Applied Mathematics and Computation 335 (2018) 264–277

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

A sequential ensemble clusterings generation algorithm for mixed data

Xingwang Zhao, Fuyuan Cao, Jiye Liang*

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China

ARTICLE INFO

Keywords:
Ensemble clustering
Base clustering
Mixed data
Information entropy

ABSTRACT

Ensemble clustering has attracted much attention for its robustness, stability, and accuracy in academic and industry communities. In order to yield base clusterings with high quality and diversity simultaneously in ensemble clustering, many efforts have been done by exploiting different clustering models and data information. However, these methods neglect correlation between different base clusterings during the process of base clusterings generation, which is important to obtain a quality and diverse clustering decision. To overcome this deficiency, a sequential ensemble clusterings generation algorithm for mixed data is developed in this paper based on information entropy. The first high quality base clustering is yield by maximizing the entropy-based criterion. Afterward, a sequential paradigm is utilized to incrementally find more base clusterings, in which the diversity between a new base clustering and the former base partitions is measured by the normalized mutual information. Extensive experiments conducted on various data sets have demonstrated the superiority of our proposal as compared to several existing base clusterings generation algorithms.

1. Introduction

Clustering analysis is one of the primary techniques in data mining and machine learning. Its aim is to partition a set of unlabeled objects into several distinct clusters so that the data objects in the same cluster are similar and dissimilar to the data objects in other clusters. It has numerous applications in such areas as customer segmentation, target marketing, bioinformatics, social network analysis, and scientific data analysis [1–4].

In real applications, analyzed data sets are often comprised of mixed numerical and categorical attributes, particularly when they are merged from different sources. In other words, data are in a mixed mode. Such data may be encountered, for instance, in medical diagnosis analyses, in the analysis of survey data, as well as in image analysis. For example, the attributes of the data about medical diagnosis may include sex, age, weight, and blood pressure of the patients, where the attribute sex is categorical, the other attributes are numerical. In the past five decades, various clustering algorithms have been proposed in the literature [2–5]. However, the primary focus of these clustering algorithms has been on the data sets with either numerical attributes or categorical attributes. It is difficult to apply traditional clustering algorithm directly into mixed data. Therefore, mixed data clustering becomes not only a difficult task but also a challenging and promising one to attract many researchers in data mining and machine learning field.

* Corresponding author.
E-mail addresses: zhaoxw84@163.com (X. Zhao), cfy@sxu.edu.cn (F. Cao), ljy@sxu.edu.cn (J. Liang).

https://doi.org/10.1016/j.amc.2018.04.035
0096-3003/

Currently, in clustering analysis, there are usually two categories of methods to process mixed data. One category is to transform either categorical attributes into numerical attributes or numerical attributes into categorical attributes. Then, the clustering methods for numerical data or categorical data can be used. However, these methods are not effective since the similarity measure of the transformed data could not represent the similarity of original mixed data. The other category is to extend the clustering algorithms for numerical data or categorical data to match with mixed data to improve the clustering result. Using these two strategies, some clustering algorithms for mixed data have been developed in the literature [7–11].

Although there are many mixed data clustering algorithms, Kuncheva et al.[12] pointed out that there is no single clustering algorithm which performs best for all data sets and can discover all types of clusters and structures. Each algorithm has its own strength and weakness. For a given mixed data set, different clustering algorithms, or even the same algorithm with different parameters, usually obtain distinct clustering results. Therefore, it is difficult for users to decide which algorithm would be a proper choice for clustering the given data set. To overcome these limitations, ensemble clustering algorithms have recently emerged as a powerful alternative to standard clustering algorithms. Their main objective is to improve the robustness as well as the quality of clustering results, by combining different clustering decisions according to some criterion. Generating a set of base clusterings is a key process in ensemble clustering [13,14]. Examples of well-known ensemble clustering generation algorithms include running a single clustering algorithm with different initialization [16–18], carrying out one or more clustering algorithms on different subspaces or subsamples of a given data set [19–21], and performing different clustering algorithms [22,27,28].

Despite notable success, these algorithms generate the different base clustering results independently. The correlation between different base clusterings during the process of base clusterings generation is neglected, which is important to obtain a quality and diverse base clustering decision. To overcome this deficiency, a sequential ensemble clusterings generation algorithm is developed in this paper based on information entropy for mixed data. The first high quality base clustering is yield by maximizing the entropy-based criterion. Afterward, a sequential paradigm is utilized to incrementally find more base clusterings, in which the diversity between a new base clustering and the former base partitions is measured by the normalized mutual information. Extensive experiments conducted on various data sets have demonstrated the superiority of our proposal as compared to several existing base clusterings generation algorithms.

The rest of this paper is organized as follows: Section 2 reviews the related work on mixed data clustering and ensemble clustering problem. The proposed sequential ensemble clusterings generation algorithm is introduced in Section 3. Then, Section 4 exhibits the evaluation of this new algorithm against other ensemble clusterings generation algorithms over real data sets. The paper is concluded in Section 5.

2. Related work

In this section, mixed data clustering algorithms and some recent developments on ensemble clustering are reviewed.

2.1. Mixed data clustering

Data sets analyzed in practice are commonly characterized by mixed numerical and categorical attributes. One of the most common approaches to cluster mixed data involves converting the data set to a single data type, and applying standard clustering algorithms to the transformed data. For example, He et al. [6] considered a numerical attribute as a category by discretization. Then they extended their earlier clustering algorithm of categorical data to cluster mixed data.

An alternative approach is to design a generalized similarity or distance measure for mixed data, and apply it to the existing clustering algorithms. K-prototype [7] is one of the most famous algorithms. It integrates the k-means and the k-modes algorithms by defining a combined dissimilarity measure to enable clustering of mixed numerical and categorical attributes. Ahmad and Dey [8] proposed a distance metric for mixed data clustering based on the co-occurrence likelihood of two categorical attribute values. Li and Biswas. [9] presented an agglomerative hierarchical clustering algorithm based on Goodall similarity measure for mixed data. Hsu et al. [10] proposed a mixed data clustering algorithm applying the idea of distance hierarchy to calculate distance for every categorical attribute. This algorithm, however, requires domain-specific knowledge to build distance hierarchy which is not available for a large number of attribute domains. Liang et al. [11] proposed an algorithm to cluster mixed data by defining two kinds of information entropy measures for numerical and categorical data, respectively. Gower [29] introduced a similarity index that measures the similarity between two mixed data. And it is used to cluster mixed data in the framework of the k-means type algorithm.

Additionally, some mixed data clustering algorithms based on statistical models are developed recently, which typically assume the observations follow a normal-multinomial finite mixture model. Readers with interests can refer to the survey paper for more comprehensive understanding [30].

2.2. Ensemble clustering

Like ensemble methods in supervised learning, ensemble clustering methods work in two steps, clustering generation and clustering combination. The quality and diversity of the base clusterings are two major factors, which affect the performance of an ensemble clustering method. As a result, several heuristics have been proposed to generate different clusterings for a given data set, which can be classified into three categories:

- *Homogeneous methods.* Base clusterings are generated by repeatedly running a single clustering algorithm with different initializations, such as the number of clusters or cluster centers [16–18].
- *Data subspaces/subsamples methods.* A set of base clustering results are obtained by projecting data onto different subspaces, choosing different subsets of features, or data sampling [19–21,31].
- *Heterogeneous methods.* Base partitions are produced with a number of different clustering algorithms on a given data set [22,28].

