu \mathbf{I}\right) \mathbf{a}}. \tag{14}$$

The extension to kernelized soft LDA can be achieved by introducing a matrix $\mathbf{\Phi}$ as shown in (15), which denotes the data in the feature space using the mapping function $\boldsymbol{\phi}$. Then, using $\mathbf{\Phi}$ to replace original data $\mathbf{X}$ leads to the kernelized soft LDA formula as shown in (16). Moreover, the projection vector $\mathbf{a}$ can be rewritten as the linear combination of data as shown in (17). Replacing the projection vector $\mathbf{a}$ with the matrix form listed in (17) can lead to (18) in which a kernel matrix $\mathbf{K}$ can be used to replace the dot product of data, that is, $\mathbf{\Phi\Phi}^T$

$$\mathbf{\Phi}^T = [\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)] \tag{15}$$

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}} \frac{\mathbf{a}^T \mathbf{\Phi}^T \mathbf{QDQ}^T \mathbf{\Phi a}}{\mathbf{a}^T \mathbf{\Phi}^T \left(\mathbf{B}^{soft} - \mathbf{QDQ}^T\right) \mathbf{\Phi a}} \tag{16}$$

$$\mathbf{a} = \sum_{i=1}^{N} \boldsymbol{\alpha}_i \phi(\mathbf{x}_i) = \mathbf{\Phi}^T \boldsymbol{\alpha} \tag{17}$$

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \frac{\boldsymbol{\alpha}^T \mathbf{\Phi\Phi}^T \mathbf{QDQ}^T \mathbf{\Phi\Phi}^T \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{\Phi\Phi}^T \left(\mathbf{B}^{soft} - \mathbf{QDQ}^T\right) \mathbf{\Phi\Phi}^T \boldsymbol{\alpha}} \\ &= \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \frac{\boldsymbol{\alpha}^T \mathbf{KQDQ}^T \mathbf{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{K} \left(\mathbf{B}^{soft} - \mathbf{QDQ}^T\right) \mathbf{K}\boldsymbol{\alpha}}. \end{aligned} \tag{18}$$

E. Semi-supervised Linear Discriminant Clustering

In information retrieval, document representation is often based on the bag-of-words model, where a document is represented as an unordered collection of words, disregarding grammar and even word order. The bag-of-words approach to document representation typically results in vectors of the order of 5,000–20,000 components as the representation of documents. Torkkola [41] has shown that LDA can be used to reduce drastically the dimension of document representation in classification tasks without sacrificing the accuracy with a small number of discriminative features obtained from latent semantic indexing (LSI) or PCA. Moreover, applying PCA first for dimensionality reduction is also a method to make the within-class scatter matrix nonsingular before the application of LDA [42]. Consequently, this paper uses PCA to reduce dimensions first, and then uses soft LDA with soft labels obtained from constrained-PLSA to find discriminative features. This paper further conducts experiments to analyze the influence of different dimensionality reduction methods on classification performance.

Algorithm 2 shows semi-supervised linear discriminant clustering (Semi-LDC) algorithm. The inputs include document term matrix $\mathbf{X}$, the number of clusters K and the seeds $S = \{S_1, \ldots, S_K\}$. The algorithm uses constrained-PLSA to obtain a soft label matrix $\mathbf{Q}$ as shown in Line 2. Line 3 shows that we use PCA to reduce dimensions of input document term matrix $\mathbf{X}$ due to the reason described above. Next, the proposed algorithm uses soft LDA to estimate the best projection matrix $\mathbf{A}$ in Line 4–8. The number of the reduced dimensions in LDA is $K-1$ conventionally, where K is the number of classes. This paper follows the same scheme to determine the number of the reduced dimensions in soft LDA. Then, the original documents can be projected to a lower dimensional space in which the classes are well separated as shown in Line 9. Finally, we apply constrained-KMeans [15] to the data points in the new space to obtain the final clustering results.

Algorithm 2: Semi-supervised Linear Discriminant Clustering Algorithm

```
    Input: A $N \times M$ document-term matrix $\mathbf{X}$, the number of
           clusters $K$ and the seeds $S = \{S_1, \ldots, S_K\}$. Without loss
           of generality, $S_k$ represents the document seeds for topic
           $k$ ($k = 1, \ldots, K$).
    Output: A $N \times K$ clustering matrix $\mathbf{U}$.
 1  begin
 2      $\mathbf{Q} \longleftarrow$ constrained-PLSA($\mathbf{X}$, $K$, $S$) ;
 3      $\mathbf{X} \longleftarrow$ PCA($\mathbf{X}$) ;
 4      Construct matrix $\mathbf{D}$ using Equation (10) ;
 5      Construct matrix $\mathbf{B}^{soft}$ using Equation (11) ;
 6      Compute $\tilde{\mathbf{S}}_b$ using Equation (12) ;
 7      Compute $\tilde{\mathbf{S}}_w$ using Equation (13) ;
 8      Solve generalized eigenvalue problem $\tilde{\mathbf{S}}_b\mathbf{a} = \lambda(\tilde{\mathbf{S}}_w + \mu\mathbf{I})\mathbf{a}$
        and use the first $K-1$ eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_{K-1}$ as
        columns to compose a projection matrix $\mathbf{A}$, where their
        corresponding eigenvalues are in descending order. ;
 9      $\tilde{\mathbf{X}} \longleftarrow \mathbf{XA}$ ;
10      $\mathbf{U} \longleftarrow$ constrained-KMeans($\tilde{\mathbf{X}}$, $S$) ;
11      return $U$
12  end
```

