Subspace Clustering of Categorical and Numerical Data With an Unknown Number of Clusters

Hong Jia and Yiu-Ming Cheung, *Senior Member, IEEE*

***Abstract*— In clustering analysis, data attributes may have different contributions to the detection of various clusters. To solve this problem, the subspace clustering technique has been developed, which aims at grouping the data objects into clusters based on the subsets of attributes rather than the entire data space. However, the most existing subspace clustering methods are only applicable to either numerical or categorical data, but not both. This paper, therefore, studies the soft subspace clustering of data with both of the numerical and categorical attributes (also simply called *mixed data* for short). Specifically, an attribute-weighted clustering model based on the definition of object-cluster similarity is presented. Accordingly, a unified weighting scheme for the numerical and categorical attributes is proposed, which quantifies the attribute-to-cluster contribution by taking into account both of intercluster difference and intracluster similarity. Moreover, a rival penalized competitive learning mechanism is further introduced into the proposed soft subspace clustering algorithm so that the subspace cluster structure as well as the most appropriate number of clusters can be learned simultaneously in a single learning paradigm. In addition, an initialization-oriented method is also presented, which can effectively improve the stability and accuracy of *k*-means-type clustering methods on numerical, categorical, and mixed data. The experimental results on different benchmark data sets show the efficacy of the proposed approach.**

***Index Terms*— Attribute weight, categorical-and-numerical data, initialization method, number of clusters, soft subspace clustering.**

I. Introduction

With the ability to extract potentially useful information from databases in unsupervised learning environment, clustering is regarded as an important technique in data mining field and has been widely applied to a variety of scientific areas, such as pattern recognition, signal processing, bioinformatics, and so on.

Manuscript received June 11, 2016; revised April 4, 2017; accepted July 9, 2017. Date of publication August 3, 2017; date of current version July 18, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61672444 and Grant 61272366, in part by the Natural Science Foundation of SZU under Grant 2017078, in part by the Faculty Research Grant of Hong Kong Baptist University under Project FRG2/16-17/051, in part by the KTO Grant of HKBU under Project MPCF-004-2017/18, and in part by SZSTI under Grant JCYJ20160531194006833. *(Corresponding author: Yiu-Ming Cheung.)*

H. Jia is with the College of Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: hongjia1102@szu.edu.cn).

Y.-M. Cheung is with the Department of Computer Science, Institute of Research and Continuing Education, Hong Kong Baptist University, Hong Kong, and also with the United International College, Beijing Normal University–Hong Kong Baptist University, Zhuhai 519085, China (e-mail: ymc@comp.hkbu.edu.hk).

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/TNNLS.2017.2728138

Generally, similar with the classification in supervised learning [1]–[3], clustering analysis requires data objects to be defined with a set of relevant attributes that represent the properties of the objects that are useful to the learning task. With the growing applications of clustering, two main problems have been raised. The first one is that the most existing clustering algorithms assume the data attributes are numerically valued [4]–[7]. However, as extensive data have been collected from medical, business, industry, and social domains, we usually have to deal with another kind of attributes that are nominal valued, such as gender, shape, hobby, and type of disease [8]. In the literature, this kind of attributes is referred to as categorical attributes or categorical features. Hereinafter, data represented with both of the numerical and categorical attributes are called *mixed data*. To handle this kind of data sets, the most straightforward way is to discretize numerical attributes or encode categorical attribute values as numerical integers such that popular clustering approaches can be applied. Nevertheless, transforming the attribute type will ignore the similarity information embedded in the original attribute values and, thus, cannot faithfully reveal the structure of the data sets [8], [9]. Under the circumstances, researchers have tried to investigate new criteria for clustering analysis of mixed data and different methods have been proposed in [8] and [10]–[15]. However, as all these methods treat data attributes equally during clustering process, they will face the other problem that different attributes may have different contributions to the detection of various clusters from the practical viewpoint. For example, if we wish to classify patients by types of diseases, gender information will be a key attribute for some special diseases but useless for the others. In general, subspace clustering is a special technique for this problem solving [16]. This type of clustering methods aims at grouping the data objects into clusters based on the subspaces of data rather than the entire data space. Although in a sense dimensionality reduction can reach similar objective, in subspace clustering, different clusters are usually learned from different attribute subsets. In practice, if the data dimensionality is extremely large, dimensionality reduction techniques [17]–[21] can also be utilized as preprocessing of subspace clustering to improve the learning results. According to the different ways with which the feature subspaces of clusters are determined, the subspace clustering methods can be divided into two categories. The first one is hard subspace clustering, which attempts to find the exact attribute subsets for different clusters. Typical examples include [22]–[29]. The other category is soft subspace clustering, which learns the

data clusters from the entire feature space, but assigns different weights to each attribute during the clustering process to control its contribution to the discovering of different clusters. Representative methods are proposed in [30]–[35]. However, all the above-mentioned subspace clustering methods are only applicable to numerical data only, which limits their applications in the domain where categorical attributes exist. By contrast, some other researchers have concentrated on the subspace clustering of data with purely categorical attributes. Typical work includes the hard subspace clustering approaches for categorical data proposed in [36]–[39] and the soft methods proposed in [40]–[42]. Nevertheless, to the best of our knowledge, the study on subspace clustering for mixed data is scarce thus far.

In addition, the most existing subspace clustering methods need to prespecify the true number of clusters. Unfortunately, this importance information is not always available from the practical viewpoint. Therefore, how to conduct clustering analysis without knowing the true number of clusters is also a significant work in clustering area [43]. To address this issue, researchers have presented a variety of methods that can estimate the number of clusters for purely numerical or categorical data [44]–[48]. Nevertheless, how to automatically select the number of clusters for mixed data during clustering process has been seldom studied. Moreover, this problem can be even more challenging in subspace clustering as the optimal feature subspace and the optimal number of clusters are interrelated, i.e., different clustering results might be obtained on different feature subspaces [49].

This paper will therefore focus on soft subspace clustering of mixed data without knowing the number of clusters. Specifically, within the learning framework of object-cluster similarity-based clustering method, an attribute-weighted learning model is presented and a new attribute weighting scheme is proposed accordingly. These attribute weights quantify the contribution of different categorical and numerical attributes to the detection of various clusters. Specifically, the weight of an attribute to a particular cluster is determined by two factors. The first one is intercluster difference, which measures the ability of the attribute in distinguishing this cluster from the other ones. The other factor is intracluster similarity, which evaluates whether the cluster along this attribute has a compact structure or not. This weighting scheme in nature is consistent with the basic concept of clustering analysis. Meanwhile, unified criteria are defined to quantify the intercluster difference and intracluster similarity for numerical and categorical attributes so that the proposed method can be directly applied to data sets with categorical, numerical, or mixed attributes. Moreover, a rival penalized competitive learning mechanism is further introduced into the proposed soft subspace clustering algorithm to enable it to learn the number of clusters automatically during clustering process. Subsequently, the subspace cluster structure as well as the most appropriate number of clusters can be learned simultaneously from a single learning paradigm. In addition, as the clustering results led by random initialization in k-means-type methods are often unpredictable and a better initialization method for mixed data clustering has not been studied in the literature, we further present an initialization-oriented method that can lead much more stable clustering results and improve the performance of k-means-type methods on numerical, categorical, and mixed data. The effectiveness of the proposed methods has been experimentally investigated on different benchmark data sets in comparison with the existing counterparts. The main contributions of this paper can be summarized as follows.

1) An attribute-weighted clustering model based on object-cluster similarity is presented for soft subspace clustering on data with numerical and categorical attributes.
2) A new attribute weighting scheme is proposed for mixed data, which adopts a unified criterion to quantify the contribution of each categorical or numerical attribute to the detection of every cluster. This weighting scheme is the first one that simultaneously considers the cluster-distinguishing ability and intracluster compactness of different attributes based on probability distribution model.
3) The rival penalized competitive learning mechanism is introduced into the soft subspace clustering of mixed data so that the number of clusters can be automatically determined.
4) This is the first attempt to study the initialization problem of clustering algorithm on mixed data type. Accordingly, an initialization-oriented method, which is applicable to numerical, categorical, and mixed data, is proposed. This method can obviously improve the stability and accuracy of k-means-type clustering methods on different types of data sets.

The rest of this paper is organized as follows. In Section II, we will overview some related work on mixed data analysis and subspace clustering. Section III proposes a new attribute-weighted clustering method for soft subspace clustering of mixed data. Section IV further introduces a competitive learning model with penalization mechanism to learn the number of clusters. Moreover, Section V presents a new initialization method. Then, Section VI shows the experimental results on real data sets. Finally, we draw a conclusion in Section VII.

II. Overview of Related Work

This section reviews the related work on: 1) clustering analysis of mixed data and 2) subspace clustering.

A. Clustering Analysis of Mixed Data

The existing methods that are applicable to clustering analysis of data with both of the numerical and categorical attributes can be grouped into three categories. In the first category, different techniques have been presented to transform categorical attribute values into numerical ones such that a wide variety of algorithms for numerical data can be directly applied. For example, the dummy variable coding method [50] transforms each value of categorical attribute A_j with m_j different values into a m_j-dimensional vector, where a single 1 in a particular position represents the attribute value and all the rest dimensions are 0. In [51], a supervised method, namely, density-based logistic regression framework,

has been proposed, which replaces each categorical value with the histogram of class labels associated with it. Moreover, Zhang *et al*. [52] proposed a multiple transitive distance learning and embedding method to learn the pairwise dissimilarity among the categorical attribute values. Subsequently, each categorical symbol can be endowed with a numerical representation using manifold learning techniques.

In the second category, the algorithms are essentially designed for purely categorical data, although they have been applied to the mixed data as well by transforming the numerical attributes to categorical ones via a discretization method. This kind of methods usually tries to find the cluster structure of categorical data based on the perspective of similarity metric, graph partitioning, or information entropy. For example, ROCK algorithm [53] is an agglomerative hierarchical clustering procedure based on the concepts of similarity-based neighbors and links. By contrast, CLICKS algorithm [54] mines subspace clusters for categorical data set by encoding the data set into a weighted graph structure. In addition, the COOLCAT algorithm, an entropy-based method proposed in [55], utilizes the information entropy to measure the closeness between objects and presents a scheme to find a clustering structure via minimizing the expected entropy of clusters. Furthermore, a scalable algorithm for categorical data clustering called LIMBO [56], which is proposed based on the information bottleneck framework [57], employs the concept of mutual information to find a clustering with minimum information loss. In general, all of the above-stated algorithms can be applied to mixed data via a discretization process, which may, however, cause loss of important information, e.g., the difference between numerical values.

By contrast, the work in the third category attempts to design a generalized clustering criterion for numerical-and-categorical attributes. For example, the similarity-based agglomerative clustering algorithm [8] utilizes Goodall similarity metric [58] to quantify the similarity between data objects with mixed attributes. He *et al*. [12] extended the Squeezer algorithm to cluster mixed data and proposed the usm-squeezer method. In [13], an evidence-based spectral clustering algorithm has been proposed for mixed data clustering by integrating the evidence-based similarity metric into the spectral clustering structure. More recently, a general clustering framework based on object-cluster similarity has been proposed, through which a unified similarity metric for both categorical and numerical attributes has been presented [15]. In addition, a clustering model within Bayesian predictive framework has also been presented for mixed data [59], in which clustering solutions corresponding to random partitions of given data and optimal partition are found via a greedy search. Recently, to solve the labeling problem in sampling clustering on large mixed data, Sangam and Om [60] proposed a hybrid similarity coefficient to find the resemblance between a data point and a cluster, based on which a hybrid data labeling algorithm was further presented to designate appropriate cluster labels to the data points that are not sampled. Among this category of approaches, the most popular one would be the distance-based method, which studies distance metric or dissimilarity measure for categorical values and then utilizes k-means paradigm for clustering analysis. A typical example is the k-prototypes algorithm [11], in which the distance between categorical values is measured with Hamming distance metric. Other distance measures available in the literature include the association-based dissimilarity measure [61], Ahmad's distance measure [14], context-based distance [62], and the distance metric proposed in [63].

B. Subspace Clustering

Generally, the purpose of subspace clustering is to identify the feature subsets where clusters can be found and explore different clusters from different feature subsets [41]. According to the ways with which the feature subsets are identified, the existing subspace clustering methods can be divided into two categories: hard subspace clustering and soft subspace clustering [35], [64].