In the second step, given a set of base clusterings, a consensus function is used to combine them into the final clustering results. During the past decade, many clustering ensemble methods have been proposed. Roughly speaking, these methods can be classified into the following four categories:

- *Feature-based methods.* These methods transform the problem of clustering combination to categorical data clustering [6,19].
- *Similarity-based methods.* These methods express pairwise similarity among data points as similarity matrices, on which any similarity-based clustering algorithms can be used to obtained a final clustering [16,32].
- *Graph-based methods.* These methods describe the base clustering information as an undirected graph and then derives the ensemble clustering via graph partitioning [20,33,34].
- *Relabeling-based methods.* These methods express the base clustering information as label vectors and then aggregates via label alignment [22,35].

In recent years, the increasing size and complexity of data sets have made most of above mentioned ensemble clustering algorithms unworkable. In order to deal with this problem, various novel ensemble clustering technologies have emerged. For example, ensemble selection clustering methods improve the clustering quality by evaluating and selecting a subset of base partitions according to the contribution of base clusterings in the integration process [14,15]. Semi-supervised ensemble clustering methods use some prior knowledge of the data sets provided by experts in the consensus functions [23,24]. Structure ensemble, firstly proposed by Yu et al. [25,26], can integrate multiple cluster structures extracted from different base clusterings into a unified structure for large-scale data.

This paper mainly focuses on the research of base clusterings generation strategy. Different from the above methods, we will take the correlation between different base clusterings into account in the process of base clusterings generation.

3. Proposed ensemble clusterings generation algorithm

3.1. Problem formulation

Suppose that $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is a set of N objects. Each object $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$ is characterized by m attributes or features, e.g. $A = \{A_1^r, A_2^r, \ldots, A_p^r, A_1^c, A_2^c, \ldots, A_q^c\}$, where $p+q=m$, $\{A_1^r, A_2^r, \ldots, A_p^r\}$ represents p numerical attributes and $\{A_1^c, A_2^c, \ldots, A_q^c\}$ represents q categorical attributes. Therefore, a mixed data point $x_i \in \mathbf{X}$ can be expressed by a vector $x_i = (x_i^r, x_i^c)$, where $x_i^r = (x_{i,1}^r, x_{i,2}^r, \ldots, x_{i,p}^r)$ is the numerical part, $x_i^c = (x_{i,1}^c, x_{i,2}^c, \ldots, x_{i,q}^c)$ represents the values of categorical data.

Let $\Pi = \{\pi_1, \ldots, \pi_M\}$ be a cluster ensemble with M base clusterings for the given data set $\mathbf{X}$, each of which is also referred as an "ensemble member". Each base clustering consists of a set of clusters $\pi_g = \{C_{g,1}, C_{g,2}, \ldots, C_{g,k_g}\} (1 \le g \le M)$, such that $\bigcup_{j=1}^{k_g} C_{g,j} = X$, where k_g is the number of clusters in the gth base clustering. A final clustering solution $\pi^* = \{C_1, C_2, \ldots, C_k\}$ of the given data set $\mathbf{X}$ based on the base clusterings Π is formed using a consensus function.

The problem of sequential ensemble clusterings generation can be roughly stated as follows. Given a mixed data set $\mathbf{X}$ and a set of base clusterings $\Pi_s = \{\pi_1, \ldots, \pi_s\} (1 \le s < M)$, generate a new base clustering $\pi_{new} = \{C_{new,1}, C_{new,2}, \ldots, C_{new,k_{new}}\}$, such that $quality(\pi_{new})$ and $\sum_{i=1}^{s} diversity(\pi_i, \pi_{new})$ are simultaneously maximized. The task here corresponds to generating a new base clusterings with respect to the previous base clusterings, where the new clustering has high quality and the pairwise diversity between the new clustering and each exist base clusterings is high.

In view of the effectiveness of information entropy in clustering analysis [11,36–38], using it to study the quality and diversity of base clusterings provides a new way to sequential ensemble clusterings generation for mixed data. Entropy-based criterion can evaluate the orderliness of a given cluster. The quality of clustering result is naturally evaluated by the entropy of all clusters, namely, the expected entropy [38]. We expect that in a good clustering the objects in the same cluster will be similar, whereas dissimilar objects will be assigned to different clusters. This intuition is obtained by minimizing the expected entropy. The lower the expected entropy is, the higher quality of clustering results is. In ensemble clustering, suppose that one of the base clustering for the given data set $\mathbf{X}$ is $\pi = \{C_1, C_2, \ldots, C_k\}$, where k is the number of clusters and n_i represent the number of objects in C_i . Thus, $H(\mathbf{X})$ and $H(C_i)$ are used to represent the data set entropy and the ith cluster entropy, respectively. The entropy-based clustering criterion tries to find the optimal partition by maximizing the following entropy criterion:

$$O(\pi) = H(\mathbf{X}) - \frac{1}{N}\sum_{i=1}^{k} n_i H(C_i). \tag{1}$$

Since $H(\mathbf{X})$ is fixed for the given data set $\mathbf{X}$, maximizing $O(\pi)$ is equivalent to minimizing the item $E(\pi) = \frac{1}{N}\sum_{i=1}^{k} n_i H(C_i)$, which is named as the expected entropy of clustering π. Owing to the difference in data types, expected entropies for numerical data and categorical data will be introduced in the following, respectively.

About the diversity measure of base clustering results, there are several different measures in the literature. Because the normalized mutual information (NMI) has been shown to impact the clustering ensemble performance and it is easy to compute, it is usually used to measure diversity between base clusterings. The lower the NMI value, the higher is the diversity. Let $\pi_s = \{C_{s,1}, C_{s,2}, \ldots, C_{s,k_s}\}$ and $\pi_t = \{C_{t,1}, C_{t,2}, \ldots, C_{t,k_t}\}$ be two base clusterings for the data set $\mathbf{X}$, the NMI between them is given by:

$$NMI(\pi_s, \pi_t) = \frac{\sum_{i=1}^{k_s}\sum_{j=1}^{k_t} N_{ij} \log \frac{N \cdot N_{ij}}{N_i^s \cdot N_j^t}}{\sqrt{\sum_{i=1}^{k_s} N_i^s \log \frac{N_i^s}{N} \sum_{j=1}^{k_t} N_j^t \log \frac{N_j^t}{N}}}, \tag{2}$$

where N is the number of objects of the data set $\mathbf{X}$; N_{ij} is the number of common objects of clusters $C_{s,i}$ and $C_{t,j}$; N_i^s is the number of objects in cluster $C_{s,i}$; and N_j^t is the number of objects in cluster $C_{t,j}$.

At a high level, the objective function of sequential ensemble clusterings generation can be expressed as follows. Given a (possibly empty) set of base clusterings $\Pi_s = \{\pi_1, \ldots, \pi_s\}$ provided as background knowledge, generate the new base clustering $\pi_{new} = \{C_{new,1}, C_{new,2}, \ldots, C_{new,k_{new}}\}$, such that

$$\pi_{new} = arg\min_{\pi_t \in \mathcal{S}} \{(1-\lambda)E(\pi_t) + \lambda \sum_{i=1}^{s} NMI(\pi_t, \pi_i)\}, \tag{3}$$

where $\mathcal{S}$ represents the space of all possible clustering results of $\mathbf{X}$, when given the number of cluster k.