IV. Experiments

This paper uses three data sets to assess system performance. Besides, several semi-supervised learning algorithms are applied to the data sets to compare with the proposed approach. The experiments focus on semi-supervised learning performance, explaining why only a few labeled examples are used in the experiments. This paper randomly selected examples as the labeled ones and the rest of examples are unlabeled examples. To further evaluate the impact of the number of labeled examples on system performance, different percentages of labeled examples are used in the experiments. Each evaluation runs ten times. We present the experimental results by using the average with two standard deviations. Additionally, some methods in the experiments use PCA to reduce dimensions, and the number of dimensions should be given in PCA operation. Instead of determining the number of dimensions directly, we use the percentage of variance explained as the criterion, since the eigenvalues obtained from PCA are equal to the variance explained by each of the principal components in decreasing order of importance. The experiments retain the number of dimensions which can account for 90% of the total variance.

A. Data Corpora

This paper uses three data sets in the experiments. The 20 Newsgroups and Reuters-21578 are popular corpora, which

TABLE II

CiteULike Corpus

	Graphics	Databases	Programming Languages
Number of papers	741	1,289	1,364
Number of terms in abstractions	65,372	115,346	110,184
Number of tags	2,013	3,126	4,983

TABLE III

Ten Largest Classes in the Reuters-21578 Collection

class	number of documents	class	number of documents
earn	3,753	trade	449
acquisitions	2,131	interest	389
money-fx	601	ship	276
grain	528	wheat	264
crude	510	corn	207

are commonly used in text analysis experiments. Besides, this paper employs the academic paper information collected from CiteULike[1] to evaluate system performance.

1) The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. Some of the newsgroups are very closely related to each other, while others are highly unrelated. It has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.
2) CiteULike is a social bookmarking web site and is aimed to promote and to develop the sharing of scientific references amongst researchers. Scientists can annotate their interested academic papers with tags and share the information with the other people. CiteUlike fuses together two separated categories of software: the new Web 2.0 breed of social bookmarking services and traditional bibliographic management software. While web bookmarks are simple URLs, citations are a bit more complex and include meta-data like journal names, authors, and page numbers. However, meta-data information does not include paper category information, which is necessary for this paper to assess performance. This paper assigns papers to communities according to their venues, using the classification system adopted by Microsoft's academic search service [2] that provides the ranking of publications in different fields. For instance, graphics field includes TOG (ACM Transactions on Graphics), and CGA (IEEE Computer Graphics and Applications). A paper published in the TOG would be classified as graphics field. Above paper classification mechanism is also used by Shi *et al.* [43]. Obviously, some publications may belong to more than one field, explaining why this paper only focuses on the fields that are highly unrelated. This paper focuses on computer science domain and collects 3,394 articles from three fields. The paper's full text is unavailable in CiteULike, so the paper's abstract and tags annotated by users are considered as paper content. This corpus can be downloaded from http://islab.cis.nctu.edu.tw/download/. Table II summaries the information of the data set.
3) Reuters-21578 is one of the most widely used test collections for text classification research. The data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text categorization system. The ten largest classes in the Reuters-21578 collection are used in the experiments, as it has been used by many researchers in recent years. Table III presents the number of documents in the ten largest classes.

In the preprocess stage, the stop words are removed from these data sets, since they fail to provide sufficient information for the analysis task. Additionally, punctuation marks are removed and all English letters are converted into lower case. Finally, stemming process is applied to the words.