1) Hard Subspace Clustering: Most hard subspace clustering methods utilize a grid-based clustering notion. Representative algorithms include the CLIQUE [22], ENCLUS [23], and MAFIA [24], [65]. CLIQUE [22] is the pioneering approach to subspace clustering. It first partitions the data space into equal-sized units with an axis-parallel grid. Subsequently, only the units that contain a predefined number of sample points are considered as dense, and the clusters can be explored by gradually grouping the connected dense units together. The ENCLUS algorithm [23] is an improved version of CLIQUE. It introduces the concept of subspace entropy, which is utilized to prune away subspaces with poor clustering performance. Usually, the subspace entropy is determined by three criteria of good clustering in a subspace, i.e., high data coverage, high density, and correlated dimensions. The subspaces with good clustering will have a lower entropy and will be selected to explore the clusters. Moreover, the MAFIA algorithm [24], [65] extends the CLIQUE method by utilizing the adaptive and variable-sized intervals (bins) in each dimension. These bins are then merged to explore clusters in higher subspaces with a candidate generate-and-test scheme. The usage of adaptive grid enables the MAFIA algorithm to reduce the computational cost and improve the clustering quality. Another variant of CLIQUE, called nCluster [29], utilizes a more flexible method to partition the dimensions and allows overlap between different bins of one dimension. As this method may result in much more bins, only maximal δ-nClusters will be mined to avoid generating too many clusters. Generally, the accuracy and efficiency of these grid-based methods primarily depend on the granularity and positioning of the grid. A higher grid granularity will most likely produce more accurate results, but at the same time, will result in a higher time complexity [64]. Different from the grid-based approaches, the SUBCLU algorithm [66] is based on the definition of density-connected sets and utilizes DBSCAN algorithm [67] to discover the clusters in subspaces. This enables SUBCLU to explore clusters with arbitrary shape and size. However, the workload of SUBCLU is huge, which reduces its practicability [68].

2) Soft Subspace Clustering: Different from the hard subspace clustering that identifies the exact subspaces,

soft subspace clustering usually assigns a weight vector to each attribute to measure its contribution to the formation of different clusters. Generally, soft subspace clustering can be regarded as an extension of the conventional attribute-weighted clustering methods [49], [69], [70] that only employ one weight value to adjust the contribution of an attribute to the whole clustering procedure [40]. Recently, increasing research attention has been devoted to soft subspace clustering, and different methods have been presented in [64] and [71]. Quite a lot of work among them has been dedicated to developing weighting schemes for k-means-type algorithms. For example, the locally adaptive clustering (LAC) algorithm [33] assigns local weights to the attributes according to the local correlations of data along different dimensions in each cluster. Particularly, the weights are quantified based on the average in-cluster distance along each dimension. Similar idea has been employed in [31], in which the compact attributes of a cluster where the projected distance along the corresponding dimension is less than the average projected distance values along all dimensions will be assigned relatively higher weights. Moreover, the work in [34], [35], and [72] has introduced the concept of entropy to control the attribute weights and proposed the entropy weighting subspace clustering algorithms, in which the weight of an attribute in a cluster is regarded as the contribution probability of that dimension in forming the cluster. As the most soft subspace clustering approaches only utilize within-cluster information to estimate the attribute weights, Deng *et al*. [71] have presented the enhanced soft subspace clustering (ESSC) method, which employs both within-cluster compactness and between-cluster separation to develop a new fuzzy optimization objective function.

In general, all the above-mentioned soft subspace clustering methods are only applicable to purely numerical data. By contrast, Chan *et al*. [73] have introduced an attribute-weighted distance measure into the framework of general k-means method. Subsequently, if Hamming distance [74] is adopted to measure the distance between categorical values, the proposed subspace clustering algorithm can also be applied to categorical and mixed data following the procedure of k-modes [75] and k-prototypes [11] algorithms. However, as the distance between categorical values is limited to 0 or 1 and the attribute weights are defined with within-cluster distances, the weight value of categorical attribute has a high possibility to be 0 or 1 in practice, which will not work well during clustering process [40], [41]. Therefore, Bai *et al*. [40] and Cao *et al.* [41] have proposed some other attribute-weighting techniques for k-modes algorithm to conduct soft subspace clustering on purely categorical data. Nevertheless, data mixed with categorical and numerical attributes have yet to be considered by them.

III. Attribute-Weighted OCIL Algorithm for Soft Subspace Clustering of Mixed Data

A. OCIL Clustering Algorithm for Mixed Data Analysis

The OCIL algorithm proposed in [15] is based on the concept of object-cluster similarity and follows the optimization procedure of k-means-type methods. Extensive experiments have demonstrated that this clustering algorithm can obtain satisfying performance on both categorical and mixed data sets. The general task of OCIL algorithm is to group the given data objects into several clusters such that the similarity between objects in the same cluster is maximized. Specifically, suppose we have a mixed data set X of N objects, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, to be grouped into k different clusters, denoted as $C_1, C_2, \ldots, C_k$. Each mixed data object $\mathbf{x}_i$ is represented with d different attributes $\{A_1, A_2, \ldots, A_d\}$, which consists of d_c categorical attributes $\{A_1^c, A_2^c, \ldots, A_{d_c}^c\}$ and d_u numerical attributes $\{A_1^u, A_2^u, \ldots, A_{d_c}^u\}$ ($d_c + d_u = d$). Thus, $\mathbf{x}_i$ can be denoted as $[\mathbf{x}_i^{cT}, \mathbf{x}_i^{u\,T}]^T$ with $\mathbf{x}_i^c = (x_{i1}^c, x_{i2}^c, \ldots, x_{id_c}^c)^T$ and $\mathbf{x}_i^u = (x_{i1}^u, x_{i2}^u, \ldots, x_{id_u}^u)^T$. Here, x_{ir}^u $(r = 1, 2, \ldots, d_u)$ belongs to $\mathbf{R}$ and x_{ir}^c $(r = 1, 2, \ldots, d_c)$ belongs to $\mathrm{dom}(A_r^c)$, where $\mathrm{dom}(A_r^c)$ contains all the possible values that can be chosen by attribute A_r^c. Usually, $\mathrm{dom}(A_r^c)$ with m_r elements can be represented with $\mathrm{dom}(A_r^c) = \{a_{r1}, a_{r2}, \ldots, a_{rm_r}\}$. The goal of OCIL algorithm is to find the optimal $\mathbf{Q}^*$ via the following objective function:

$$\mathbf{Q}^* = \arg\max_{\mathbf{Q}} \left[\sum_{j=1}^{k} \sum_{i=1}^{N} q_{ij}\, s(\mathbf{x}_i, C_j) \right] \tag{1}$$

where $s(\mathbf{x}_i, C_j)$ is the similarity between object $\mathbf{x}_i$ and cluster C_j, and $\mathbf{Q} = (q_{ij})$ is an $N \times k$ partition matrix satisfying

$$\sum_{j=1}^{k} q_{ij} = 1, \text{ and } 0 < \sum_{i=1}^{N} q_{ij} < N \tag{2}$$

with

$$q_{ij} \in [0, 1], \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, k. \tag{3}$$

According to [15], the object-cluster similarity for mixed data can be simply defined as

$$s(\mathbf{x}_i, C_j) = \frac{1}{d_f} \left\{ \sum_{r=1}^{d_c} s\left(x_{ir}^c, C_j\right) + s\left(\mathbf{x}_i^u, C_j\right) \right\} \tag{4}$$

with

$$s\left(x_{ir}^c, C_j\right) = \frac{\Psi_{A_r^c = x_{ir}^c}(C_j)}{\Psi_{A_r^c \neq \mathrm{NULL}}(C_j)} \tag{5}$$

and

$$s\left(\mathbf{x}_i^u, C_j\right) = \frac{\exp\left(-0.5\mathrm{Dis}\left(\mathbf{x}_i^u, \mathbf{c}_j\right)\right)}{\sum_{t=1}^{k} \exp\left(-0.5\mathrm{Dis}\left(\mathbf{x}_i^u, \mathbf{c}_t\right)\right)} \tag{6}$$

where $d_f = d_c + 1$. $\Psi_{A_r^c = x_{ir}^c}(C_j)$ counts the number of objects in cluster C_j that have the value x_{ir}^c for attribute A_r^c, NULL refers to the empty, and $\Psi_{A_r^c \neq \mathrm{NULL}}(C_j)$ means the number of objects in cluster C_j that have the attribute A_r^c. $\mathbf{c}_j$ is the center of all numerical vectors in cluster C_j and Dis(.) stands for a distance function. Specifically, if the Euclidean distance is adopted to measure the distance between numerical vectors, (6) can be rewritten as

$$s\left(\mathbf{x}_i^u, C_j\right) = \frac{\exp\left(-0.5\left\|\mathbf{x}_i^u - \mathbf{c}_j\right\|^2\right)}{\sum_{t=1}^{k} \exp\left(-0.5\left\|\mathbf{x}_i^u - \mathbf{c}_t\right\|^2\right)}. \tag{7}$$

In this paper, we concentrate on hard cluster partition only, i.e., $q_{ij} \in \{0, 1\}$. Thus, the optimal $\mathbf{Q}^* = \{q_{ij}^*\}$ in (1) can be obtained with

$$q_{ij}^* = \begin{cases} 1, & \text{if } s(\mathbf{x}_i, C_j) \geq s(\mathbf{x}_i, C_t) \quad \forall 1 \leq t \leq k \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, k$. Subsequently, the OCIL algorithm for mixed data clustering can be described as Algorithm 1 [15].

Algorithm 1 OCIL Algorithm for Mixed Data Clustering

Input: data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and the number of clusters k
Output: cluster label $Y = \{y_1, y_2, \ldots, y_N\}$
Set $Y = \{0, 0, \ldots, 0\}$ and select k initial objects, one for each cluster
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** N **do**
    $y_i^{(new)} = \arg \max_{j \in \{1,\ldots,k\}} [s(\mathbf{x}_i, C_j)]$
    if $y_i^{(new)} \neq y_i^{(old)}$ **then**
      $noChange = false$
      Update the information of clusters $C_{y_i^{(new)}}$ and $C_{y_i^{(old)}}$, including the frequency of each categorical value and the centroid of numerical vectors.
    end if
  end for
until $noChange$ is $true$

B. Attribute-Weighted OCIL Algorithm

The basic idea of soft subspace clustering is to introduce the attribute-cluster weights into the clustering process to measure the contribution of each attribute in forming different clusters. Therefore, the object-cluster similarity-based soft subspace clustering for mixed data is to maximize the following objective function:

$$\mathbf{Q}^* = \arg \max_{\mathbf{Q}, \mathbf{W}} \left[\sum_{j=1}^{k} \sum_{i=1}^{N} q_{ij} s_w(\mathbf{x}_i, C_j) \right] \tag{9}$$

where $s_w(\mathbf{x}_i, C_j)$ denotes the attribute-weighted object-cluster similarity and $\mathbf{W} = (w_{rj})$ is a $d \times k$ weight matrix. According to (4), $s_w(\mathbf{x}_i, C_j)$ can be defined as

$$s_w(\mathbf{x}_i, C_j) = \frac{1}{d_f} \left\{ \sum_{r=1}^{d_c} s_w\left(x_{ir}^c, C_j\right) + s_w\left(\mathbf{x}_i^u, C_j\right) \right\} \tag{10}$$

where the attribute-weighted similarity $s_w(x_{ir}^c, C_j)$ is calculated by $s_w(x_{ir}^c, C_j) = w_{rj}^c s(x_{ir}^c, C_j)$. Here, w_{rj}^c, $r = 1, 2, \ldots, d_c$ denotes the weight of categorical attribute A_r^c to cluster C_j, which satisfies $0 \leq w_{rj}^c \leq 1$. Moreover, based on (7), the attribute-weighted object-cluster similarity on numerical part can then be expressed as

$$s_w\left(\mathbf{x}_i^u, C_j\right) = \frac{\exp\left(-0.5 \sum_{r=1}^{d_u} w_{rj}^u \left(x_{ir}^u - c_{jr}\right)^2\right)}{\sum_{t=1}^{k} \exp\left(-0.5 \sum_{r=1}^{d_u} w_{rt}^u \left(x_{ir}^u - c_{tr}\right)^2\right)} \tag{11}$$

where w_{rj}^u $(r = 1, 2, \ldots, d_u)$ denotes the weight of the rth numerical attribute to cluster C_j and satisfies $0 \leq w_{rj}^u \leq 1$. In general, the sum of all the attribute weights to a particular cluster C_j should be 1. That is, w_{rj}^c and w_{rj}^u satisfy

$$\sum_{r=1}^{d_c} w_{rj}^c + \sum_{r=1}^{d_u} w_{rj}^u = 1, \quad j = 1, 2, \ldots, k. \tag{12}$$

Utilizing a simplified symbol w_{rj} to represent the weight of an arbitrary attribute A_r to cluster C_j, we will have

$$\sum_{r=1}^{d} w_{rj} = 1, \quad j = 1, 2, \ldots, k. \tag{13}$$

To well measure $w_{rj} (r = 1, 2, \ldots, d, \ j = 1, 2, \ldots, k)$, the contribution of the rth attribute to the detection of cluster C_j, denoted as H_{rj}, will be investigated. According to the general task of clustering analysis, two important factors should be considered when analyzing the contribution of a particular attribute. The first one is the intercluster difference, denoted as F_{rj}, which measures the ability of attribute A_r in distinguishing cluster C_j from the other clusters. The other factor is intracluster similarity, denoted as M_{rj}, which evaluates whether the cluster C_j along the attribute A_r has a compact structure or not.