In the right part of the objective function, the first term is the expected entropy of a new base clustering results which is used to generate high quality clustering by minimization. The second term is the sum of normalized mutual information between a new clustering and the previously generated clusterings that we want to minimize. And it is used to ensure the diversity of the new clustering compared with the exist base clusterings. The λ is a tradeoff parameter, which is used to trade off the quality and diversity of the objective function. Note that, in the above objective function, when $\Pi_s = \emptyset$, generating the first base clustering only considers the cluster quality.

3.2. Expected entropy for numerical data

In order to compute the expected entropy for numerical data, kernel-based probability density functions estimations, such as Renyi entropy [11], are the most commonly used methods [39]. Their use is usually restricted to one- or two-dimensional probability density functions. They are not satisfactory in terms of dealing with high-dimensional problems. An additional difficulty in kernel based estimation lies in the choice of kernel function. Because there is no any prior knowledge about the cluster distribution, utilizing any one kernel function to describe the density is not always a good choice.

Recently, a new smooth estimator for the entropy evaluation, called MeanNN differential entropy estimator, is proposed in [40]. In this paper, we use it to computer expected entropy due to its smoothness with respect to the coordinates of data points. This estimator computes the entropy based on the pair-wise distances between all the given data points in one cluster. Suppose that $\pi^r = \{C_1^r, C_2^r, \ldots, C_k^r\}$ is a base clustering for numerical data $\mathbf{X^r}$ consisting of N objects which are described by p attributes. The information entropy for $C_i^r \in \pi^r$ is given as [40]:

$$RH(C_i^r) \approx \frac{p}{n_i^r(n_i^r - 1)} \sum_{x,y \in C_i^r, x \neq y} \log ||x - y||. \tag{4}$$

where n_i^r is the number of objects in C_i^r. In order to measure the quality consistently, the values of $RH(C_i^r)(C_i^r \in \pi^r)$ are normalized to [0,1] by $NRH(C_i^r) = 1 - \frac{1}{1+exp(RH(C_i^r))}$. Plugging Eq. (4) into the expected entropy function yields the following form of the quality measure:

$$RE(\pi^r) = \frac{1}{N}\sum_{i=1}^{k} NRH(C_i^r). \tag{5}$$

3.3. Expected entropy for categorical data

In a categorical domain, Liang et.al [41] used the complement entropy to measure information content and uncertainty for a categorical data table. Unlike the logarithmic behavior of Shannons entropy, the complement entropy can measure both uncertainty and fuzziness. Let $\mathbf{X^c}$ be a categorical data set with N objects described by q categorical attributes A^c. For any one attribute $a_j (1 \le j \le q)$, its domain D_{a_j} is defined as $D_{a_j} = \{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(v_j)}\}$, where v_j is the number of categories

of attribute a_j for $1 \le j \le q$. An object $\mathbf{x_i^c} \in \mathbf{X^c}$ can be represented as a vector $[x_{i,1}^c, x_{i,2}^c, \ldots, x_{i,q}^c]$, where $x_{i,j}^c \in D_{a_j}$, for $1 \le j \le q$. The complement entropy of the given data $\mathbf{X^c}$ with respect to attribute a_j is defined as:

$$LE(\mathbf{X^c}, \{a_j\}) = \sum_{l=1}^{v_j} \frac{W_{j,l}^{X^c}}{N}\left(1 - \frac{W_{j,l}^{X^c}}{N}\right), \tag{6}$$

where $W_{j,l}^{X^c} = |\{\mathbf{x_r^c} | x_{r,j}^c = a_j^l, \mathbf{x_r^c} \in \mathbf{X^c}\}|$ denotes the number of objects whose values are equal to domain value $a_j^l \in D_{a_j}$ for attribute a_j.

Based on the complement entropy, the expected entropy for categorical data is given in the following. Suppose that $\pi^c = \{C_1^c, C_2^c, \ldots, C_k^c\}$ is a base clustering for categorical data $\mathbf{X^c}$ consisting of N objects which are described by q attributes. The information entropy for $C_i^c \in \pi^c$ is given as:

$$CH(C_i^c) = \sum_{j=1}^{q} LE(C_i^c, \{a_j\}) = \sum_{j=1}^{q}\sum_{l=1}^{v_j} \frac{W_{j,l}^{C_i^c}}{n_i^c}\left(1 - \frac{W_{j,l}^{C_i^c}}{n_i^c}\right). \tag{7}$$

where $W_{j,l}^{C_i^c} = |\{\mathbf{x_r^c} | x_{r,j}^c = a_j^l, \mathbf{x_r^c} \in C_i^c\}|$ and n_i^c is the number of objects in C_i^c.

It is found that there is a quantitative relation between $CH(C_i^c)$ and $d(x, y)$ [11], i.e.,

$$CH(C_i^c) = \frac{1}{(n_i^c)^2}\sum_{j=1}^{q}\sum_{x,y \in C_i^c} d(x_j, y_j), \tag{8}$$

where $d(x_i, y_j) = \begin{cases} 0, & x_j = y_j, \\ 1, & x_j \ne y_j. \end{cases}$

The above derivation means that the within-cluster entropy can be expressed with the average dissimilarity between objects with in a cluster for categorical data.

In order to measure the quality consistently, the values of $CH(C_i^c)(C_i^c \in \pi^c)$ are normalized to [0,1] by $NCH(C_i^c) = \frac{n_i^c}{q(n_i^c - 1)} CH(C_i^c)$. So, the expected entropy of the base clustering $\pi^c = \{C_1^c, C_2^c, \ldots, C_k^c\}$ for categorical data is given as follows:

$$CE(\pi^c) = \frac{1}{N}\sum_{i=1}^{k} n_i^c NCH(C_i^c). \tag{9}$$

By integrating the $E(\pi^r)$ and $E(\pi^c)$ together, the expected entropy of a base clustering $\pi = \{C_1, C_2, \ldots, C_k\}$ for mixed data can be calculated as follows:

$$E(\pi) = \frac{p}{p+q} RE(\pi^r) + \frac{q}{p+q} CE(\pi^c). \tag{10}$$

With this measure of entropy, our objective in (2) becomes:

$$\pi_{new} = arg\min_{\pi_t \in \mathcal{S}} \{(1-\lambda)E(\pi_t) + \lambda \sum_{i=1}^{s} NMI(\pi_t, \pi_i)\}. \tag{11}$$

3.4. Algorithm description

To optimize the objective function mentioned in Eq. (11), we can apply an iterative cluster-and-re-cluster algorithm creating different base clusterings. The iterative clustering algorithm starts with a random partition of the given data into clusters. Then, it goes over all the data objects in a cyclical manner and for each object checks whether moving it from its current cluster to another one decreases the objective function. This loop may be iterated until either there is no possible single membership change that decreases the objective function or the local decrease of the score function become sufficiently small. In order to give a simple algorithm description, some notations are defined in the following:

Now suppose that we are considering moving a data object $x \in \mathbf{X}$ for its current cluster C_i to cluster C_j when generating the base clustering π. If π is the first base clustering, only the expected entropy of the objective function is affected in this operation. The change of the objective value in the Eq. (11) is

$$\triangle E(\pi | x, C_i \to C_j) = \frac{p}{p+q} \triangle RE(\pi^r | x, C_i^r \to C_j^r) + \frac{q}{p+q} \triangle CE(\pi^c | x, C_i^c \to C_j^c), \tag{12}$$

where $\triangle RE(\pi^r | x, C_i^r \to C_j^r) = \frac{1}{N}\{NRH(C_j^r \bigcup \{x\})(n_j^r + 1) + NRH(C_i^r \backslash \{x\})(n_i^r - 1) - (NRH(C_j^r)n_j^r - (NRH(C_i^r)n_i^r\}$ and $\triangle CE(\pi^c | x, C_i^c \to C_j^c) = \frac{1}{N}\{NCH(C_j^c \bigcup \{x\})(n_j^c + 1) + NCH(C_i^c \backslash \{x\})(n_i^c - 1) - (NCH(C_j^c)n_j^c - (NCH(C_i^c)n_i^c\}$

If there exists a set of base clustering $\Pi_s = \{\pi_1, \ldots, \pi_s\}(\Pi_s \ne \emptyset)$, when moving a data object $x \in \mathbf{X}$ for its current cluster C_i to cluster C_j in the process of generating a new base clustering π, the objective function value will change as follows:

$$\triangle F(\pi|x, C_i \to C_j) = (1-\lambda) \triangle E(\pi|x, C_i \to C_j) + \lambda \sum_{i=1}^{s} \triangle I(\pi, \pi_i) \tag{13}$$

, where $\triangle I(\pi, \pi_i)$ means the changing value of the $NMI(\pi, \pi_i)$ after moving the object x to cluster C_j form cluster C_i.

Suppose a clustering result $\pi = \{C_1, C_2, \ldots, C_k\}$ of the given data **X**, the potential target cluster for the object x is $C_l = arg\min_{C\in\pi} \triangle E(\pi|x, member(x) \to C)$ or $C_l = arg\min_{C\in\pi} \triangle F(\pi|x, member(x) \to C)$, where $member(x)$ means the cluster that the object x belongs to. This means that the membership re-assignment should result in the largest objective decrease. Based on the above mentioned formulations and notations, the developed **s**equential **e**nsemble **c**lusterings **g**eneration algorithm for mixed data (abbreviated as SECG) is shown in Algorithm 1.

Algorithm 1 The SECG Algorithm.

Input:
  X: a mixed data set; k: the number of clusters in base clusterings;
  M: the number of base clusterings to be generated.
  λ: the tradeoff parameter
Output:
  $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$: the generated base clusterings.
Method:
$\Pi = \varnothing$;
repeat
  t=1;
  Generate an initial clustering result $\pi_t = \{C_1, C_2, \ldots, C_k\}$ randomly;
  $continue \leftarrow TRUE$
  while *continue* **do**
    $continue \leftarrow FALSE$
    for every $x \in X$ **do**
      if t ==1 **then**
        $C' = arg\min_{C\in\pi_t} \triangle E(\pi_t|x, member(x) \to C)$;
        G = $\triangle E(\pi_t|x, member(x) \to C')$;
      else
        $C' = arg\min_{C\in\pi_t} \triangle F(\pi_t|x, member(x) \to C)$;
        G = $\triangle F(\pi_t|x, member(x) \to C')$;
      end if
      if $G < 0$ **then**
        Move the object x to the cluster C' from its current cluster;
        $continue \leftarrow TRUE$;
      end if
    end for
  end while
  Update $\Pi = \Pi \bigcup \{\pi_t\}$;
  t = t+1;
until $|\Pi| = M$;
return Π.

The proposed sequential base clusterings generating algorithm starts with a random partition of the data points into clusters. Then, it goes over all the points in a cyclical manner and for each point checks whether moving it from its current cluster to another one decreases the objective function. During the process of generating one base clustering, to find the cluster re-assignment of an object, we need to compute the updated entropy of each cluster and NMI value after adding this object to that cluster. To do so, the distance of this object to all the other members of that cluster is needed to calculate. Hence the complexity of reassigning one object to a new cluster is $O(N)$ and the computational complexity of this process is $O(TN^2)$, where N is the number of objects for data **X** and T is the number of iterations. In order to generate M base clusterings, the overall computational complexity of the algorithm is $O(MTN^2)$.

4. Experimental analysis

This section presents the effectiveness evaluation of the proposed algorithm over 8 real-world data sets in terms of some benchmark evaluation criteria.

Table 1
Characteristics of the mixed data sets.

Data sets	# objects	# numerical attributes	# categorical attributes	# classes
TAE	151	1	4	3
Flag	194	10	18	8
SHeart	270	7	6	2
CHeart	303	5	8	2
Credit	690	6	8	2
GCredit	1000	7	13	2
CMC	1473	2	7	3
Adult	44842	6	8	2

4.1. Data sets

The characteristics of the mixed data sets are shown in Table 1. These data sets are downloaded from the UCI machine learning repository [42]. Note that all these data sets are labeled and contain supervised class information. However, the class labels were not used in the processes of base clusterings generation and only used in evaluating the final clustering results.

4.2. Evaluation criteria

In order to give comprehensive results, three popular external criteria are used to evaluate the effectiveness of the clustering algorithms. They are Clustering Accuracy (CA), Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI), which measure the agreement of the clustering results produced by an algorithm and the ground truth.

Suppose that $C = \{c_1, c_2, \ldots, c_k\}$ and $P = \{p_1, p_2, \ldots, p_{k'}\}$ represent the clustering results and pre-defined classes of the data set with N objects, respectively. k and k' are the number of clusters C and classes P; $N_{i,j}$ is the number of common objects of cluster c_i and pre-defined class p_j; N_i^c is the number of data points in cluster c_i ; and N_j^p is the number of data points in class p_j. Then the three popular external criteria are given as follows:

• *Clustering Accuracy (CA).* CA measures the percentage of correctly classified data points in the clustering solution compared to pre-defined class labels. The CA is defined as:

$$CA = \frac{\sum_{i=1}^{k} max_{j=1}^{k'} N_{i,j}}{N}. \tag{14}$$

• *Normalized Mutual Information (NMI).* This is one of the common external clustering validation metrics that estimate the quality of the clustering with respect to a given class labels of the data. More formally, NMI can effectively measure the amount of statistical information shared by random variables representing the cluster assignments and the pre-defined label assignments of the objects. Thus, NMI is defined and computed according to the following formula:

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'} N_{i,j} log\frac{N \cdot N_{i,j}}{N_i^c \cdot N_j^p}}{\sqrt{\sum_{i=1}^{k} N_i^c \cdot log\frac{N_i^c}{N} \cdot \sum_{j=1}^{k'} N_j^p \cdot log\frac{N_j^p}{N}}}. \tag{15}$$

• *Adjusted Rand Index (ARI).* ARI takes into account the number of objects that exist in the same cluster and different clusters [43]. The ARI is defined as:

$$ARI = \frac{\binom{N}{2}\sum_{i=1}^{k}\sum_{j=1}^{k'}\binom{N_{i,j}}{2} - \left[\sum_{i=1}^{k}\binom{N_i^c}{2}\sum_{j=1}^{k'}\binom{N_j^p}{2}\right]}{\frac{1}{2}\binom{N}{2}\left[\sum_{i=1}^{k}\binom{N_i^c}{2} + \sum_{j=1}^{k'}\binom{N_j^p}{2}\right] - \left[\sum_{i=1}^{k}\binom{N_i^c}{2}\sum_{j=1}^{k'}\binom{N_j^p}{2}\right]}. \tag{16}$$

The maximum value of the three external criteria is 1. If the clustering result is close to the true class distribution, then the values of them are high. The higher the values of the three measures for a clustering result, the better the clustering performance is.