B. Evaluation Measurements

For each class, the correctness of a classification can be assessed by calculating the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or were not recognized as class examples (false negatives) [44]. Equation (19) shows the definition of precision, recall and F_1 score, where TP represents the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives. Meanwhile, numerous classification tasks employed in the experiments are multiclass problems, so the evaluation should consider the prediction results for every class. Macro-average F_1, which is the average of the F_1 scores of all the classes, is used to assess system performance. Equation (20) shows the definition of the macro-average F_1 score, where K denotes the number of classes and F_{1i} represents the F_1 score of the ith class

$$\begin{aligned}\text{Precision} &= \frac{TP}{TP+FP}\\ \text{Recall} &= \frac{TP}{TP+FN}\\ F_1 &= \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \qquad (19)\\ \text{Macro-average } F_1 &= \frac{\sum_{i=1}^{K} F_{1i}}{K}. \qquad (20)\end{aligned}$$

C. Comparison Methods

1) *Graph-Based Semi-Supervised Learning*

Graph-based approach has been widely used in

[1] CiteULike: http://www.citeulike.org/

[2] Microsoft Academic Search: http://academic.research.microsoft.com

TABLE IV
EXPERIMENTAL RESULTS ON COMPUTER NEWSGROUPS (FIVE NEWSGROUPS)

	Semi-LDC	Graph-based	TSVM	C-KMeans	PCA + C-KMeans	Label Propagation
1%	0.5321 ± 0.0791	0.2842 ± 0.0561	0.3622 ± 0.0559	0.2838 ± 0.0582	0.3015 ± 0.0517	0.3901 ± 0.0574
2%	0.6467 ± 0.0505	0.3800 ± 0.0756	0.4789 ± 0.0736	0.3084 ± 0.0789	0.3207 ± 0.0742	0.4434 ± 0.0461
3%	0.6710 ± 0.0402	0.4456 ± 0.0314	0.5167 ± 0.0514	0.3096 ± 0.0489	0.3300 ± 0.0366	0.4721 ± 0.0297
4%	0.7091 ± 0.0189	0.4710 ± 0.0466	0.5716 ± 0.0337	0.3257 ± 0.0395	0.3296 ± 0.0551	0.4999 ± 0.0196
5%	0.7150 ± 0.0184	0.4862 ± 0.0513	0.6132 ± 0.0411	0.3388 ± 0.0276	0.3357 ± 0.0395	0.5219 ± 0.0353

TABLE V
EXPERIMENTAL RESULTS ON TALK NEWSGROUPS (FOUR NEWSGROUPS)

	Semi-LDC	Graph-based	TSVM	C-KMeans	PCA + C-KMeans	Label Propagation
1%	0.6402 ± 0.1107	0.3051 ± 0.1617	0.4794 ± 0.0549	0.3734 ± 0.0854	0.3757 ± 0.0203	0.5012 ± 0.0439
2%	0.7328 ± 0.0611	0.5118 ± 0.0629	0.5546 ± 0.0843	0.3992 ± 0.0745	0.3382 ± 0.1075	0.5680 ± 0.0284
3%	0.7660 ± 0.0382	0.5317 ± 0.0633	0.6197 ± 0.0580	0.3944 ± 0.0113	0.3840 ± 0.0152	0.6172 ± 0.0410
4%	0.7840 ± 0.0201	0.6061 ± 0.0380	0.6249 ± 0.0584	0.3952 ± 0.0114	0.3619 ± 0.0646	0.6357 ± 0.0225
5%	0.7928 ± 0.0157	0.6080 ± 0.0541	0.6745 ± 0.0211	0.3972 ± 0.0066	0.3596 ± 0.1011	0.6442 ± 0.0402

semi-supervised learning methods, so this paper uses graph-based method in the experiments to compare with the proposed method. The one used in the experiments is proposed by Goldberg *et al.* [45]. Similar to the other graph-based semi-supervised learning approaches, their approach uses a graph to represent labeled and unlabeled data. Each document is a node in the graph, and each node is connected with an observed node called dongle. The edge weight between a labeled document and its dongle is a large number M, while the weight between an unlabeled document and its dongle is 1. Each unlabeled document x_i connects to k nearest labeled documents and k' nearest unlabeled documents. Different weight coefficients are given in the above two cases. Then, the original problem can be transformed into an optimization with constraints problem. Goldberg *et al.* [45] used support vector regression (SVR) in their proposed graph-based semi-supervised learning approach to perform an initial prediction. However, our experimental results show that graph-based semi-supervised with support vector machines (SVM) outperforms graph-based semi-supervised with SVR, explaining why this paper employs graph-based semi-supervised with SVM. This paper conducts graph-based semi-supervised learning experiments using libsvm [46] package with radial basis function (RBF) kernel function. Moreover, the value of k' is 5, and k is the number of seeds divided by 10 in the experiments.

2) *TSVM*
TSVM is an extension of standard SVM with unlabeled data. This paper employs SVMlight [11] with the RBF kernel function to conduct experiments. For multiclass classification, one-against-all approach is used in the experiments. The TSVM is a state-of-the-art method in semi-supervised learning, explaining why TSVM is used in the experiments.

3) *Constrained-KMeans: (Abbreviated as C-KMeans)*
Basu *et al.* [15] proposed two semi-supervised variants of KMeans clustering that use initial labeled data for seeding. These two algorithms are Seeded-KMeans and constrained-KMeans. Their experimental results showed that constrained-KMeans outperforms Seeded-KMeans. Meanwhile, constrained-KMeans also outperforms COP-KMeans [14]. The proposed algorithm uses constrained-KMeans to cluster the data points in the reduced feature space, so this paper conducts experiments with constrained-KMeans in the original input space to see whether the proposed method can benefit from the proposed dimensionality reduction steps.