First, to evaluate the intercluster difference F_{rj}, we can compare the value distribution of attribute A_r within the cluster C_j with the distribution outside cluster C_j. The larger difference these two value distributions have, the better the cluster C_j can be distinguished from the other clusters with attribute A_r. Thus, we introduce the Hellinger distance, which is derived from the Bhattacharyya coefficient [76] and can work as an effective metric to quantify the dissimilarity between two probability distributions [77], [78]. Let P_1 and P_2 denote two probability distributions, then the Hellinger distance $\mathrm{HD}(P_1, P_2)$ between them satisfies $0 \leq \mathrm{HD}(P_1, P_2) \leq 1$. If A_r is a categorical attribute, we denote its value distribution within cluster C_j as $P_1^c(r, j)$ and the value distribution outside cluster C_j as $P_2^c(r, j)$. Then, the Hellinger distance between $P_1^c(r, j)$ and $P_2^c(r, j)$ is calculated by

$$\mathrm{HD}\left(P_1^c(r, j), P_2^c(r, j)\right) = \frac{1}{\sqrt{2}} \sqrt{\sum_{t=1}^{m_r} \left(\frac{\Psi_{A_r^c = a_{rt}}(C_j)}{\Psi_{A_r^c \neq \mathrm{NULL}}(C_j)} - \frac{\Psi_{A_r^c = a_{rt}}(X \setminus C_j)}{\Psi_{A_r^c \neq \mathrm{NULL}}(X \setminus C_j)} \right)^2}. \tag{14}$$

If A_r is a numerical attribute, for simplicity, we will utilize Gaussian distribution to estimate its within-cluster distribution and outside-cluster distribution. That is, we assume $P_1^u(r, j) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $P_2^u(r, j) \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with

$$\begin{aligned} \mu_1 &= \frac{1}{N_j} \sum_{\mathbf{x}_i \in C_j} x_{ir}, \quad \mu_2 = \frac{1}{N - N_j} \sum_{\mathbf{x}_i \notin C_j} x_{ir} \\ \sigma_1^2 &= \frac{1}{N_j - 1} \sum_{\mathbf{x}_i \in C_j} (x_{ir} - \mu_1)^2 \\ \sigma_2^2 &= \frac{1}{N - N_j - 1} \sum_{\mathbf{x}_i \notin C_j} (x_{ir} - \mu_2)^2 \end{aligned} \tag{15}$$

where N_j stands for the number of data objects in cluster C_j. Subsequently, the Hellinger distance between $P_1^u(r, j)$ and $P_2^u(r, j)$ is calculated by

$$\mathrm{HD}\big(P_1^u(r, j), P_2^u(r, j)\big) = \sqrt{1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right)}. \quad (16)$$

Thus, the intercluster difference F_{rj} ($r = 1, 2, \ldots, d$, $j = 1, 2, \ldots, k$) can be quantified with

$$F_{rj} = \begin{cases} \mathrm{HD}\big(P_1^c(r, j), P_2^c(r, j)\big), & \text{if } A_r \text{ is categorical} \\ \mathrm{HD}\big(P_1^u(r, j), P_2^u(r, j)\big), & \text{if } A_r \text{ is numerical.} \end{cases} \quad (17)$$

Second, to investigate the intracluster similarity M_{rj} for attribute A_r and cluster C_j, the object-cluster similarity concept [15] can be utilized. Specifically, M_{rj} can be estimated with the average object-cluster similarity in cluster C_j along attribute A_r, that is

$$M_{rj} = \frac{1}{N_j} \sum_{\mathbf{x}_i \in C_j} s(x_{ir}, C_j) \quad (18)$$

where

$$s(x_{ir}, C_j) = \begin{cases} \dfrac{\Psi_{A_r = x_{ir}}(C_j)}{\Psi_{A_r \neq \mathrm{NULL}}(C_j)}, & \text{if } A_r \text{ is categorical} \\ \exp(-0.5(x_{ir} - c_{jr})^2), & \text{if } A_r \text{ is numerical.} \end{cases} \quad (19)$$

As both of F_{rj} and M_{rj} reaching large values imply the important contribution of attribute A_r to the detection of cluster C_j, H_{rj} of attribute A_r to cluster C_j can be calculated by

$$H_{rj} = F_{rj} M_{rj}. \quad (20)$$

Since $0 \leq F_{rj} \leq 1$ and $0 \leq M_{rj} \leq 1$, we have $0 \leq H_{rj} \leq 1$. Subsequently, the attribute weight w_{rj} can be defined as

$$w_{rj} = \frac{H_{rj}}{\sum_{t=1}^{d} H_{tj}}, \quad r = 1, 2, \ldots, d, \ \ j = 1, 2, \ldots, k. \quad (21)$$

To conduct soft subspace clustering analysis, an iterative algorithm will be designed to optimize (9). As global statistic information of different clusters is needed by the estimate of attribute weights, the value of w_{rj} should be updated after each learning epoch (i.e., scanning the whole data set once) rather than the input of every single data object. Initially, w_{rj}s can be set at an average value. That is, let $w_{rj} = (1/d)$ for $r = 1, 2, \ldots, d$, $j = 1, 2, \ldots, k$. Consequently, the corresponding attribute-weighted OCIL algorithm is summarized as Algorithm 2. According to the analysis in [15], the time complexity of OCIL algorithm is $O(TkdN)$, where T stands for the total number of learning epoches. Comparing with the OCIL algorithm, the additional procedure needed by WOCIL algorithm is to update the attribute-cluster weights after each learning epoch. To calculate w_{rj}, we will scan the whole data set at most once. Thus, updating the matrix W needs $O(kdN)$ time. Subsequently, the total time cost of Algorithm 2 is also $O(TkdN)$. As k and T are usually much smaller than N in practice, this algorithm is efficient for data set with a large sample size. In the general case of solving (9), the optimization of the cluster partition matrix Q is nonconvex. Therefore, it is a nontrivial task to theoretically analyze the convergence property of the learning algorithm. Under the circumstances, in the experiments, we empirically investigated the convergence of Algorithm 2 and found that the WOCIL usually converges quickly to a good accuracy. The details can be found in Section VI.

Algorithm 2 Attribute-Weighted OCIL Algorithm (WOCIL)

1: **Input:** data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and the number of clusters k
2: **Output:** cluster label $Y = \{y_1, y_2, \ldots, y_N\}$ and attribute-cluster weights matrix W
3: Set $Y = \{0, 0, \ldots, 0\}$, $w_{rj} = \frac{1}{d}$, and select k initial objects, one for each cluster.
4: **repeat**
5: Initialize $noChange = true$.
6: **for** $i = 1$ **to** N **do**
7: $y_i^{(new)} = \arg \max_{j \in \{1, \ldots, k\}} [s_w(\mathbf{x}_i, C_j)]$
8: **if** $y_i^{(new)} \neq y_i^{(old)}$ **then**
9: $noChange = false$
10: Update the information of clusters $C_{y_i^{(new)}}$ and $C_{y_i^{(old)}}$, including the frequency of each categorical value and the centroid of numerical vectors.
11: **end if**
12: **end for**
13: Update w_{rj} according to equations (21), (20), (18), and (17).
14: **until** $noChange$ is $true$

IV. Learning the Number of Clusters

Similar to the most existing soft subspace clustering algorithms, the above presented WOCIL algorithm still needs the number of clusters k to be preassigned exactly equal to the true one; otherwise, an incorrect clustering result will be obtained. To overcome this problem, in this section, we further investigate an attribute-weighted clustering method with the capability of learning the number of clusters automatically.

The studies in [46] have indicated that competitive learning with penalization mechanism can enable the EM and k-means clustering algorithms to select the number of clusters automatically during the learning process by gradually fading out the redundant clusters. Therefore, in this paper, we also adopt this approach to solve the selection problem of the number of clusters in soft subspace clustering. Specifically, a set of cluster weights will be introduced into the objective function expressed by (9), thus resulting in the following equation:

$$\mathbf{Q}^* = \arg \max_{\mathbf{Q}} \left[\sum_{j=1}^{k} \sum_{i=1}^{N} q_{ij} g_j s_w(\mathbf{x}_i, C_j) \right] \quad (22)$$

where g_j ($j = 1, 2, \ldots, k$) is the weight of cluster C_j satisfying $0 \leq g_j \leq 1$. This weight measures the importance of cluster C_j to the whole cluster structure. Specifically, all clusters with a weight value approaching to 1 mean that each

of them is a component of the whole cluster structure. In case a cluster has a very low weight, the similarity between data objects and this cluster will be reduced. Then, according to the cluster assignment criterion, the number of data objects assigned to this cluster will decrease rapidly and finally this cluster will be eliminated.

Generally, the basic idea of competitive learning with penalization mechanism is that, for each input $\mathbf{x}_i$, not only the winning cluster selected from the initialized cluster candidates is updated toward $\mathbf{x}_i$, but also the rival nearest to the winner (i.e., the runner-up) is penalized according to a specific criterion. Thus, in this kind of method, the number k of clusters can be initialized not less than the true one (i.e., $k \geq k^*$) and the redundant clusters can be gradually eliminated during the clustering process. According to the clustering objective represented by (22), for a given data object $\mathbf{x}_i$, the winner C_v among the k clusters should satisfy

$$v = \arg \max_{1 \leq j \leq k} [g_j s_w(\mathbf{x}_i, C_j))]. \tag{23}$$

However, iterative learning according to this criterion will result in that some seed points located in marginal positions will immediately become dead without learning chance any more in the subsequent learning process [79]. To overcome this problem, Ahalt *et al.* [79] proposed to gradually reduce the winning chance of a frequent winning seed point. That is, the relative winning frequency of different clusters will be further utilized to adjust the selection of winner. Thus, (23) will be revised as

$$v = \arg \max_{1 \leq j \leq k} [(1 - \gamma_j) g_j s_w(\mathbf{x}_i, C_j))]. \tag{24}$$

The relative winning frequency γ_j of cluster C_j is defined as

$$\gamma_j = \frac{n_j}{\sum_{t=1}^{k} n_t} \tag{25}$$

where n_j is the winning times of cluster C_j in the past. At the same time, the nearest rival C_r to the winner is determined by

$$r = \arg \max_{1 \leq j \leq k, j \neq v} [(1 - \gamma_j) g_j s_w(\mathbf{x}_i, C_j))]. \tag{26}$$

For each data object $\mathbf{x}_i$, when the winning cluster and its nearest rival are determined, on the one hand, the data object $\mathbf{x}_i$ will be assigned to the winner C_v and the statistic information of this cluster as well as its winning times will be updated accordingly. On the other hand, we will further reward the winner by increasing its weight while penalize the nearest rival by decreasing its weight. As the values of cluster weights g_j, $j = 1, 2, \ldots, k$ are limited to the interval [0, 1], we can update them through an indirect way. Specifically, we utilize sigmoid function to conduct the transformation and let

$$g_j = \frac{1}{1 + e^{(-10\beta_j + 5)}}, \quad j = 1, 2, \ldots, k. \tag{27}$$

The mappings between β_j and g_j are illustrated in Fig. 1. Thus, the updating of g_j can be accomplished by changing the value of β_j instead. Subsequently, following the penalized competitive learning model proposed in [46], the winner C_v will be awarded with:

$$\beta_v^{(\text{new})} = \beta_v^{(\text{old})} + \eta \tag{28}$$

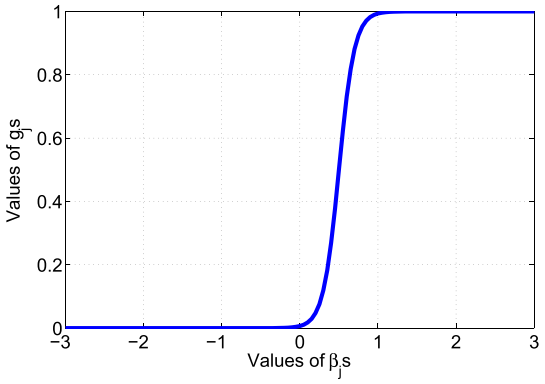


Fig. 1. Illustration of the mappings between the values of β_js and g_js.

and meanwhile, the nearest rival C_r will be penalized with

$$\beta_r^{(\text{new})} = \beta_r^{(\text{old})} - \eta s_w(\mathbf{x}_i, C_r) \tag{29}$$

where η is a small learning rate. From (29), we can see that the rival-penalized strength increases with the similarity between $\mathbf{x}_i$ and the rival. It can be noted that, once the weight of some cluster reaches or approaches to 0, no data objects will be assigned to this cluster during the subsequent learning epoches. According to (20), the contribution of each attribute to this empty cluster will also become 0. Then, the corresponding attribute-cluster weights become meaningless. Finally, the main procedure of the attribute-weighted clustering with the automatic selection of the number of clusters can be summarized as Algorithm 3.