4.3. Experimental setups

To fully investigate the performance of the proposed SECG algorithm, it is compared with a number of the state-of-the-art base clusterings generation algorithms. Details of these compared algorithms are described in the following:

- Full-space based base clusterings generation algorithm (Fullspace): In this kind algorithm, base clusterings are created using repeated runs of the modified k-prototypes algorithm [11] on the full features space of the given mixed data set, with different initial cluster centers.

- Subspace based base clusterings generation algorithm (Subspace): In this kind algorithm, some mixed data set with different subspaces are used to generate multiple ensemble members with the modified k-prototypes algorithm [11]. Each data subspace X' is generated by firstly defining $m' = m'_{min} + \lfloor \alpha(m'_{max} - m'_{min}) \rfloor$; where $\alpha \in [0, 1]$ is a uniform random variable, d'_{min} and d'_{max} are the lower and upper bounds of the generated subspace, respectively. They are set to $0.75m$ and $0.85m$. An attribute is selected one by one from the pool of m attributes, until the collection of m' is obtained.
- Subsample based base clusterings generation algorithm(Subsample): This kind algorithm produces each base clustering with a data subset that contains randomly selected 20% of original data objects. Firstly, the sampled data are gathered into different clusters with the modified k-prototypes algorithm [11]. Then, the out-of-sample data objects obtain their cluster labels utilizing the nearest neighbor labeling technique based on the partial clustering results of the sampled data.
- Random base clusterings generation algorithm(Random): For generating the base clusterings, a simply random partition algorithm is performed on the given mixed data set.

Other related setups of experimental analysis are described in the following:

- For the modified k-prototypes algorithm [11], the number of clusters k is set equal to the true number of clusters. And the initial cluster centers are different.
- In the aggregation of base clustering results, two types of consensus functions will be used in our experiments: the co-association similarities based consensus functions [44] and the graph based consensus functions. The first type firstly constructs an $N \times N$ similarity matrix between each pair of objects, which is been computed based on the number of objects shared in the base partitions. Next, based on this co-association similarity, three agglomerative clustering methods, namely single-link (SL), complete-link (AL), and average-link (CL) [3] are used to generate the final partition. In the second type, the three consensus methods: Cluster-based Similarity Partitioning Algorithm (CSPA), the Hyper-Graph Partitioning Algorithm (HGPA), and the Meta-CLustering Algorithm (MCLA) [20] are used in our experiments.
- In all the experiments, unless otherwise mentioned, we set the size of base clustering $M = 20$ and the tradeoff parameter $\lambda = 0.5$.
- The reported experimental results are the average values with 10 runs.
- The proposed algorithm and the compared algorithms were implemented in the MATLAB computing environment and all experiments were conducted on a workstation with Intel Xeon CPU E5-2650@2.60 GHz and 128 GB RAM.

Therefore, the combinations of 5 kind algorithms for generating the base clusterings, and 6 consensus functions result in 30 kinds of ensemble clustering results.

4.4. Results on effectiveness analysis

In this subsection, we focus on the clustering performance of different base clustering generation algorithms on above mentioned 8 mixed data sets with three evaluation criteria. And the related statistical tests are carried out.

Tables 2–4 show the experiment results. For each data set, the rank values of different base clustering generation algorithms using each consensus function are calculated. It ranks the algorithms for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2..., as shown in the parentheses. When the evaluation indices are tied, the average ranks are assigned. Based on the performances on different data sets, the average ranks of each base clustering generating algorithm with the same consensus function are calculated in the bottom. In addition, the values of the best performance for each data set are highlighted in boldface. Firstly, the clustering accuracies (CA) and ranks of the compared base clustering generating algorithms with different consensus functions are listed in Table 2. According to the average ranks, we find that the proposed SECG algorithm always outperforms other algorithms. From Table 2, the SECG algorithm achieves the best performance on 5 of the 8 data sets with SL, CL and AL consensus functions. This superiority is more evident for the TAE and Credit data sets. In addition, these results indicate that the quality of clustering results produced by the random base clustering generating algorithm are worse than those obtained by the other algorithms. Surprisingly, the CA values of clustering results producing by different base clusterings generating algorithms on GCredit and Adult data sets are the same. It may be due to the following two reasons. On the one hand, the clustering structures of these two mixed data sets revealed by the different algorithms are similar. On the other hand, the CA index has weak differentiation ability. Similar experimental results with these algorithms are observed using NMI and ARI evaluation indices in Tables 3–4. From these two tables, we can see that the advantages of the proposed algorithm are more obvious. In addition, in term of NMI and ARI evaluation indices, the proposed SECG algorithm is better than that of other algorithms on GCredit and Adult data sets. In general, the proposed SECG algorithm is the most suitable for generating base clusterings when compared with most of the existing algorithms.

In order to give a comprehensive comparison, we further perform the Friedman test and Nemenyi test [45] to analyze the differences between the proposed SECG algorithm and the other algorithms. For Friedman test, there are $A = 5$ algorithms, $B = 18$ cases (i.e., 3 evaluation indices, 6 consensus functions). Let r_i^j be the rank of the jth of the A algorithms on the ith of the B cases. For example, according to the average rank values in Table 2, the proposed SECG algorithm ranks 1.88 under the single-link (SL) consensus function with respect to CA. The Friedman test compares the average ranks of algorithms for all the cases, $R_j = (\frac{1}{B}) \sum_{i=1}^{B} r_i^j$ representing the average rank of the jth algorithm for all the cases, where B is the number of

Table 2
Results of CA values for the compared base clusterings generation algorithms.