4) *PCA + constrained-KMeans: (Abbreviated as PCA + C-KMeans)*
As described above, document collection is generally a high-dimensional data set, and many machine learning methods may benefit from dimensionality reduction process. As a result, this method uses PCA to reduce dimensions and then uses constrained-KMeans to cluster documents in the new space.

5) *Label Propagation*
Label propagation is a commonly used technique in many graph based semi-supervised learning algorithms [47], [48], so this paper uses label propagation algorithm in the experiments and compares with the proposed algorithm. The label information is propagated from labeled examples to unlabeled ones iteratively. The experiments use the algorithm proposed by Nie *et al.* [36].

D. Semi-supervised Learning Experiments

The first data set is 20 Newsgroups data set. This paper focuses on the newsgroups which are highly related. Two combinations of newsgroups are used in the experiments, including computer subject and talk subject. The purposes of the experiments focus on two issues. The first one focuses on whether these methods can function well on multiclass problems. Some algorithms are designed for binary class classification or clustering, so these experiments can be used to evaluate whether these methods can be extended to multiclass problems. For instance, TSVM is a binary classifier, and this

TABLE VI

EXPERIMENTAL RESULTS ON CITEULIKE

	Semi-LDC	Graph-based	TSVM	C-KMeans	PCA + C-KMeans	Label Propagation
1%	0.8056 ± 0.0822	0.4944 ± 0.2208	0.8460 ± 0.0226	0.4676 ± 0.1921	0.5447 ± 0.0767	0.6357 ± 0.0681
2%	0.8590 ± 0.0569	0.6220 ± 0.0830	0.8701 ± 0.0231	0.5320 ± 0.1623	0.5653 ± 0.0073	0.6959 ± 0.0472
3%	0.8845 ± 0.0449	0.6953 ± 0.0531	0.8831 ± 0.0116	0.5022 ± 0.1143	0.5695 ± 0.0068	0.7205 ± 0.0467
4%	0.8990 ± 0.0135	0.7146 ± 0.0692	0.8890 ± 0.0159	0.4795 ± 0.0067	0.5737 ± 0.0059	0.7379 ± 0.0298
5%	0.8991 ± 0.0227	0.7449 ± 0.0662	0.8961 ± 0.0126	0.5111 ± 0.1616	0.5764 ± 0.0071	0.7465 ± 0.0303

TABLE VII

EXPERIMENTAL RESULTS ON REUTERS-21578

	Semi-LDC	Graph-based	TSVM	C-KMeans	PCA + C-KMeans	Label Propagation
1%	0.5460 ± 0.0735	0.3082 ± 0.0352	0.4870 ± 0.0561	0.3035 ± 0.1166	0.2398 ± 0.0794	0.4939 ± 0.0331
2%	0.5788 ± 0.0842	0.3854 ± 0.0486	0.5640 ± 0.0429	0.3472 ± 0.1262	0.2584 ± 0.0786	0.5088 ± 0.0184
3%	0.5975 ± 0.0449	0.3980 ± 0.0178	0.5769 ± 0.0318	0.3791 ± 0.1544	0.2385 ± 0.0362	0.5428 ± 0.0322
4%	0.6127 ± 0.0144	0.4307 ± 0.0322	0.5688 ± 0.0497	0.4188 ± 0.2115	0.2703 ± 0.0690	0.5607 ± 0.0219
5%	0.6243 ± 0.0369	0.4426 ± 0.0183	0.5907 ± 0.0367	0.4086 ± 0.1182	0.2931 ± 0.0985	0.5570 ± 0.0372

paper employs one-against-all scheme for multiclass problems. The second one is to evaluate whether these methods can function properly when the boundaries among clusters are unclear. Tables IV and V present the experimental results on computer subject and talk subject, respectively.

The second data set is CiteULike, including three categories. The content of each document includes paper abstract and the tags annotated by researchers. The abstract is similar to the summary of a paper; while tags can be thought of as the keywords of a paper. Thus, abstract and tags can be viewed as condensed information of a paper. The purpose of this experiment focuses on whether these methods can perform well on the data set in which only condensed information is available. Table VI summaries the experimental results.

The final data set is Reuters-21578, including the ten largest classes in the collection. It is apparent that this is a multiclass and imbalanced data set as shown in Table III. As a result, the purpose of this experiment focuses on whether the methods can perform well on the data set with imbalanced cluster sizes. Experimental results are listed in Table VII.