Algorithm 3 Attribute-Weighted Clustering With Rival Penalized Mechanism (RP-WOCIL)

1: **Input:** data set X, learning rate η, and an initial value of k ($k \geq k^*$)
2: **Output:** cluster label $Y = \{y_1, y_2, \ldots, y_N\}$, the number k^* of clusters, and attribute-cluster weights matrix W
3: Select k initial objects, one for each cluster, and set $Y = \{0, 0, \ldots, 0\}$, $w_{rj} = \frac{1}{d}$, $n_j = 1$ and $\beta_j = 1$.
4: **repeat**
5: Initialize $noChange = true$.
6: **for** $i = 1$ **to** N **do**
7: Determine v and r according to (24) and (26).
8: Let $y_i^{(new)} = v$, $n_v^{(new)} = n_v^{(old)} + 1$, and update the statistic information of C_v based on $\mathbf{x}_i$.
9: Update β_v and β_r using (28) and (29).
10: **if** $y_i^{(new)} \neq y_i^{(old)}$ **then**
11: $noChange = false$
12: **end if**
13: **end for**
14: Update w_{rj} according to equations (21), (20), (18), and (17).
15: **until** $noChange$ is $true$

V. Initialization-Oriented Method

In the first step of k-means-type clustering methods, including the WOCIL and RP-WOCIL algorithms, we have to

initialize k clusters. In general, random initialization is the simplest and most popular method. However, due to the uncertainty of the randomly initialized seed points, the clustering results are often somewhat unpredictable and cannot be relied with confidence. Under the circumstances, in order to evaluate the performance of an algorithm, we usually have to repeat the clustering procedure many times with the different initializations to get statistic information. This problem can be even more serious in soft subspace clustering because a poor initialization will degrade the clustering performance meanwhile mislead the learning of attribute-cluster weights. Thus, the whole clustering process will suffer much more negative effects. In the literature, some improved initialization methods for k-means-type clustering have been presented, such as [80]–[83] for numerical data clustering and [84], [85] for categorical data clustering. However, to the best of our knowledge, such initialization refinement for mixed data clustering has yet to be studied. In this section, we will, therefore, propose a new initialization method that conducts a selection process of initial seed points for mixed data clustering.

In general, if we want to get a good clustering result, we often expect that the dissimilarity between the initial seed points is high so that they are more likely from different clusters. Nevertheless, if we only seek for a higher dissimilarity, some outliers that have negative influence on the clustering process may be preferred. Therefore, when we estimate the priority of an object to be selected as an initial seed point, we should simultaneously consider two factors: the dissimilarity between this object and the already selected seed points and also the similarity between this object and the whole data set. Given the mixed data set X for clustering analysis, we let $U = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_l\}$ be the set of seed points that have already been selected, where $l < k$, and k is the desired number of seed points. Similarly, each individual $\boldsymbol{\mu}_j$ $(j \in \{1, 2, \ldots, l\})$ can also be represented by $[\boldsymbol{\mu}_j^c, \boldsymbol{\mu}_j^u]^T$, where $\boldsymbol{\mu}_j^c$ and $\boldsymbol{\mu}_j^u$ denote the categorical and numerical parts, respectively. In the selection process, if we want to choose another seed point and add it to U, we should first estimate the selection priority for each object in X and then select the one with the largest priority. Let $\mathrm{Pry}(\mathbf{x}_i)$ denote the selection priority of object $\mathbf{x}_i$. Since $\mathbf{x}_i$ contains both the categorical and numerical attributes, $\mathrm{Pry}(\mathbf{x}_i)$ can be decomposed as

$$\mathrm{Pry}(\mathbf{x}_i) = \mathrm{Pry}\big(\mathbf{x}_i^c\big) + \mathrm{Pry}\big(\mathbf{x}_i^u\big). \tag{30}$$

Thus, given the estimation of $\mathrm{Pry}(\mathbf{x}_i^c)$ and $\mathrm{Pry}(\mathbf{x}_i^u)$, we can get the priority of $\mathbf{x}_i$.

First, to get the priority of $\mathbf{x}_i^c$, we should estimate the dissimilarity between $\mathbf{x}_i^c$ and U as well as the similarity between $\mathbf{x}_i^c$ and X. To this end, the object-cluster similarity metric for categorical attributes can be utilized to calculate the object-set similarity here. Therefore, analogous to (4) and (5), the similarity between $\mathbf{x}_i^c$ and U can be calculated by

$$\mathrm{Sim}\big(\mathbf{x}_i^c, U\big) = \frac{1}{d_c}\sum_{r=1}^{d_c}\frac{\Psi_{A_r^c=x_{ir}^c}(U)}{\Psi_{A_r^c\neq \mathrm{NULL}}(U)}. \tag{31}$$

Also, the similarity between $\mathbf{x}_i^c$ and X is given by

$$\mathrm{Sim}(\mathbf{x}_i^c, X) = \frac{1}{d_c}\sum_{r=1}^{d_c}\frac{\Psi_{A_r^c=x_{ir}^c}(X)}{\Psi_{A_r^c\neq \mathrm{NULL}}(X)}. \tag{32}$$

Actually, the result of (32) is equivalent to the average density of an object in X defined in [85]. Subsequently, the priority of $\mathbf{x}_i^c$ can be estimated by

$$\begin{aligned}\mathrm{Pry}(\mathbf{x}_i^c) &= \mathrm{DSim}\big(\mathbf{x}_i^c, U\big) + \mathrm{Sim}\big(\mathbf{x}_i^c, X\big)\\ &= \big(1 - \mathrm{Sim}\big(\mathbf{x}_i^c, U\big)\big) + \mathrm{Sim}\big(\mathbf{x}_i^c, X\big)\\ &= \left(1 - \frac{1}{d_c}\sum_{r=1}^{d_c}\frac{\Psi_{A_r^c=x_{ir}^c}(U)}{\Psi_{A_r^c\neq NULL}(U)}\right)\\ &\quad + \frac{1}{d_c}\sum_{r=1}^{d_c}\frac{\Psi_{A_r^c=x_{ir}^c}(X)}{\Psi_{A_r^c\neq \mathrm{NULL}}(X)}\end{aligned} \tag{33}$$

where $\mathrm{DSim}(\mathbf{x}_i, U) = 1 - \mathrm{Sim}(\mathbf{x}_i, U)$ denotes the dissimilarity between $\mathbf{x}_i$ and U.

Next, we discuss the estimation of priority for numerical part of (30), i.e., $\mathrm{Pry}(\mathbf{x}_i^u)$. For any two numerical instances, as the dissimilarity between them is usually quantified by their distance, we can give an estimation of the selection priority for numerical vectors based on the distance metric. First, we find the extremum of all numerical vectors in the whole data set X. Let $\mathbf{x}_{\max}^u = (x_{\max,1}^u, x_{\max,2}^u, \ldots, x_{\max,d_u}^u)^T$ and $\mathbf{x}_{\min}^u = (x_{\min,1}^u, x_{\min,1}^u, \ldots, x_{\min,d_u}^u)^T$ be the upper and lower bounding vectors, respectively. Then, these two specific vectors can be calculated by

$$x_{\max,r}^u = \max_{1\le i\le N}\left(x_{i,r}^u\right) \text{ and } x_{\min,r}^u = \min_{1\le i\le N}\left(x_{i,r}^u\right) \tag{34}$$

where $r \in \{1, 2, \ldots, d_u\}$. Suppose $D(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u)$ denotes the Euclidean distance between $\mathbf{x}_{\max}^u$ and $\mathbf{x}_{\min}^u$. For any two numerical vectors $\mathbf{x}_i^u$ and $\mathbf{x}_j^u$ in X, we then have

$$D\big(\mathbf{x}_i^u, \mathbf{x}_j^u\big) \le D\big(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u\big). \tag{35}$$

Subsequently, similar to the max–min principle in [86], the dissimilarity between $\mathbf{x}_i^u$ and the set of already selected seed points U can be defined as

$$\mathrm{DSim}\big(\mathbf{x}_i^u, U\big) = \frac{\min_{1\le j\le p} D\big(\mathbf{x}_i^u, \boldsymbol{\mu}_j^u\big)}{D\big(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u\big)}. \tag{36}$$

That is, $\mathrm{DSim}(\mathbf{x}_i^u, U)$ is determined by the minimum distance between $\mathbf{x}_i^u$ and the numerical vectors in U. The regulating term $D(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u)$ in (36) is to guarantee a value from interval [0, 1], which will be consistent with the result on categorical attributes.

On the other hand, the similarity between $\mathbf{x}_i^u$ and X can be estimated via investigating whether $\mathbf{x}_i^u$ has a relatively high local density in X. Generally, a standard way to get the local density of $\mathbf{x}_i^u$ is to calculate the distance between $\mathbf{x}_i^u$ and each of other numeric vectors. However, the time complexity of this method is at least $O(N^2)$, which will result in plenty of additional computation cost compared to the random initialization. Therefore, we present an approximation method to estimate the local density. Specifically, all numerical

vectors in X will first be grouped into k clusters with standard k-means algorithm [4]. Let $C^u = \{\mathbf{c}_1^u, \mathbf{c}_2^u, \ldots, \mathbf{c}_k^u\}$ be the centers of these k clusters. Then, the similarity between $\mathbf{x}_i^u$ and X can be estimated with the similarity between $\mathbf{x}_i^u$ and the cluster center set C^u, i.e.

$$\begin{aligned}\mathrm{Sim}\big(\mathbf{x}_i^u, X\big) &= \mathrm{Sim}\big(\mathbf{x}_i^u, C^u\big) = 1 - \mathrm{DSim}\big(\mathbf{x}_i^u, C^u\big) \\ &= 1 - \frac{\min_{1\le j\le k} D\big(\mathbf{x}_i^u, \mathbf{c}_j^u\big)}{D\big(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u\big)}. \end{aligned} \tag{37}$$

Subsequently, we have

$$\begin{aligned}\mathrm{Pry}\big(\mathbf{x}_i^u\big) &= \mathrm{DSim}\big(\mathbf{x}_i^u, U\big) + \mathrm{Sim}\big(\mathbf{x}_i^u, X\big) \\ &= \frac{\min_{1\le j\le p} D\big(\mathbf{x}_i^u, \boldsymbol{\mu}_j^u\big)}{D\big(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u\big)} \\ &\quad + \left(1 - \frac{\min_{1\le j\le k} D\big(\mathbf{x}_i^u, \mathbf{c}_j^u\big)}{D\big(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u\big)}\right). \end{aligned} \tag{38}$$

Finally, based on (30), (33), and (38), the priority of mixed data $\mathbf{x}_i$ to be selected as a clustering seed point is calculated by

$$\begin{aligned}&\mathrm{Pry}(\mathbf{x}_i) \\ &= \mathrm{Pry}\big(\mathbf{x}_i^c\big) + \mathrm{Pry}\big(\mathbf{x}_i^u\big) \\ &= \left(1 - \frac{1}{d_c}\sum_{r=1}^{d_c} \frac{\Psi_{A_r^c = x_{ir}^c}(U)}{\Psi_{A_r^c \neq \mathrm{NULL}}(U)}\right) + \frac{1}{d_c}\sum_{r=1}^{d_c} \frac{\Psi_{A_r^c = x_{ir}^c}(X)}{\Psi_{A_r^c \neq \mathrm{NULL}}(X)} \\ &\quad + \frac{\min_{1\le j\le p} D\big(\mathbf{x}_i^u, \boldsymbol{\mu}_j^u\big)}{D\big(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u\big)} + \left(1 - \frac{\min_{1\le j\le k} D\big(\mathbf{x}_i^u, \mathbf{c}_j^u\big)}{D\big(\mathbf{x}_{\max}^u, \mathbf{x}_{\min}^u\big)}\right). \end{aligned} \tag{39}$$

Particularly, when selecting the first seed point, as U is an empty set, only the similarity between $\mathbf{x}_i$ and X should be investigated. That is, the first seed point $\boldsymbol{\mu}_1$ is selected according to

$$\boldsymbol{\mu}_1 = \arg\max_{\mathbf{x}_i \in X} \big[\mathrm{Sim}\big(\mathbf{x}_i^c, X\big) + \mathrm{Sim}\big(\mathbf{x}_i^u, X\big)\big]. \tag{40}$$

Consequently, the procedure of initialization-oriented method for mixed data clustering can be summarized as Algorithm 4, where $X^u = \{\mathbf{x}_1^u, \mathbf{x}_2^u, \ldots, \mathbf{x}_N^u\}$ and $kmeans(X^u, k)$ means executing k-means clustering on the set X^u. In Algorithm 4, the computation cost from State 3 to State 5 is $O(Nd_u + kNd_u + kNd)$. For each selection, the cost of State 8 and State 9 is $O(Nd_c + (1/2)kNd_u)$. Thus, the total time cost is $O(kNd + (1/2)k^2Nd_u))$. It can be seen that this computation time approaches to that cost by a single iteration of OCIL algorithm. Therefore, the proposed initialization method will not cost much additional computation.