Data sets	SL					CL				
	SECG	Fullspace	Subspace	Subsample	Random	SECG	Fullspace	Subspace	Subsample	Random
TAE	**0.3899(1)**	0.3834(2)	0.3775(3)	0.3642(4)	0.3556(5)	**0.4576(1)**	0.4053(4)	0.4265(2.5)	0.4265(2.5)	0.3901(5)
Flag	**0.3505(1)**	0.3278(3)	0.3263(5)	0.3268(4)	0.3361(2)	**0.5052(1)**	0.4531(3)	0.4381(4)	0.4624(2)	0.3371(5)
SHeart	**0.7541(1)**	0.7333(2)	0.6363(4)	0.6404(3)	0.5570(5)	0.7254(2)	**0.7333(1)**	0.7033(4)	0.7081(3)	0.557(5)
CHeart	0.7541(2)	**0.7941(1)**	0.5436(4)	0.5446(3)	0.5416(5)	**0.8125(1)**	0.7944(3)	0.7624(4)	0.7983(2)	0.5432(5)
Credit	**0.6348(1)**	0.5786(2)	0.5552(5)	0.5561(3)	0.5554(4)	**0.7826(1)**	0.7203(2)	0.6867(4)	0.7157(3)	0.5557(5)
GCredit	0.7000(5)	**0.7007(1.5)**	0.7004(3.5)	**0.7007(1.5)**	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)
CMC	**0.4384(1)**	0.4276(4.5)	0.4276(4.5)	0.428(2)	0.4278(3)	**0.4838(1)**	0.4329(4)	0.4353(2)	0.4336(3)	0.4272(5)
Adult	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)
Average ranks	1.88	2.38	4.00	2.94	3.81	1.63	2.88	3.31	2.69	4.50
Data sets	AL					CSPA				
	SECG	Fullspace	Subspace	Subsample	Random	**SECG**	Fullspace	Subspace	Subsample	Random
TAE	**0.4576(1)**	0.4245(2)	0.4106(4)	0.4225(3)	0.4040(5)	**0.4450(1)**	0.4265(3)	0.4026(4)	0.4391(2)	0.3901(5)
Flag	0.4505(2)	**0.4582(1)**	0.4407(4)	0.4485(3)	0.3387(5)	0.4536(3)	**0.4747(1)**	0.4407(4)	0.4588(2)	0.3304(5)
SHeart	**0.7541(1)**	0.7333(2)	0.7011(4)	0.7230(3)	0.5556(5)	0.7560(2)	0.6785(4)	**0.7567(1)**	0.7000(3)	0.5600(5)
CHeart	**0.8125(1)**	0.7941(3)	0.7739(4)	0.7957(2)	0.5462(5)	0.7603(4)	0.7789(3)	0.7828(2)	**0.7868(1)**	0.5432(5)
Credit	**0.7563(1)**	0.7552(2)	0.75(3)	0.7352(4)	0.5551(5)	0.7638(2)	0.7372(4)	**0.7819(1)**	0.7561(3)	0.5551(5)
GCredit	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)
CMC	**0.4377(1)**	0.4329(3)	0.4331(2)	0.43(4)	0.427(5)	**0.4688(1)**	0.4342(4)	0.4358(3)	0.4363(2)	0.427(5)
Adult	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)
Average ranks	1.63	2.38	3.38	3.13	4.50	2.38	3.13	2.63	2.38	4.50
Data sets	HGPA					MCLA				
	SECG	Fullspace	Subspace	Subsample	Random	**SECG**	Fullspace	Subspace	Subsample	Random
TAE	**0.4834(1)**	0.4033(3)	0.4026(4)	0.4132(2)	0.3934(5)	**0.4510(1)**	0.4265(3)	0.4205(4)	0.4311(2)	0.3901(5)
Flag	**0.4887(1)**	0.4381(3)	0.4459(2)	0.434(4)	0.3335(5)	0.4649(2)	**0.4830(1)**	0.4454(3)	0.4418(4)	0.334(5)
SHeart	0.5556(3)	0.5556(3)	0.5556(3)	0.5556(3)	0.5556(3)	**0.7541(1)**	0.7333(3)	0.7448(2)	0.7248(4)	0.5574(5)
CHeart	**0.5454(1)**	0.5413(3)	0.5413(3)	0.5413(3)	0.4125(5)	0.7954(2)	0.7937(3)	0.7845(4)	**0.7970(1)**	0.5442(5)
Credit	**0.5572(1)**	0.5515(3)	0.5551(2)	0.5155(4)	0.4555(5)	0.7563(3)	0.759(2)	**0.7736(1)**	0.7455(4)	0.5551(5)
GCredit	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)	0.7000(3)
CMC	0.4277(3)	0.4281(2)	**0.4291(1)**	0.427(4.5)	0.427(4.5)	**0.4702(1)**	0.4299(2)	0.4289(3.5)	0.4289(3.5)	0.427(5)
Adult	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)	0.7607(3)
Average ranks	2.00	2.88	2.63	3.31	4.19	2.00	2.50	2.94	3.06	4.50

Table 3
Results of NMI values for the compared base clusterings generation algorithms.

Data sets	SL					CL				
	SECG	Fullspace	Subspace	Subsample	Random	SECG	Fullspace	Subspace	Subsample	Random
TAE	**0.0603(1)**	0.0441(4.5)	0.0441(4.5)	0.0593(2)	0.047(3)	**0.0595(1)**	0.0258(4)	0.0412(2)	0.0409(3)	0.0164(5)
Flag	0.1400(3)	**0.1638(1)**	0.1287(4)	0.1596(2)	0.1014(5)	0.2104(3)	**0.2363(1)**	0.2090(4)	0.2289(2)	0.0766(5)
SHeart	0.1631(2)	**0.1641(1)**	0.0767(4)	0.0844(3)	0.0177(5)	0.1631(2)	**0.1641(1)**	0.1295(4)	0.1353(3)	0.0023(5)
CHeart	**0.0631(1)**	0.0264(2)	0.0219(3)	0.0215(4)	0.0147(5)	**0.3134(1)**	0.2645(3)	0.2147(4)	0.2755(2)	0.0018(5)
Credit	**0.0284(1)**	0.0255(2)	0.0107(4)	0.0127(3)	0.0087(5)	**0.2838(1)**	0.1700(2)	0.1497(4)	0.1588(3)	0.0017(5)
GCredit	**0.0151(1)**	0.0137(2)	0.01(5)	0.0133(3)	0.0103(4)	**0.0015(1)**	0.0003(4.5)	0.0009(3)	0.0011(2)	0.0003(4.5)
CMC	0.0131(2)	0.0124(3)	0.0119(5)	**0.0134(1)**	0.0121(4)	0.0131(4)	**0.0343(1)**	0.0213(3)	0.0324(2)	0.0015(5)
Adult	**0.0731(1)**	0.066(2)	0.0009(3.5)	0.0008(5)	0.0009(3.5)	**0.3134(1)**	0.1145(2)	0.0451(3)	0.0273(4)	0.0000(5)
Average ranks	1.50	2.19	4.13	2.88	4.31	1.75	2.31	3.38	2.63	4.94
Data sets	AL					CSPA				
	SECG	Fullspace	Subspace	Subsample	Random	**SECG**	Fullspace	Subspace	Subsample	Random
TAE	**0.0595(1)**	0.0312(4)	0.0317(3)	0.0339(2)	0.0172(5)	0.0258(2)	0.023(3)	0.0182(4)	**0.0296(1)**	0.0124(5)
Flag	**0.2386(1)**	0.2339(2)	0.2148(4)	0.224(3)	0.0798(5)	**0.2859(1)**	0.2290(2)	0.1993(4)	0.2137(3)	0.0677(5)
SHeart	0.1634(2)	**0.1641(1)**	0.1261(4)	0.1478(3)	0.0029(5)	**0.2354(1)**	0.0955(4)	0.2037(2)	0.1228(3)	0.0053(5)
CHeart	0.2634(3)	0.2639(2)	0.2278(4)	**0.2689(1)**	0.0033(5)	**0.3544(1)**	0.2405(4)	0.2484(3)	0.2554(2)	0.0021(5)
Credit	**0.2838(1)**	0.2077(3)	0.2163(2)	0.1849(4)	0.0015(5)	0.2084(2)	0.1727(4)	**0.2504(1)**	0.2025(3)	0.0009(5)
GCredit	**0.0015(1)**	0.0004(4.5)	0.0004(4.5)	0.0007(3)	0.0008(2)	**0.0005(1.5)**	0.0003(4)	0.0004(3)	0.0002(5)	**0.0005(1.5)**
CMC	0.0131(4)	**0.0373(1)**	0.0295(3)	0.0351(2)	0.001(5)	0.0283(4)	0.0295(3)	0.0296(2)	**0.0298(1)**	0.0012(5)
Adult	**0.3134(1)**	0.1207(3)	0.0986(4)	0.1220(2)	0.0000(5)	**0.0354(1)**	0.0000(3.5)	0.0000(3.5)	0.0000(3.5)	0.0000(3.5)
Average ranks	1.75	2.56	3.56	2.50	4.63	1.69	3.44	2.81	2.69	4.38
Data sets	HGPA					MCLA				
	SECG	Fullspace	Subspace	Subsample	Random	**SECG**	Fullspace	Subspace	Subsample	Random
TAE	**0.0291(1)**	0.0194(3)	0.0183(4)	0.0201(2)	0.0132(5)	**0.0489(1)**	0.0308(3)	0.03(4)	0.0405(2)	0.0151(5)
Flag	0.1802(3)	**0.1859(1)**	0.1816(2)	0.1775(4)	0.0696(5)	**0.2499(1)**	0.2478(2)	0.2069(4)	0.2124(3)	0.0698(5)
SHeart	**0.0018(1.5)**	0.0017(3)	**0.0018(1.5)**	0.0015(4)	0.0006(5)	**0.1854(1)**	0.1641(3)	0.1810(2)	0.1502(4)	0.0035(5)
CHeart	0.0000(3)	0.0000(3)	0.0000(3)	0.0000(3)	0.0000(3)	**0.2854(1)**	0.2633(3)	0.2471(4)	0.2715(2)	0.0024(5)
Credit	0.0000(3)	0.0000(3)	0.0000(3)	0.0000(3)	0.0000(3)	**0.2838(1)**	0.2101(3)	0.242(2)	0.1916(4)	0.0009(5)
GCredit	**0.0008(1)**	0.0005(3.5)	0.0005(3.5)	0.0005(3.5)	0.0005(3.5)	**0.0015(1)**	0.0004(4.5)	0.0009(2)	0.0004(4.5)	0.0006(3)
CMC	0.0164(2)	**0.0165(1)**	0.013(4)	0.0162(3)	0.0007(5)	**0.0369(1)**	0.0341(3)	0.0281(4)	0.0358(2)	0.0014(5)
Adult	**0.0001(1)**	0.0000(3.5)	0.0000(3.5)	0.0000(3.5)	0.0000(3.5)	**0.2854(1)**	0.1240(4)	0.1334(2)	0.1255(3)	0.0000(5)
Average ranks	1.94	2.63	3.06	3.25	4.13	1.00	3.19	3.00	3.06	4.75