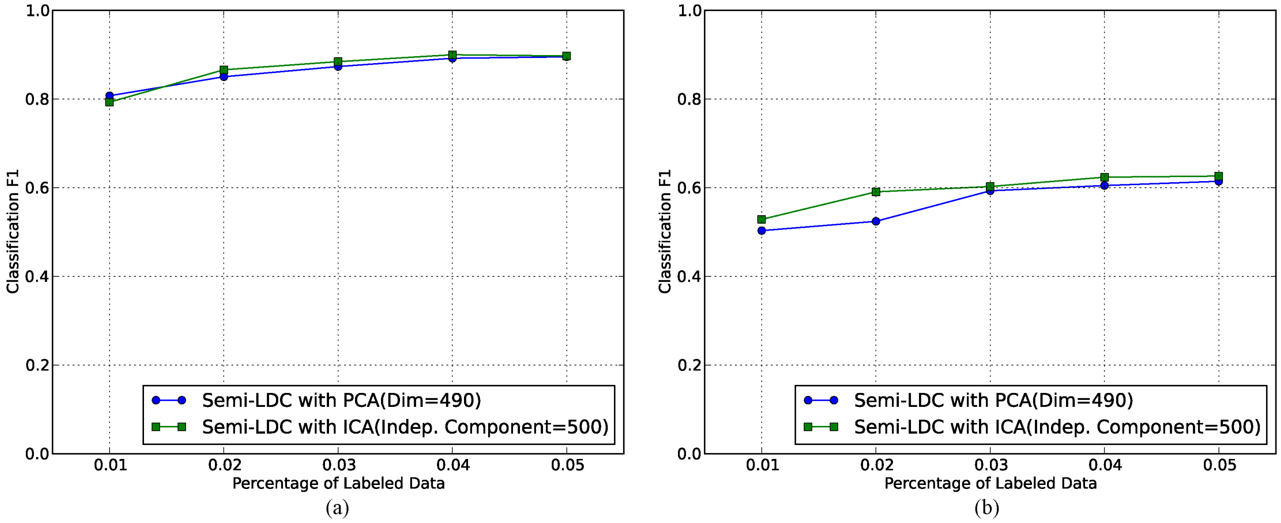
E. Discussion

This paper conducts experiments on three data sets, and the experimental results indicate that the proposed method generally outperforms the other methods. Even though the cluster boundaries are unclear, the proposed method can perform well. We further analyze and discuss the experiments in this section. Although Torkkola [41] has shown that the combination of PCA and LDA can improve classification performance, and the experimental results presented above conform to his conclusion. We further analyze whether clustering can benefit from the dimensionality reduction performed by PCA. Tables VI and VII present the experimental results. The combination with PCA + C-KMeans outperforms C-KMeans on CiteULike data set. However, the experiments show different results on Reuters-21578 data set. Although PCA can preserve as much of the variance in the reduced space as possible when performing dimensionality reduction, clustering in the reduced space may fail to perform well. The main reason is that the PCA does not consider classification or clustering criteria when reducing dimensions.

As PCA considers second order moments only, it lacks information on higher order statistics. Independent component analysis (ICA) is a data analysis technique accounting for higher order statistics [49]. ICA has great potential in applications such as the separation of sound signals, in telecommunication or in biomedical engineering [50], [51]. This paper applies ICA to text analysis problem. In the experiments, it is hard to estimate the number of independent components for ICA, so we use the number of principal components used in PCA as the criterion. Fig. 1(a) and (b) presents experimental results on CiteULike and Reuters-21578 data sets, respectively. The experimental results indicate that the two methods can have almost identical performance results. As a result, ICA can also be used in the proposed algorithm.