Algorithm 4 Oriented Initialization for Mixed Data Clustering

1: **Input:** data set X and the number of clusters k
2: **Output:** set of initial seed points U
3: Find the extremum vectors $\mathbf{x}_{max}^u$ and $\mathbf{x}_{min}^u$ according to (34).
4: Let $C^u = kmeans(X^u, k)$
5: For each $\mathbf{x}_i \in X$, calculate $Sim(\mathbf{x}_i^c, X)$ and $Sim(\mathbf{x}_i^u, X)$ according to (32) and (37), respectively.
6: Select $\boldsymbol{\mu}_1$ from X based on (40), let $U = \boldsymbol{\mu}_1$ and $Counter = 1$.
7: **while** $Counter < k$ **do**
8: For each $\mathbf{x}_i \in \{X \setminus U\}$, calculate $Pry\ (\mathbf{x}_i)$ with (39).
9: Let $\mathbf{x}_{i*} = \arg\max_{\mathbf{x}_i \in \{X\setminus U\}} [Pry\ (\mathbf{x}_i)]$, $U = U \cup \{\mathbf{x}_{i*}\}$, and $Counter = Counter + 1$.
10: **end while**

VI. Experiments

This section is to investigate the effectiveness of the proposed approaches for soft subspace clustering of data with different types and compare their performance with the existing counterparts. The algorithms were coded with MATLAB and all the experiments were implemented by a desktop PC computer with Intel(R) Core2 Quad CPU, 2.40-GHz main frequency, and 4 GB DDR2 667 RAM.

A. Utilized Data Sets

Experiments were conducted on various mixed, purely categorical, and purely numerical data sets obtained from UCI machine learning data repository (URL: *http://archive.ics.uci.edu/ml/*). Specifically, five mixed data sets have been utilized and the detailed information is as follows.

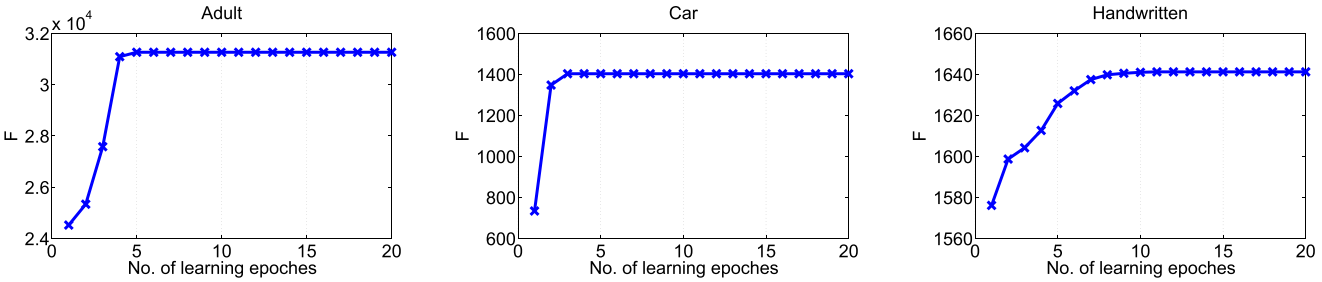
1) *Heart Disease Database:* This data set contains 303 instances concerning heart disease diagnosis, which are characterized by seven categorical attributes and six numeric attributes. All the instances can be grouped into two classes: *healthy* (164 instances) and *sick* (139 instances).
2) *Credit Approval Data Set:* There are 690 data instances about credit card application. Each of them is described by nine categorical attributes and six numeric attributes. In the experiments, 653 instances without missing values were utilized, which have been labeled with *positive* (296 instances) or *negative* (357 instances).
3) *Statlog German Credit Data:* It contains the information of 1000 people described by 13 categorical attributes and 7 numerical attributes. All people are grouped as good or bad credit risk.
4) *Adult Data Set:* This data set was extracted from the census database. 45222 instances without unknown values were adopted. They are described by eight categorical attributes and six numerical attributes, and can be grouped into two clusters.
5) *Dermatology Data Set:* This data set contains 366 instances about six types of Eryhemato-Squamous Disease. All instances are characterized by 34 attributes, 33 of which are numerical and the other one is categorical.

Moreover, the utilized five purely categorical data sets are Soybean, Voting, Wisconsin Breast Cancer Database (WBCD), Car, and Zoo. Their information is listed as follows.

1) *Small Soybean Database:* There are 47 instances characterized by 35 multivalued categorical attributes. According to the different kind of diseases, all the instances should be divided into four groups.

TABLE I
STATISTICS OF THE UTILIZED DATA SETS

Mixed data sets				Categorical data sets				Numerical data sets			
Data set	N	d ($d_c + d_u$)	k^*	Data set	N	d	k^*	Data set	N	d	k^*
Heart	303	7 + 6	2	Soybean	47	35	4	Iris	150	4	3
Credit	653	9 + 6	2	Voting	435	16	2	Wine	178	13	3
German	1000	13 + 7	2	WBCD	699	9	2	Ionosphere	351	34	2
Adult	30162	8 + 6	2	Car	1728	6	4	Handwritten	5620	64	10
Dermatology	366	33 + 1	6	Zoo	101	16	7	Sonar	208	60	2



(a) (b) (c)

Fig. 2. Convergence curve of the WOCIL algorithm on (a) adult data set, (b) car data set, and (c) handwritten data set.

2) *Congressional Voting Records Data Set:* There are 435 votes based on 16 key features and each vote comes from one of the two different party affiliations: *democrat* (267 votes) and *republican* (168 votes).
3) *WBCD:* This data set has 699 instances described by nine categorical attributes with the values from 1 to 10. Each instance belongs to one of the two clusters labeled by *benign* (contains 458 instances) and *malignant* (contains 241 instances).
4) *Car Evaluation Database:* It contains 1728 car samples derived from a simple decision model that evaluates cars according to six different aspects. Each sample is labeled with one of the four categories: unacceptable, acceptable, good, and very good.
5) *Zoo Data Set:* This data set consists of 101 instances represented by 16 attributes, in which each instance belongs to one of the 7 animal categories.

Besides, more experiments were conducted on the following five purely numerical data sets.

1) *Iris Data Set:* It contains 150 instances characterized by four attributes from three classes, where each class refers to a type of iris plant.
2) *Wine Data Set:* There are 178 data samples from a chemical analysis of wines that determined the quantities of 13 constituents found in each of the three types of wines.
3) *Ionosphere Data Set:* This data set is about the classification of radar returns from the ionosphere. It contains 351 instances described by 34 attributes. Each instance has been labeled with *good* or *bad*.
4) *Optical Recognition of Handwritten Digits Data:* The data are extracted from the bitmaps of ten different handwritten digits. Each of the 5620 instance is represented by 64 attributes.
5) *Sonar Data Set:* This data set is utilized to discriminate between sonar signals bounced off a metal cylinder and a roughly cylindrical rock. It contains 208 instances described by 60 attributes.

In addition, Table I has briefly summarized the general information of all utilized data sets for quick check.

B. Study of WOCIL Algorithm

1) Convergence of WOCIL Algorithm: To investigate the convergence property of WOCIL algorithm, this experiment executed it on data sets with different types and recorded the value of objective function after each learning iteration. The adopted data sets were Adult, Car, and Handwritten, which have relatively larger sample sizes among the others. The curves in Fig. 2 show the variation trend of objective function values with the number of learning iterations. Here, F stands for the value of objective function $\sum_{j=1}^{k}\sum_{i=1}^{N} q_{ij} s_w(\mathbf{x}_i, C_j)$. It can be observed that the performance of the WOCIL algorithm was converged within five learning iterations on Adult and Car data sets, and for the Handwritten data sets, about ten learning epoches were spent. These results demonstrate that the WOCIL algorithm has a good convergence rate in practice and can conduct efficient learning on data with large sample size.

2) Performance Evaluation of WOCIL Algorithm: To evaluate the performance of WOCIL algorithm, we applied it to different real data sets and compared its performance with the existing counterparts. As the number of clusters is preassigned in this kind of method, we adopted three popular validity indices, i.e., clustering accuracy (ACC) [87], rand index (RI), and normalized mutual information (NMI), to evaluate the clustering results. The definitions of these three indices are as follows.

TABLE II
CLUSTERING PERFORMANCE IN TERMS OF ACC OF DIFFERENT ALGORITHMS ON MIXED DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM
Heart	0.8356±0	0.8152±0.0517	0.7695±0.0083	0.7337±0.0739	0.7565±0.0803
Credit	0.8537±0	0.7442±0.1278	0.6979±0.1298	0.7572±0.1361	0.7727±0.0798
German	0.6956±0	0.6930±0.0038	0.6670±0.0477	0.5423±0.0274	0.5702±0.0377
Adult	0.7510±0	0.7509±0.0001	0.7509±0.0001	0.7277±0.0001	0.7511±0.0001
Dermatology	0.7860±0	0.7327±0.0865	0.7253±0.0842	0.2832±0.0176	0.6364±0.0944

TABLE III
CLUSTERING PERFORMANCE IN TERMS OF RI OF DIFFERENT ALGORITHMS ON MIXED DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM
Heart	0.7245±0	0.6894±0.0553	0.6473±0.0112	0.6183±0.0608	0.6431±0.0643
Credit	0.7542±0	0.6546±0.1032	0.6068±0.0951	0.6670±0.1155	0.6584±0.0716
German	0.5761±0	0.5741±0.0030	0.5598±0.0262	0.5045±0.0096	0.5116±0.0122
Adult	0.6259±0	0.6258±0.0001	0.6247±0.0001	0.6037±0.0001	0.6261±0.0001
Dermatology	0.9022±0	0.8712±0.0452	0.8838±0.0431	0.6985±0.0068	0.8535±0.0388

TABLE IV
CLUSTERING PERFORMANCE IN TERMS OF NMI OF DIFFERENT ALGORITHMS ON MIXED DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM
Heart	0.3535±0	0.3065±0.0865	0.2333±0.0194	0.1838±0.0937	0.2252±0.0993
Credit	0.4454±0	0.2361±0.1740	0.1816±0.1454	0.2869±0.1963	0.2449±0.1132
German	0.0095±0	0.0063±0.0054	0.0025±0.0018	0.0038±0.0011	0.0125±0.0062
Adult	0.0052±0	0.0054±0.0003	0.0038±0.0011	0.0921±0.0005	0.0012±0.0003
Dermatology	0.6005±0	0.5914±0.1701	0.7964±0.0666	0.0961±0.0118	0.5074±0.1174

1) *Clustering Accuracy*

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N} \delta\ (c_i, \mathrm{map}(l_i))}{N}$$

where N is the number of objects in the data set, c_i stands for the provided label, $\mathrm{map}(l_i)$ is a mapping function that maps the obtained cluster label l_i to the equivalent label from the data corpus, and the delta function $\delta\ (c_i, \mathrm{map}(l_i)) = 1$ only if $c_i = \mathrm{map}(l_i)$, otherwise 0.

2) *Rand Index*

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

3) *Normalized Mutual Information*

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k^*} N_{i,j} \log\left(\frac{N \cdot N_{i,j}}{N_i \cdot N_j}\right)}{\sqrt{\left(\sum_{i=1}^{k} N_i \log \frac{N_i}{N}\right)\left(\sum_{j=1}^{k^*} N_j \log \frac{N_j}{N}\right)}}$$

where k^* stands for the true number of classes, k is the number of clusters obtained by the algorithm, $N_{i,j}$ denotes the number of agreements between cluster i and class j, N_i is the number of data objects in cluster i, N_j is the number of objects in class j, and N is the number of objects in the whole data set.

In general, all of ACC, RI, and NMI have values from interval $[0, 1]$ and larger values of them indicate better performance.