Table 4
Results of ARI values for the compared base clusterings generation algorithms.

Data sets	SL					CL				
	SECG	Fullspace	Subspace	Subsample	Random	SECG	Fullspace	Subspace	Subsample	Random
TAE	**0.0067(1)**	0.0065(2)	0.0024(3)	0.0006(4)	0.0002(5)	**0.0186(1)**	0.0092(4)	0.0183(2)	0.0178(3)	0.0018(5)
Flag	**0.0040(1)**	−0.0186(5)	−0.0119(3)	−0.0151(4)	0.0009(2)	**0.0988(1)**	0.0816(3)	0.0723(4)	0.0896(2)	−0.0049(5)
SHeart	0.1209(2)	**0.2141(1)**	0.0914(4)	0.0982(3)	0(5)	0.1209(4)	**0.2141(1)**	0.1706(3)	0.1788(2)	−0.0011(5)
CHeart	0.2085(2)	**0.3437(1)**	0.0001(4)	0.0011(3)	−0.0009(5)	**0.3986(1)**	0.3445(3)	0.2775(4)	0.3545(2)	−0.0013(5)
Credit	**0.0397(1)**	0.0282(2)	−0.0005(5)	0.0003(3)	−0.0002(4)	**0.3971(1)**	0.216(2)	0.1756(4)	0.2004(3)	0.0004(5)
GCredit	**0.0063(1)**	0.0015(2)	0.0004(4)	0.0014(3)	0.0003(5)	**0.0063(1)**	0.0015(2)	−0.0019(4)	−0.0057(5)	0.0004(3)
CMC	**0.0001(1.5)**	−0.0002(4)	−0.0003(5)	**0.0001(1.5)**	−0.0001(3)	**0.0598(1)**	0.0333(3)	0.0245(4)	0.0364(2)	0.0003(5)
Adult	**0.0516(1)**	0.0389(2)	0(4)	0(4)	0(4)	**0.0852(1)**	0.0646(2)	0.0232(3)	−0.0126(5)	0.0008(4)
Average ranks	1.31	2.38	4.00	3.19	4.13	1.38	2.50	3.50	3.00	4.63
Data sets	AL					CSPA				
	SECG	Fullspace	Subspace	Subsample	Random	**SECG**	Fullspace	Subspace	Subsample	Random
TAE	**0.0286(1)**	0.0149(2)	0.0096(4)	0.0126(3)	0.0065(5)	**0.0227(1)**	0.0126(3)	0.0067(4)	0.0203(2)	0.0000(5)
Flag	**0.0884(1)**	0.0865(2)	0.0812(4)	0.0843(3)	0.0048(5)	**0.1063(1)**	0.1056(2)	0.0889(4)	0.0978(3)	−0.0024(5)
SHeart	0.1209(4)	**0.2141(1)**	0.1662(3)	0.1952(2)	−0.0015(5)	0.2449(2)	0.1246(4)	**0.2611(1)**	0.1593(3)	0.0035(5)
CHeart	**0.3986(1)**	0.3437(3)	0.2986(4)	0.3477(2)	0.0017(5)	**0.4488(1)**	0.309(4)	0.3184(3)	0.327(2)	−0.0005(5)
Credit	**0.2971(1)**	0.2586(2)	0.2577(3)	0.2192(4)	0.0006(5)	0.3004(2)	0.2245(4)	**0.3188(1)**	0.2614(3)	−0.0002(5)
GCredit	**0.0063(1)**	0(4.5)	0.0001(3)	0.0000(4.5)	0.0008(2)	**−0.0002(1)**	−0.0005(4)	−0.0004(3)	−0.0006(5)	−0.0003(2)
CMC	**0.0264(1)**	0.0247(2)	0.0225(3.5)	0.0225(3.5)	0(5)	**0.0276(1)**	0.0248(4)	0.0249(3)	0.0253(2)	0.0000(5)
Adult	**0.0859(1)**	0.0565(3)	0.0462(4)	0.0624(2)	0.0003(5)	**0.0449(1)**	0.0006(4)	0.0006(4)	0.0068(2)	0.0006(4)
Average ranks	1.38	2.44	3.56	3.00	4.63	1.25	3.63	2.88	2.75	4.50
Data sets	HGPA					MCLA				
	SECG	Fullspace	Subspace	Subsample	Random	**SECG**	Fullspace	Subspace	Subsample	Random
TAE	0.0079(2)	0.0066(4)	0.0067(3)	**0.0089(1)**	0.0009(5)	**0.0270(1)**	0.0137(3)	0.0123(4)	0.0207(2)	0.0018(5)
Flag	**0.0876(1)**	0.0755(4)	0.087(2)	0.0777(3)	−0.0016(5)	0.0944(2)	**0.1062(1)**	0.0829(3)	0.0772(4)	−0.0013(5)
SHeart	**−0.0003(1)**	−0.0014(3)	−0.0013(2)	−0.0017(4)	−0.0028(5)	0.2099(3)	0.2141(2)	**0.2375(1)**	0.1986(4)	0.0013(5)
CHeart	**−0.0011(1)**	−0.0033(3.5)	−0.0033(3.5)	−0.0033(3.5)	−0.0033(3.5)	0.3432(2)	0.3429(3)	0.3221(4)	**0.3512(1)**	0.0003(5)
Credit	**−0.0011(1)**	−0.0014(4)	−0.0013(2)	−0.0014(4)	−0.0014(4)	**0.3971(1)**	0.2664(3)	0.3001(2)	0.239(4)	−0.0003(5)
GCredit	**−0.0002(1)**	−0.0003(3)	−0.0003(3)	−0.0003(3)	−0.0007(5)	**0.0063(1)**	0.0003(2)	−0.0045(5)	−0.0017(4)	−0.0002(3)
CMC	**0.0093(1)**	0.009(2)	0.006(4)	0.0083(3)	−0.0007(5)	**0.0267(1)**	0.0216(4)	0.0218(3)	0.0226(2)	0.0001(5)
Adult	**0.0311(1)**	0.0014(2)	−0.0011(4)	−0.0012(5)	0.0008(3)	0.0773(2)	0.0578(4)	**0.0955(1)**	0.0596(3)	0.0000(5)
Average ranks	1.13	3.19	2.94	3.31	4.44	1.63	2.75	2.88	3.00	4.75