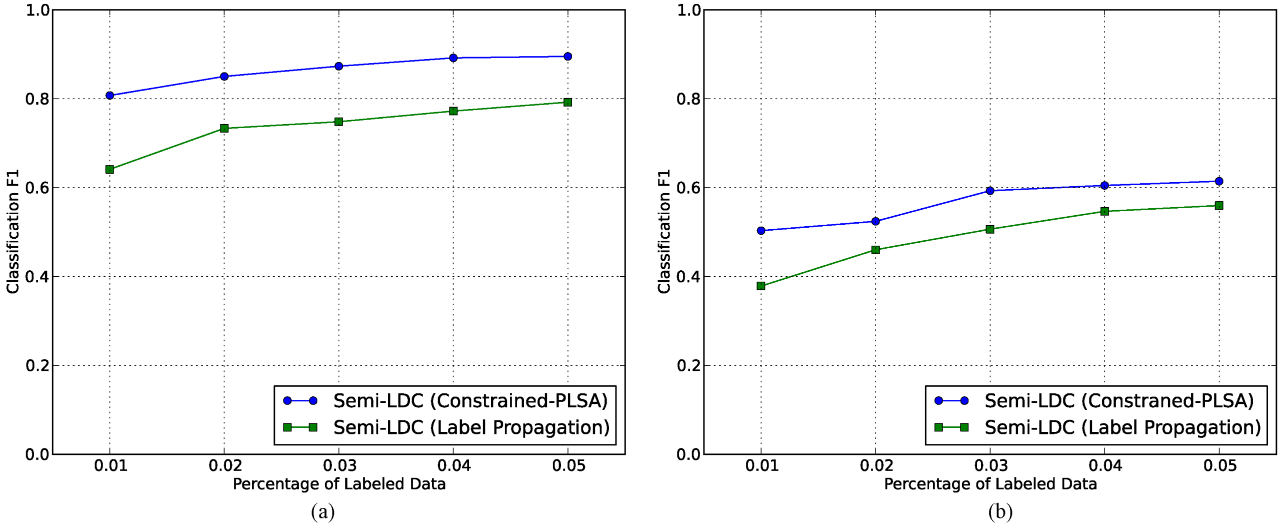
Practically, the constrained-PLSA used in the proposed algorithm can be replaced by the other algorithms. This paper focuses on text analysis problem and PLSA model is derived from text analysis problem, making it appropriate to use constrained-PLSA to obtain soft labels. Nie *et al.* [36] devised a label propagation algorithm to obtain soft labels, so we conduct experiments using Semi-LDC with different soft label computation methods for performance comparison. Fig. 2(a) and (b) summarizes the experimental results on CiteULike and Reuters-21578 data sets, respectively. The Semi-LDC with constrained-PLSA outperforms Semi-LDC with label propagation on the two data sets. The experimental results indicate that constrained-PLSA is a better choice than label propagation in text analysis problem. Practically, the proposed method can use the other soft label computation methods according to different application requirements.

This paper further analyzes soft label impact on classification performance with different percentages of labeled examples. We use LDA as the comparison method, meaning that the dimensionality reduction performed by LDA only uses hard labels of labeled examples to find the projection vector.



(a) (b)

Fig. 1. Semi-LDC Classification Performance with Different Dimensionality Reduction Methods. (a) Semi-LDC Classification Performance on CiteULike Data Set. (b) Semi-LDC Classification Performance on Reuters-21578 Data Set.

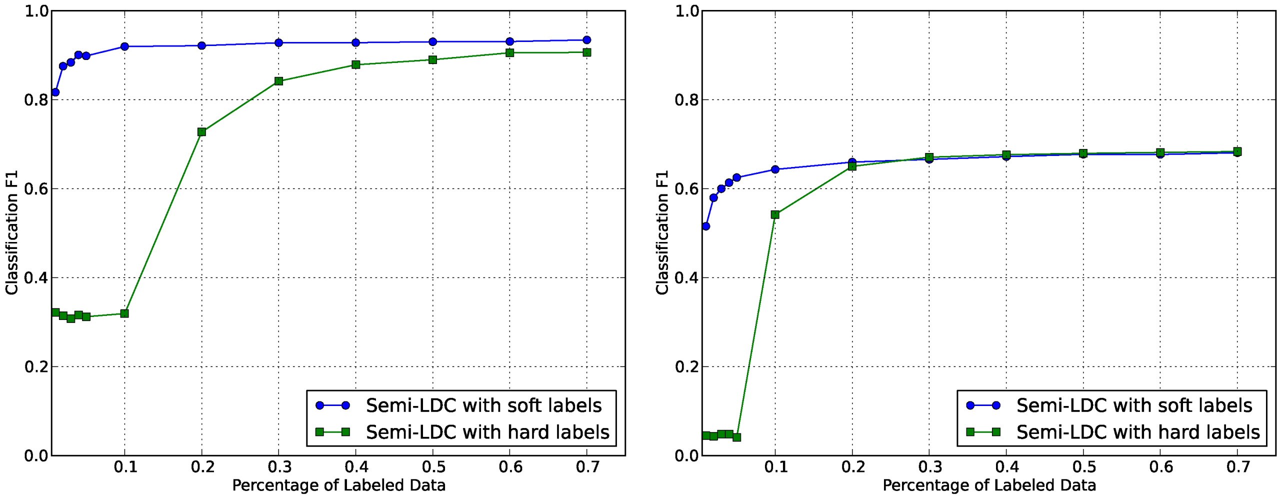


(a) (b)

Fig. 2. Semi-LDC Classification Performance with Different Soft Label Computation Methods. (a) Classification Performance on CiteULike Data Set. (b) Classification Performance on Reuters-21578 Data Set.

We call the comparison approach as Semi-LDC with hard labels, since LDA can only use hard labels. Fig. 3(a) and (b) summarizes the experimental results on CiteULike and Reuters-21578 data sets, respectively. The two experiments exhibit similar results, namely, the proposed method with soft labels can perform very well when only a few labeled examples are available. The experimental results conform to the research results presented in semi-supervised learning, that is, the semi-supervised learning can use unlabeled data with model assumptions and available labeled examples to improve performance in certain problems. The experimental results also indicate that hard label approach and soft label approach can achieve almost identical performance when the number of labeled examples is sufficient for LDA to find an appropriate projection vector.