First, we investigated the performance of WOCIL algorithm on mixed data. For comparative studies, the performance of simple OCIL algorithm [15], WKM algorithm [70], and EWKM algorithm [35] on these data sets has also been investigated. All these algorithms are k-means-type method. The WKM and EWKM algorithms are applicable to numerical, categorical, and mixed data if the learning model of k-means, k-modes [75], and k-prototypes [11] methods are utilized, respectively. In the experiment, each algorithm has been executed 50 times on each data set and the clustering results are statistically summarized. According to the authors' recommendation in [70], the parameter β in WKM algorithm was set at 2 and the parameter γ in EWKM algorithm was set at 1.5. Tables II–IV present the clustering results obtained by different methods in the form of the means and standard deviations of ACC, RI, and NMI, respectively. In the tables, WOCIL+OI stands for the WOCIL algorithm initialized with the initialization-oriented method, while the other algorithms are implemented with random initializations. From the results, we can find that the WOCIL algorithm outperforms the OCIL, WKM, and EWKM algorithms in the most cases. This indicates that appropriate attribute weights can improve the clustering performance on the most data sets. The results obtained by different algorithms on the Adult data set are similar to each other, which would indicate that this data set does not have subspace structures. Moreover, comparing the results of the WOCIL+OI and WOCIL methods, we can find that the proposed initialization-oriented method can lead to much better and more stable clustering performance. In addition, to investigate the effectiveness of the proposed attribute

TABLE V
ATTRIBUTE-CLUSTER WEIGHTS LEARNED BY THE WOCIL METHOD ON HEART DATA SET

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
C_1	0.0255	0.0460	0.0077	0.0306	0.1685	0.1224	0.0969	0.0605	0.0292	0.0204	0.1117	0.1769	0.1037
C_2	0.0324	0.1043	0.0105	0.0320	0.1219	0.1452	0.0811	0.0728	0.0258	0.0167	0.1391	0.1426	0.0757

TABLE VI
ATTRIBUTE-CLUSTER WEIGHTS LEARNED BY THE WOCIL METHOD ON CREDIT DATA SET

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
C_1	0.0045	0.0216	0.0216	0.0046	0.0189	0.3283	0.0943	0.0145	0.0174	0.0418	0.0586	0.1363
C_2	0.0043	0.0304	0.0304	0.0056	0.0212	0.3917	0.0910	0.0182	0.0215	0.0457	0.0637	0.1356
	A_{13}	A_{14}	A_{15}									
C_1	0.1969	0.0040	0.0370									
C_2	0.0864	0.0141	0.0403									

TABLE VII
CLUSTERING PERFORMANCE IN TERMS OF ACC OF DIFFERENT ALGORITHMS ON CATEGORICAL DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM	wk-Modes
Soybean	1±0	0.8609±0.1558	0.8349±0.1741	0.7681±0.1623	0.7604±0.1262	0.8572±0.1329
Voting	0.8767±0	0.8747±0.0011	0.8676±0.0019	0.7749±0.1279	0.7834±0.1108	0.8335±0.0513
WBCD	0.8998±0	0.8935±0.0026	0.8796±0.0013	0.7894±0.1229	0.7502±0.1569	0.8339±0.0468
Car	0.4097±0	0.3784±0.0264	0.3499±0.0303	0.3863±0.0625	0.3395±0.0314	0.3779±0.0356
Zoo	0.7624±0	0.6897±0.0883	0.6479±0.0865	0.6844±0.1028	0.6941±0.0919	0.6852±0.0869

TABLE VIII
CLUSTERING PERFORMANCE IN TERMS OF RI OF DIFFERENT ALGORITHMS ON CATEGORICAL DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM	wk-Modes
Soybean	1±0	0.9202±0.0946	0.9093±0.1075	0.8524±0.0920	0.8509±0.0724	0.8965±0.0845
Voting	0.7884±0	0.7803±0.0017	0.7698±0.0028	0.6825±0.1151	0.6840±0.1001	0.7328±0.0621
WBCD	0.8082±0	0.8021±0.0015	0.7882±0.0021	0.6967±0.1393	0.6730±0.1520	0.7469±0.0333
Car	0.5291±0	0.5056±0.0134	0.5002±0.0231	0.5083±0.0311	0.4890±0.0102	0.5033±0.0189
Zoo	0.9097±0	0.8787±0.0370	0.8642±0.0426	0.8747±0.0488	0.8770±0.0412	0.8775±0.0415

TABLE IX
CLUSTERING PERFORMANCE IN TERMS OF NMI OF DIFFERENT ALGORITHMS ON CATEGORICAL DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM	wk-Modes
Soybean	1±0	0.8974±0.1162	0.8831±0.1311	0.7404±0.1707	0.7592±0.1222	0.8751±0.1031
Voting	0.4967±0	0.4836±0.0024	0.4757±0.0036	0.3334±0.1831	0.3449±0.1369	0.4408±0.0775
WBCD	0.5249±0	0.5138±0.0025	0.5024±0.0056	0.3260±0.2474	0.3086±0.2504	0.4552±0.1338
Car	0.1029±0	0.0895±0.0602	0.0817±0.0661	0.0778±0.0543	0.0440±0.0192	0.0835±0.0646
Zoo	0.8290±0	0.7748±0.0475	0.7452±0.0605	0.7482±0.0848	0.7458±0.0795	0.7471±0.0552

weighting scheme, we took the Heart data set as an example to show the changes of attribute weights during the learning process. Initially, all attribute-cluster weights were set at an equal value, i.e., $1/d$. After the learning algorithm converged, the obtained attribute-cluster weights were the values shown in Table V. It can be seen that the proposed weighting scheme can well distinguish the contribution of different attributes to the detection of each cluster. Besides, one more example has been shown in Table VI, from which we can draw a similar conclusion.

Moreover, to investigate the performance of WOCIL algorithm on purely categorical data, further clustering analysis was conducted on five categorical data sets. Besides the WKM and EWKM algorithms, the proposed method has also been compared with the wk-Modes algorithm [41], which is a subspace clustering method for purely categorical data. The average clustering performance as well as the standard deviation evaluated with the three criteria has been recorded in Tables VII–IX. It can be observed that, with random initializations, the WOCIL algorithm has competitive advantage in terms of ACC compared to the OCIL, WKM, EWKM, and wk-Modes methods on the first three data sets. Although the WKM and EWKM algorithms have better performance on the Car and Zoo data sets, respectively, the results of the WOCIL algorithm are still comparable to theirs. In addition, the problem of large standard deviation in ACC caused by random initializations on categorical data can also be solved by the initialization-oriented method. Particularly, the true label

TABLE X
CLUSTERING PERFORMANCE IN TERMS OF ACC OF DIFFERENT ALGORITHMS ON NUMERICAL DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM	LAC	ESSC
Iris	0.9067±0	0.8537±0.14	0.7957±0.154	0.8383±0.153	0.7240±0.166	0.8212±0.128	0.8409±0.131
Wine	0.9607±0	0.9404±0.099	0.6745±0.053	0.9502±0.004	0.6863±0.102	0.9215±0.030	0.9335±0.327
Ionosphere	0.7223±0	0.7174±0.015	0.7102±0.009	0.7115±0.001	0.6460±0.026	0.6957±0.011	0.7116±0.007
Handwritten	0.7328±0	0.6902±0.066	0.7125±0.054	0.3018±0.053	0.1619±0.027	0.3519±0.048	0.3356±0.069
Sonar	0.5488±0	0.5457±0.031	0.5218±0.007	0.5413±0.023	0.5227±0.006	0.5113±0.008	0.5345±0.005

TABLE XI
CLUSTERING PERFORMANCE IN TERMS OF RI OF DIFFERENT ALGORITHMS ON NUMERICAL DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM	LAC	ESSC
Iris	0.8923±0	0.8701±0.064	0.8323±0.073	0.8639±0.085	0.7486±0.135	0.8506±0.115	0.8695±0.052
Wine	0.9467±0	0.9356±0.076	0.7126±0.012	0.9348±0.005	0.7090±0.062	0.8953±0.032	0.9157±0.022
Ionosphere	0.5934±0	0.5889±0.012	0.5854±0.007	0.5863±0.001	0.5427±0.021	0.5668±0.009	0.5869±0.004
Handwritten	0.9281±0	0.9041±0.086	0.9226±0.011	0.7122±0.075	0.3340±0.130	0.7619±0.054	0.7337±0.082
Sonar	0.5063±0	0.5019±0.006	0.5004±0.002	0.5021±0.004	0.4987±0.001	0.4993±0.004	0.5016±0.002

TABLE XII
CLUSTERING PERFORMANCE IN TERMS OF NMI OF DIFFERENT ALGORITHMS ON NUMERICAL DATA SETS

Data sets	WOCIL+OI	WOCIL	OCIL	WKM	EWKM	LAC	ESSC
Iris	0.8058±0	0.7636±0.076	0.7057±0.074	0.7559±0.099	0.5550±0.219	0.7205±0.106	0.7423±0.069
Wine	0.8610±0	0.8535±0.116	0.4271±0.004	0.8281±0.008	0.4015±0.123	0.8312±0.061	0.8408±0.057
Ionosphere	0.1428±0	0.1322±0.020	0.1253±0.014	0.1249±0.0004	0.0291±0.059	0.0733±0.013	0.1225±0.001
Handwritten	0.7426±0	0.6944±0.056	0.7182±0.021	0.4048±0.052	0.0308±0.003	0.4843±0.061	0.5506±0.063
Sonar	0.0136±0	0.0085±0.005	0.0053±0.004	0.0146±0.010	0.0347±0.009	0.0089±0.003	0.0101±0.002

of the Soybean data set could always be obtained by the WOCIL+OI method in our experiments.

Besides, the performance of WOCIL algorithm on purely numerical data sets was further studied. The evaluations of clustering outcomes obtained by different algorithms have been listed in Tables X–XII in the form of the means and standard deviations of ACC, RI, and NMI, respectively. Two subspace clustering methods for purely numerical data, i.e., LAC [33] and ESSC [71], were investigated as well for comparative study. From the statistical results, we can find that the best results on different data sets have also been obtained by the WOCIL+OI method. Among the other random-initialization-based methods, the WOCIL method outperforms the others in the most cases. An additional observation is that, for the Handwritten data set, the OCIL algorithm without attribute weighting has a better performance than the WOCIL, WKM, EWKM, LAC, and ESSC methods. This result indicates that subspace clustering is not suitable for all kinds of data sets. Some cases prefer the attributes to be treated equally during clustering analysis.

In addition, we further investigated the real execution time of WOCIL algorithm on different types of data sets. Table XIII reports the average execution time of five main algorithms over 50 repeats on mixed, categorical, and numerical data sets. Comparing the results, we can find that the WOCIL algorithm generally has the same time complexity level with the other algorithms. Due to the additional computation costs by the calculation of attribute weights, the WOCIL algorithm needs more execution time than the OCIL method on each data set. However, the time difference between them is not large and will not reduce the practicability of WOCIL method. Moreover, it can also be observed that the real execution time of WOCIL+OI method is just a little more than that of the WOCIL algorithm with random initialization. This result validates that the proposed initialization-oriented method will not bring much additional computation load.

TABLE XIII
COMPARISON OF THE AVERAGE EXECUTION TIME BETWEEN DIFFERENT ALGORITHMS (IN SECONDS)

Data sets	WOCIL +OI	WOCIL	OCIL	WKM	EWKM
Heart	0.2352	0.1548	0.0863	0.1711	0.0866
Credit	2.7554	2.4305	0.1030	2.5670	0.1970
German	3.3136	2.8907	0.4376	9.3610	0.3307
Adult	11.6588	7.2252	3.2949	14.4700	5.7764
Dermatology	0.8505	0.7113	0.1864	0.3172	0.6098
Soybean	0.0049	0.0082	0.0051	0.1559	0.0333
Voting	0.0505	0.0520	0.0327	0.1039	0.0894
WBCD	0.0930	0.0927	0.0756	0.0981	0.1257
Car	0.1660	0.1626	0.1456	0.8153	0.4965
Zoo	0.0179	0.0145	0.0109	0.4707	0.0817
Iris	0.0733	0.0448	0.0318	0.0102	0.0194
Wine	0.1208	0.0686	0.0213	0.0148	0.0354
Ionosphere	0.1428	0.0867	0.0373	0.0292	0.0206
Handwritten	4.6313	1.8675	1.1350	2.9293	0.6893
Sonar	0.2225	0.1510	0.0339	0.0241	0.0274

C. Performance Evaluation of RP-WOCIL Algorithm

1) Convergence of RP-WOCIL Algorithm: In order to investigate the convergence performance of RP-WOCIL algorithm

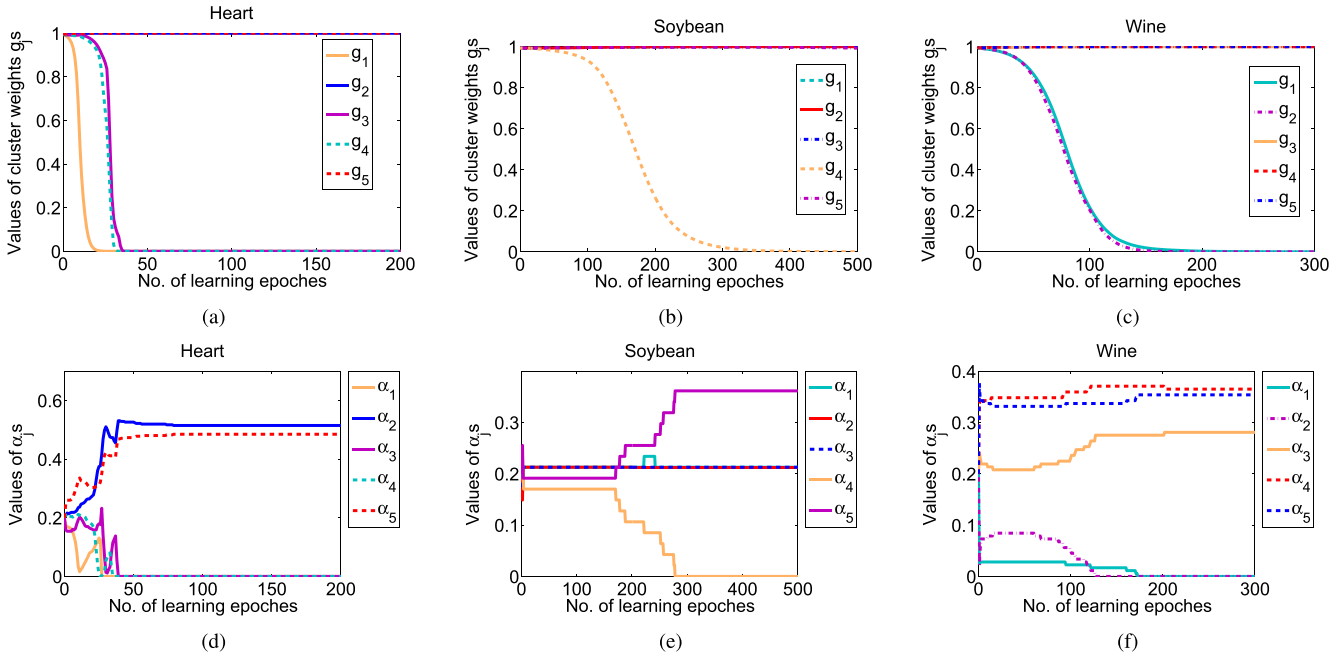


Fig. 3. Learning curves of g_js and α_js obtained by the RP-WOCIL algorithm on Heart, Soybean, and Wine data sets.

and its ability of automatically learning the number of clusters, this experiment applied it to the Heart, Soybean, and Wine data sets, which belong to different date types and have a varied number of clusters. Fig. 3 shows the detailed learning process of RP-WOCIL algorithm on each data set. In the experiment, the initial number of clusters k was set at five and the variable α_j, $j = 1, 2, \ldots, k$, was utilized to record the proportion of objects among the whole data set that had been assigned to the jth cluster. The learning curves of α_js and cluster weights g_js over the epoch numbers obtained by the RP-WOCIL algorithm on the three data sets are shown in Fig. 3. It can be observed from Fig. 3(a) that the values of three g_js converged to around 0 after about 40 learning epoches while the other two converged to 1. This means that the learning algorithm has identified the true number of clusters by eliminating the three redundant clusters from the initial cluster structure. Meanwhile, it can be found from Fig. 3(d) that α_js of the three redundant clusters also converged to 0 as no data objects will be assigned to a cluster with zero weight. The final obtained values of the other two α_js were 0.5149 and 0.4851, which are approximate to the data-member proportions of the two true clusters, i.e., 54.13% and 45.87%. For the categorical data set, Soybean data, the weight value of one cluster converged to 0, while the other four converged to 1. The correct cluster structure has also been leant. Moreover, similar result has been obtained on the Wine data set, where the two redundant clusters have been successfully eliminated and the obtained nonzero α_js are 0.3652, 0.3539, and 0.2809.

2) Parameter Analysis on Learning Rate η: The setting of learning rate η plays an important rule in the RP-WOCIL algorithm. Generally, a too small value of η may lead to an insufficient penalization process by which the redundant clusters cannot be completely eliminated from the initial cluster structure. Conversely, a too large value of η may cause an excessive penalization, such that the initialized clusters are overeliminated and the obtained number of clusters will less than the true one. To investigate a good setting method for this parameter, we have performed the RP-WOCIL algorithm on different data sets with the varied settings of learning rate η. From these experiments we find that the robustness of RP-WOCIL's performance to the setting of η has obvious difference on different data sets. A plausible reason is that the cluster structures of different data sets usually have the varied levels of partition difficulty. Fig. 4 presents the experimental results on Heart and Soybean data sets for comparison, where the initial value of k was set at 5. As shown in Fig. 4(a), for the Heart data sets, the RP-WOCIL algorithm can always obtain the correct number of clusters with η setting from 0.0001 to 0.005. By contrast, for the Soybean data set, if the value of η is set larger than 0.0006, the RP-WOCIL algorithm tends to overpenalize the initialized clusters and the obtained number of clusters will less than the true one. Moreover, from Fig. 4(b) and (d), we can find that the convergence speed of RP-WOCIL algorithm will slow up as the value of η decreases. Based on this empirical study, it is appropriate to set the value of η between 0.0001 and 0.0005 in practice.

3) Performance Evaluation of the RP-WOCIL Algorithm: This experiment was to investigate the performance of the RP-WOCIL algorithm on different data sets. Specifically, six data sets were utilized, which consisted of two mixed data sets, two categorical data sets, and two numerical data sets. To study the ability of the RP-WOCIL algorithm in learning the number of clusters, we implemented the experiments with the varied settings of k, and the statistic result over 20 runs in each case has been recorded. Moreover, according to [44], the performance of clustering algorithms with the different

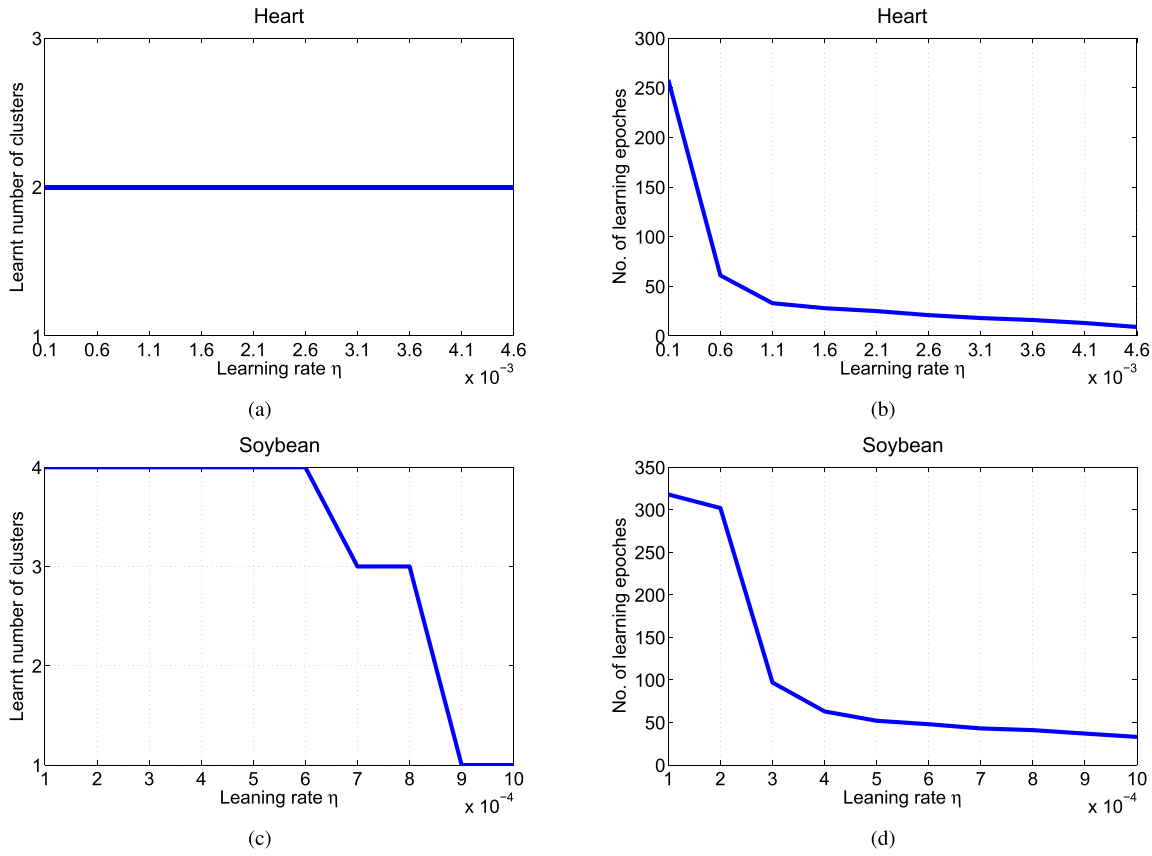


Fig. 4. Learning performance of RP-WOCIL algorithm on Heart and Soybean data sets with different settings of learning rate η.

TABLE XIV
CLUSTERING RESULTS OBTAINED BY DIFFERENT ALGORITHMS WITH THE VARIED SETTINGS OF k

Data sets	k^*	k	Cluster no.		PQ	
			RP-WOCIL	WOCIL	RP-WOCIL	WOCIL
Heart	2	3	1.95±0.22	3±0	0.7251±0.0523	0.6248±0
		4	2.15±0.36	4±0	0.7036±0.1255	0.5719±0
		5	2.15±0.52	5±0	0.7012±0.1380	0.5201±0
Credit	2	3	2.05±0.13	3±0	0.7582±0.0105	0.6745±0
		4	2.15±0.38	4±0	0.7351±0.0556	0.6238±0
		5	2.35±0.86	5±0	0.7039±0.1154	0.5663±0
Soybean	4	5	4.05±0.22	5±0	0.9052±0.0412	0.8052±0
		6	4.25±0.36	6±0	0.8563±0.0896	0.7210±0
		7	4.40±0.47	7±0	0.8028±0.1568	0.6502±0
Voting	2	3	2.0±0	3±0	0.7785±0.0002	0.6218±0
		4	2.0±0	4±0	0.7779±0.0012	0.5344±0
		5	2.05±0.22	5±0	0.7608±0.0215	0.4861±0
Iris	3	4	2.95±0.22	4±0	0.8106±0.0413	0.7549±0
		5	3.1±0.45	5±0	0.8209±0.0524	0.6438±0
		6	3.25±0.58	6±0	0.7754±0.1023	0.5686±0
Wine	3	4	3.1±0.45	4±0	0.8573±0.0106	0.7503±0
		5	3.25±0.55	5±0	0.8109±0.0721	0.6774±0
		6	3.25±0.58	6±0	0.7886±0.1103	0.5529±0

settings of the number of clusters can be evaluated by the following partition quality (PQ) index:

$$PQ = \begin{cases} \dfrac{\sum_{i=1}^{k^*}\sum_{j=1}^{k'}[p(i,j)^2 \cdot (p(i,j)/p(j))]}{\sum_{i=1}^{k^*} p(i)^2}, & \text{if } k' > 1 \\ 0, & \text{otherwise} \end{cases}$$

where k' is the number of clusters learned by the algorithm. The term $p(i, j)$ calculates the frequency-based probability that a data point is labeled i by the true label and labeled j by the obtained label. This PQ index achieves the maximum value 1 when the obtained labels induce the same partition as the true ones. That is, all data points in each cluster have the same true label and the estimated k' is equal to k^*. The learning

rate η in the RP-WOCIL algorithm was set at 0.0003 and the initialization-oriented method was adopted in the experiments. Table XIV summarizes the clustering results obtained by the RP-WOCIL and WOCIL algorithms in each situation. It can be observed that the RP-WOCIL algorithm has given a good estimation of the number of clusters and obtained better partition qualities. By contrast, for the WOCIL algorithm without cluster-number-learning mechanism, the obtained number of clusters was always equal to the preassigned one and the clustering performance degraded seriously when the initialized number of clusters was not appropriately chosen.

VII. Conclusion

In this paper, we have presented a new soft subspace clustering method, which is applicable to data with numerical, categorical, and mixed data. This method follows the learning model of object-cluster similarity-based clustering analysis. A unified weighting scheme for numerical and categorical attributes has been proposed, which quantifies the contributions of different attributes to the detection of various clusters with two factors, i.e., intercluster difference and intracluster similarity. Moreover, to solve the selection problem regarding the number of clusters, a rival penalized competitive learning mechanism has been introduced, which enables the number of clusters to be determined automatically during clustering process. In addition, a new initialization-oriented method has been proposed to improve the performance of k-means-type clustering methods on numerical, categorical, and mixed data sets. Experiments on benchmark data sets have shown the effectiveness of the proposed method in comparison with the existing algorithms.

References

[1] T. C. Silva and L. Zhao, "Network-based high level data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 954–970, Jun. 2012.

[2] Y. Wang, S. Chen, and Z.-H. Zhou, "New semi-supervised classification method based on modified cluster assumption," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 689–702, May 2012.

[3] J. Gui, T. Liu, D. Tao, Z. Sun, and T. Tan, "Representative vector machines: A unified framework for classical classifiers," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1877–1888, Aug. 2016, doi: 10.1109/TCYB.2015.2457234.