The proposed method is a framework, which involves several algorithms, including PCA, constrained-PLSA, soft LDA and constrained-KMeans. As described above, the PCA in the proposed method is mainly for dimensionality reduction so as to make the within-class scatter matrix nonsingular before soft LDA process, so the PCA can be viewed as a preprocess step to make the whole framework robust. The previous experiments also indicate that the PCA can be replaced by ICA. The goal of the proposed framework is to use dimensionality reduction technique to find an appropriate feature space, so that the clustering can perform well in the new feature space. Meanwhile, the soft LDA relies on constrained-PLSA to obtain the soft labels of unlabeled examples, so constrained-PLSA is a critical process in the framework. Practically, different application domains may require different soft label estimation algorithms. The previous experiments also indicate that although label propagation can replace constrained-PLSA, label propagation method degrades clustering performance in text analysis domain. The main reason is that constrained-PLSA is derived from PLSA, explaining why it outperforms label propagation in text analysis problems. Additionally, the soft LDA aims at finding a feature representation by which the within-class distance is minimized and the between-class



(a)

(b)

Fig. 3. Semi-LDC Classification Performance with Soft Labels and Hard Labels. (a) Classification Performance on CiteULike Data Set. (b) Classification Performance on Reuters-21578 Data Set.

distance is maximized. In other words, the goal is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. When the dimensionality reduction performs well, clustering in the new feature space becomes an easy task. The framework uses constrained-KMeans, which is derived from K-means, but other clustering algorithms can function properly.

V. Conclusion

This paper devises a semi-supervised learning algorithm, which connects K-means and LDA seamlessly, to consider clustering and dimensionality reduction simultaneously. The K-means is an unsupervised learning method, while LDA is a supervised learning method. Central to the proposed method is using soft LDA with soft labels of unlabeled examples to find an appropriate feature space, in which we argue text classification performance can be improved effectively in the reduced space. The proposed method can be viewed as a framework, since different soft label estimation methods can be used in the framework according to application requirements. The experimental results indicate that the proposed method can generally outperform several alternative methods. This paper also demonstrates that the proposed method can benefit from soft label representation particularly when only a few labeled examples are available.

Although the proposed framework is a semi-supervised learning method, the implementation assumes that a complete data set is given in advance, and the learning process is carried out in one batch. In many application domains, data is presented as a data stream, so we often confront difficult situation where a complete set of data set is not given in advance. The future work is to incorporate online learning in the proposed framework. Moreover, latent Dirichlet allocation [52] has been successfully applied to text analysis problem, but it is an unsupervised learning. Another research direction is to investigate how to devise an effective semi-supervised latent Dirichlet allocation algorithm to replace the proposed constrained-PLSA in the framework.

References

[1] L. Wang, L. Bo, and L. Jiao, "A modified k-means clustering with a density-sensitive distance metric," in *Proc. 1st Int. Conf. Rough Sets Knowl. Technol.*, 2006, pp. 544–551.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[3] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Mach. Learn.*, vol. 39, pp. 103–134, May 2000.

[4] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1530, 2005.

[5] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 1135–1142.

[6] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr., and T. M. Mitchell, "Coupled semi-supervised learning for information extraction," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 101–110.

[7] C.-L. Liu, T.-H. Chang, and H.-H. Li, "Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans," *Fuzzy Sets Syst.*, vol. 221, pp. 48–64, Jun. 2013.

[8] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1-2, pp. 177–196, 2001.

[9] T. Hofmann, J. Puzicha, and M. I. Jordan, "Learning from dyadic data," in *Proc. 1998 Conf. Adv. Neural Inf. Process. Syst. II*, 1999, pp. 466–472.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[11] T. Joachims, *Making Large-Scale Support Vector Machine Learning Practical*. Cambridge, MA, USA: MIT Press, 1999, pp. 169–184.

[12] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 19–26.

[13] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in *Proc. 1st Workshop Graph Based Methods Natural Lang. Process.*, 2006, pp. 45–52.

[14] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained K-means clustering with background knowledge," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 577–584.

[15] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *Proc. 19th ICML*, 2002, pp. 27–34.

[16] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proc. 10th ACM SIGKDD Int. Conf. KDD*, 2004, pp. 59–68.

[17] F. Nie, D. Xu, and X. Li, "Initialization independent clustering with actively self-training method," *Trans. Syst., Man, Cybern. B*, vol. 42, no. 1, pp. 17–27, Feb. 2012.

[18] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 27, no. 5, pp. 787–795, Sep. 1997.