[4] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[5] Y.-M. Cheung, "k-means: A new generalized k-means clustering algorithm," *Pattern Recognit. Lett.*, vol. 24, no. 15, pp. 2883–2893, Aug. 2003.

[6] H. Zeng and Y.-M. Cheung, "Learning a mixture model for clustering with the completed likelihood minimum message length criterion," *Pattern Recognit.*, vol. 47, no. 5, pp. 2011–2030, 2014.

[7] H. Jia, Y.-M. Cheung, and J. Liu, "Cooperative and penalized competitive learning with application to kernel-based clustering," *Pattern Recognit.*, vol. 47, no. 9, pp. 3060–3069, 2014.

[8] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 673–690, Jul./Aug. 2002.

[9] C.-C. Hsu, "Generalizing self-organizing map for categorical data," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 294–304, Mar. 2006.

[10] P. Cheeseman and J. Stutz, "Bayesian classification (autoclass): Theory and results," in *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996.

[11] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific–Asia Conf. Knowl. Discovery Data Mining*, 1997, pp. 21–34.

[12] Z. He, X. Xu, and S. Deng, "Scalable algorithms for clustering large datasets with mixed type attributes," *Int. J. Intell. Syst.*, vol. 20, no. 10, pp. 1077–1089, 2005.

[13] H. Luo, F. Kong, and Y. Li, "Clustering mixed data based on evidence accumulation," in *Advanced Data Mining and Applications* (Lecture Notes in Computer Science), vol. 4093, X. Li, O. R. Zaiane, and Z. Li, Eds. New York, NY, USA: Springer, 2006, pp. 348–355.

[14] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, 2007.

[15] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," *Pattern Recognit.*, vol. 46, no. 8, pp. 2228–2238, 2013.

[16] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 90–105, Jun. 2004.

[17] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[18] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[19] J. Gui, S.-L. Wang, and Y.-K. Lei, "Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data," *Artif. Intell. Med.*, vol. 50, no. 3, pp. 181–191, 2010.

[20] H. Zeng and Y.-M. Cheung, "Feature selection and kernel learning for local learning-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, Aug. 2011.

[21] J. Gui, Z. Sun, J. Cheng, S. Ji, and X. Wu, "How to estimate the regularization parameter for spectral regression discriminant analysis and its kernel version?" *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 211–223, Feb. 2014.

[22] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1998, pp. 94–105.

[23] C.-H. Cheng, A. W. Fu, and Y. Zhang, "Entropy-based subspace clustering for mining numerical data," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Data Mining*, 1999, pp. 84–93.

[24] S. Goil, H. Nagesh, and A. Choudhary, "MAFIA: Efficient and scalable subspace clustering for very large data sets," Northwest Univ., Kirkland, WA, USA, Tech. Rep. CPDC-TR-9906-010, 1999.

[25] C. C. Aggarwal, J. L. Wolf, P. S. Yu, and J. S. Park, "Fast algorithms for projected clustering," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1999, pp. 61–72.

[26] C. C. Aggarwal and P. S. Yu, "Finding generalized projected clusters in high dimensional spaces," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 70–81.

[27] K. Woo and J. Lee, "FINDIT: A fast and intelligent subspace clustering algorithm using dimension voting," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Korea Adv. Inst. Sci. Technol., Daejeon, South Korea, 2002.

[28] J. Yang, W. Wang, H. Wang, and P. Yu, "δ-clusters: Capturing subspace correlation in a large data set," in *Proc. 18th Int. Conf. Data Eng.*, 2002, pp. 517–528.

[29] G. Liu, J. Li, K. Sim, and L. Wong, "Distance based subspace clustering with flexible dimension partitioning," in *Proc. 23rd Int. Conf. Data Eng.*, 2007, pp. 1250–1254.

[30] D. S. Modha and W. S. Spangler, "Feature weighting in k-means clustering," *Mach. Learn.*, vol. 52, no. 3, pp. 217–237, 2003.

[31] H. Frigui and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognit.*, vol. 37, no. 3, pp. 567–581, 2004.

[32] H. Frigui and O. Nasraoui, "Simultaneous clustering and dynamic keyword weighting for text documents," in *Survey of Text Mining*, M. Berry, Ed. New York, NY, USA: Springer, 2004, pp. 45–70.

[33] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma, "Subspace clustering of high dimensional data," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 517–521.

[34] J. H. Friedman and J. J. Meulman, "Clustering objects on subsets of attributes," *J. Roy. Statist. Soc. B*, vol. 66, no. 4, pp. 815–849, Nov. 2004.

[35] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007.

[36] G. Gan and J. Wu, "Subspace clustering for high dimensional categorical data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 2, pp. 87–94, 2004.

[37] M. Kim and R. S. Ramakrishna, "Projected clustering for categorical datasets," *Pattern Recognit. Lett.*, vol. 27, no. 12, pp. 1405–1417, Sep. 2006.

[38] G. Gan, J. Wu, and Z. Yang, "PARTCAT: A subspace clustering algorithm for high dimensional categorical data," in *Proc. Int. Joint Conf. Neural Netw.*, 2006, pp. 16–21.

[39] M. J. Zaki, M. Peters, I. Assent, and T. Seidl, "CLICKS: An effective algorithm for mining subspace clusters in categorical datasets," *Data Knowl. Eng.*, vol. 60, no. 1, pp. 51–70, Jan. 2007.

[40] L. Bai, J. Liang, C. Dang, and F. Cao, "A novel attribute weighting algorithm for clustering high-dimensional categorical data," *Pattern Recognit.*, vol. 44, no. 12, pp. 2843–2861, Dec. 2011.

[41] F. Cao, J. Liang, D. Li, and X. Zhao, "A weighting k-modes algorithm for subspace clustering of categorical data," *Neurocomputing*, vol. 108, pp. 23–30, May 2013.

[42] L. Chen and S. Wang, "Central clustering of categorical data with automated feature weighting," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1260–1266.

[43] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[44] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Proc. 17th Annu. Conf. Neural Inf. Process. Syst.*, 2003, pp. 281–288.

[45] C. A. Sugar and G. M. James, "Finding the number of clusters in a dataset: An information-theoretic approach," *J. Amer. Statist. Assoc.*, vol. 98, no. 463, pp. 750–763, Sep. 2003.

[46] Y.-M. Cheung, "Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 750–761, Jun. 2005.

[47] L. Wang, C. Leckie, K. Ramamohanarao, and J. Bezdek, "Automatically determining the number of clusters in unlabeled data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 3, pp. 335–350, Mar. 2009.

[48] H.-Y. Liao and M. K. Ng, "Categorical data clustering with automatic selection of cluster number," *Fuzzy Inf. Eng.*, vol. 1, no. 1, pp. 5–25, 2009.

[49] H. Zeng and Y.-M. Cheung, "A new feature selection method for Gaussian mixture clustering," *Pattern Recognit.*, vol. 42, no. 2, pp. 243–250, 2009.

[50] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. Evanston, IL, USA: Routledge, 2003.

[51] W. Chen, Y. Chen, Y. Mao, and B. Guo, "Density-based logistic regression," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, 2013, pp. 140–148.

[52] K. Zhang *et al.*, "From categorical to numerical: Multiple transitive distance learning and embedding," in *Proc. SIAM Int. Conf. Data Mining*, Vancouver, BC, Canada, 2015, pp. 1–9.

[53] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Inf. Syst.*, vol. 25, no. 5, pp. 345–366, 2000.

[54] M. J. Zaki and M. Peters, "CLICK: Mining subspace clusters in categorical data via k-partite maximal cliques," in *Proc. 21st Int. Conf. Data Eng.*, 2005, pp. 355–356.

[55] D. Barbará, J. Couto, and Y. Li, "COOLCAT: An entropy-based algorithm for categorical clustering," in *Proc. 11th ACM Conf. Inf. Knowl. Manage.*, 2002, pp. 582–589.

[56] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik, "LIMBO: Scalable clustering of categorical data," in *Proc. 9th Int. Conf. Extending Database Technol.*, 2004, pp. 123–146.

[57] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control Comput.*, 1999, pp. 368–377.

[58] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 22, no. 4, pp. 882–907, 1966.

[59] P. Blomstedt, J. Tang, J. Xiong, C. Granlund, and J. Corander, "A Bayesian predictive model for clustering data of mixed discrete and continuous type," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 489–498, Mar. 2015.

[60] R. S. Sangam and H. Om, "Hybrid data labeling algorithm for clustering large mixed type data," *J. Intell. Inf. Syst.*, vol. 45, no. 2, pp. 273–293, Oct. 2015.

[61] S. Q. Le and T. B. Ho, "An association-based dissimilarity measure for categorical data," *Pattern Recognit. Lett.*, vol. 26, no. 16, pp. 2549–2557, 2005.

[62] D. Ienco, R. G. Pensa, and R. Meo, "From context to distance: Learning dissimilarity for categorical data clustering," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–25, Mar. 2012.

[63] H. Jia, Y.-M. Cheung, and J. Liu, "A new distance metric for unsupervised learning of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1065–1079, May 2016.

[64] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 1, Mar. 2009, Art. no. 1.

[65] H. S. Nagesh, S. Goil, and A. Choudhary, "Adaptive grids for clustering massive data sets," in *Proc. SIAM Int. Conf. Data Mining*, 2001, pp. 1–17.

[66] K. Kailing, H.-P. Kriegel, and P. Kröger, "Density-connected subspace clustering for high-dimensional data," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 246–257.

[67] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.

[68] Y. H. Chu, Y. J. Chen, D. N. Yang, and M. S. Chen, "Reducing redundancy in subspace clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 10, pp. 1432–1446, Oct. 2009.

[69] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[70] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 657–668, May 2005.

[71] Z. Deng, K.-S. Choi, F.-L. Chung, and S. Wang, "Enhanced soft subspace clustering integrating within-cluster and between-cluster information," *Pattern Recognit.*, vol. 43, no. 3, pp. 767–781, 2010.

[72] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos, "Locally adaptive metrics for clustering high dimensional data," *Data Mining Knowl. Discovery*, vol. 14, no. 1, pp. 63–97, Feb. 2007.

[73] E. Y. Chan, W. K. Ching, M. K. Ng, and J. Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures," *Pattern Recognit.*, vol. 37, no. 5, pp. 943–952, 2004.

[74] F. Esposito, D. Malerba, V. Tamma, and H.-H. Bock, "Classical resemblance measures," in *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information From Complex Data*, H. H. Bock and E. Diday, Eds. Berlin, Germany: Springer, 2002, pp. 139–152.

[75] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," in *Proc. SIGMOD Workshop Res. Issues Data Mining Knowl. Discovery*, 1997, pp. 1–8.

[76] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.

[77] J. Oosterhoff and W. R. van Zwet, "A note on contiguity and Hellinger distance," in *Selected Works in Probability and Statistics*. New York, NY, USA: Springer, 2011, pp. 63–72.

[78] A. Bhattacharya, *Fundamentals of Database Indexing and Searching*. Boca Raton, FL, USA: CRC Press, 2014.

[79] S. C. Ahalt, A. K. Krishnamurthy, P. Chen, and D. E. Melton, "Competitive learning algorithms for vector quantization," *Neural Netw.*, vol. 3, no. 3, pp. 277–290, 1990.

[80] J. M. Pena, J. A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the k-means algorithm," *Pattern Recognit. Lett.*, vol. 20, no. 10, pp. 1027–1040, 1999.

[81] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, Feb. 2003.

[82] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-means clustering," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1293–1302, 2004.

[83] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[84] S. S. Khan and S. Kant, "Computation of initial modes for K-modes clustering algorithm using evidence accumulation," in *Proc. 20th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2007, pp. 2784–2789.

[85] F. Cao, J. Liang, and L. Bai, "A new initialization method for categorical data clustering," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10223–10228, 2009.

[86] R. E. Higgs, K. G. Bemis, I. A. Watson, and J. H. Wikel, "Experimental designs for selecting molecules from large chemical databases," *J. Chem. Inf. Comput. Sci.*, vol. 37, no. 5, pp. 861–870, 1997.

[87] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17. Vancouver, BC, Canada, Dec. 2005, pp. 507–514.

Hong Jia received the M.Sc. degree from the School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2013.

She was a Post-Doctoral Research Fellow with the Department of Computer Science, Hong Kong Baptist University. She is currently an Assistant Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China. Her current research interests include machine learning, data mining, and pattern recognition.

Yiu-Ming Cheung (SM'06) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, visual computing, and optimization.

Prof. Cheung is an IET Fellow, BCS Fellow, and IETI Fellow. He is the Founding Chairman of the Computational Intelligence Chapter, IEEE Hong Kong Section. He also serves as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, *Knowledge and Information Systems*, and the *International Journal of Pattern Recognition and Artificial Intelligence*.