[19] A. Bouchachia and W. Pedrycz, "A semi-supervised clustering algorithm for data exploration," in *Proc. 10th Int. Fuzzy Syst. Assoc. World Congr. Conf. Fuzzy Sets Syst.*, 2003, pp. 328–337.

[20] S. Miyamoto, M. Yamazaki, and W. Hashimoto, "Fuzzy semi-supervised clustering with target clusters using different additional terms," in *Proc. IEEE Int. Conf. Granular Comput.*, 2009, pp. 444–449.

[21] A. Bouchachia and W. Pedrycz, "Data clustering with partial supervision," *Data Min. Knowl. Discov.*, vol. 12, pp. 47–78, Jan. 2006.

[22] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic, 1981.

[23] I. Fodor, "A survey of dimension reduction techniques," Center Appl. Sci. Comput., Lawrence Livermore Nat. Laboratory, Livermore, CA, USA, Tech. Rep. UCRL-ID-148494, 2002.

[24] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2008, pp. 444–452.

[25] N. Pham and R. Pagh, "A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2012, pp. 877–885.

[26] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. NIPS*, 2004, pp. 513–520.

[27] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Jun. 2009.

[28] F. Wang, "Semisupervised metric learning by maximizing constraint margin," *Trans. Syst. Man, Cybern. B*, vol. 41, no. 4, pp. 931–939, Aug. 2011.

[29] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, Tech. Rep., 2006.

[30] A. Lehrmann, M. Huber, A. Polatkan, A. Pritzkau, and K. Nieselt, "Visualizing dimensionality reduction of systems biology data," *Data Min. Knowl. Discovery*, vol. 27, no. 1, pp. 146–165, 2012.

[31] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2001, pp. 245–250.

[32] J. B. Tenenbaum, V. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[33] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[34] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, Jun. 2003.

[35] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 225–232.

[36] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2615–2627, 2009.

[37] J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu, "Efficient model selection for regularized linear discriminant analysis," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage.*, 2006, pp. 532–539.

[38] J. Peng, P. Zhang, and N. Riedel, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 2008.

[39] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized discriminant analysis and its application in microarrays," *Biostatist.*, vol. 8, no. 1, pp. 86–100, 2007.

[40] J. Peng, P. Zhang, and N. Riedel, "Discriminant learning analysis," *IEEE Trans. Syst., Man, Cybern. B*, vol. 38, no. 6, pp. 1614–1625, Dec. 2008.

[41] K. Torkkola, "Discriminative features for text document classification," *Pattern Anal. Appl.*, vol. 6, no. 4, pp. 301–308, Feb. 2003.

[42] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[43] X. Shi, B. L. Tseng, and L. A. Adamic, "Information diffusion in computer science citation networks," in *Proc. ICWSM*, 2009.

[44] M. Sokolova and L. Guy, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, pp. 427–437, Jul. 2009.

[45] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in *Proc. TextGraphs: 1st Workshop Graph Based Methods for Natural Lang. Process.*, 2006, pp. 45–52.

[46] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[47] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, 2003, pp. 912–919.

[48] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems 16*. S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA, USA: MIT Press, 2004.

[49] C. Bugli and P. Lambert, "Comparison between principal component analysis and independent component analysis in electroencephalograms modelling," *Biometrical J.*, vol. 48, no. 5, pp. 1–16, 2006.

[50] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, May 2000.

[51] P. Akhlaghi, A. R. Kashanipour, and K. Salehshoor, "Complex dynamical system fault diagnosis based on multiple ANFIS using independent component," in *Proc. 16th MED*, 2008, pp. 1798–1803.

[52] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

Chien-Liang Liu (M'13) received the M.S. and Ph.D. degrees in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2000 and 2005, respectively.

He is currently an Engineer in the Computational Intelligence Technology Center, Industrial Technology Research Institute, Hsinchu. His current research interests include machine learning, natural language processing, information retrieval, and data mining.

Wen-Hoar Hsaio received the B.S. degree from the Department of Computer Science and Information Engineering, Chung Cheng Institute of Technology, National Defense University, Taiwan, in 1980, and the M.S. degree from the Department of Computer Science from National Chiao Tung University, Hsinchu, Taiwan, in 1996.

He is currently pursuing the Ph.D. degree at the Department of Computer Science, National Chiao Tung University. His current research interests include information retrieval, web mining, and machine learning.

Chia-Hoang Lee received the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, in 1983.

He was a faculty member at the University of Maryland and Purdue University, West Lafayette, IN, USA. Currently, he is a Professor in the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan. His current research interests include artificial intelligence, human machine interface systems, natural language processing, and opinion mining.

Fu-Sheng Gou received the M.S. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2012.

He is currently a Software Engineer with the Chung Shan Institute of Science and Technology, Taoyuan, Taiwan. His current research interests include machine learning, natural language processing, and pattern recognition.