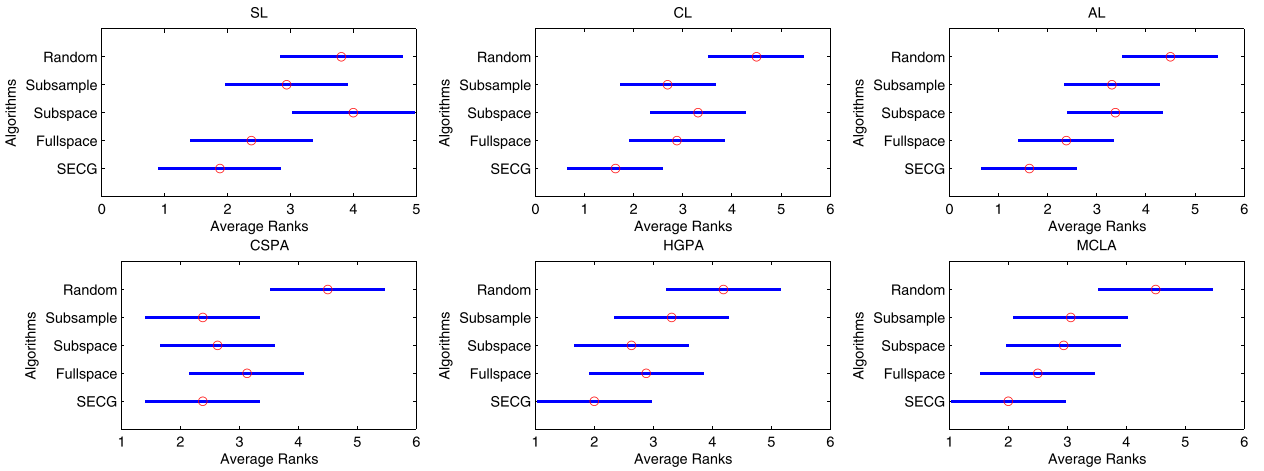

Fig. 1. Nemenyi tests for different base clusterings generation algorithms in terms of CA. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

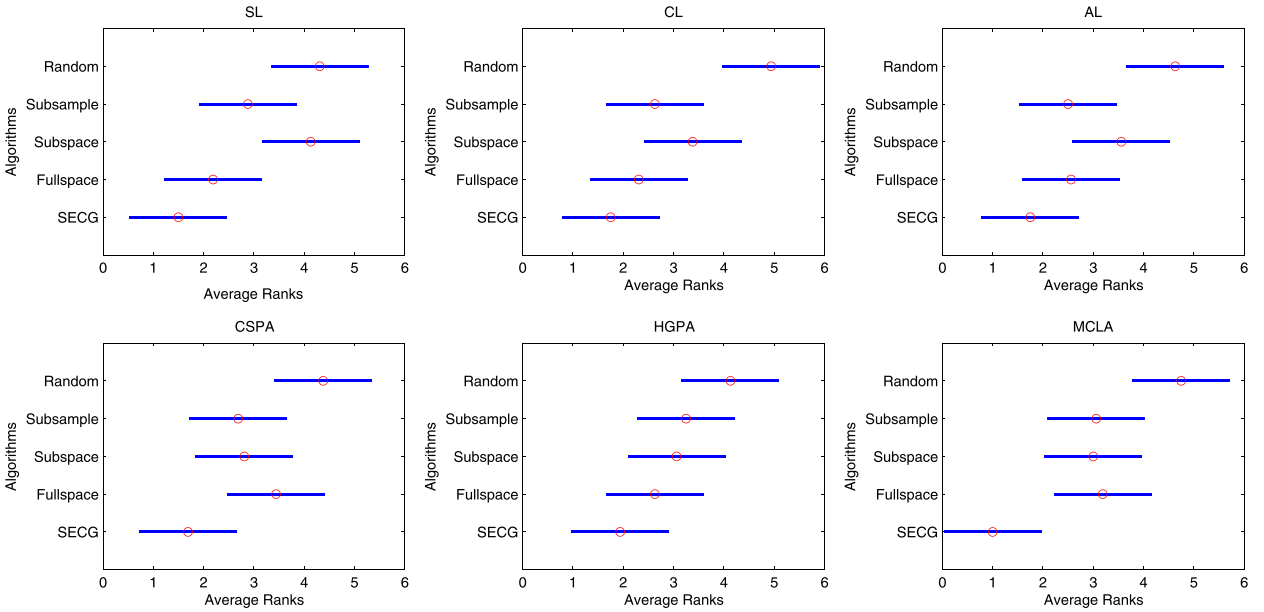

Fig. 2. Nemenyi tests for different base clusterings generation algorithms in terms of NMI. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

cases of the problem considered. Then, the average ranks of the seven algorithms over all 36 cases are calculated to be 1.03, 2.67, 3.39, 2.97 and 4.94 for SECG, Fullspace, Subspace, Subsample and Random algorithms, respectively.

Under the null-hypothesis, which states that all the algorithms are equivalent and so their ranks R_j should be equal, the Friedman statistic

$$\mathcal{X}_F^2 = \frac{12B}{A(A+1)} \sum_{j=1}^{A} R_j^2 - 3B(A+1), \tag{17}$$

is distributed according to $\mathcal{X}_F^2$ with $A-1$ degrees of freedom. According to the Friedman test, a p-value is 7.9660×10^{-12}, which indicates that the null hypothesis can be rejected with high confidence. One can observe that the compared five algorithms are not equivalent and there are significant differences among different algorithms.

Then, the Nemenyi tests are used to reveal the significant differences. The critical difference between two algorithms is defined as

$$CD = q_\alpha \sqrt{\frac{A(A+1)}{6D}}, \tag{18}$$

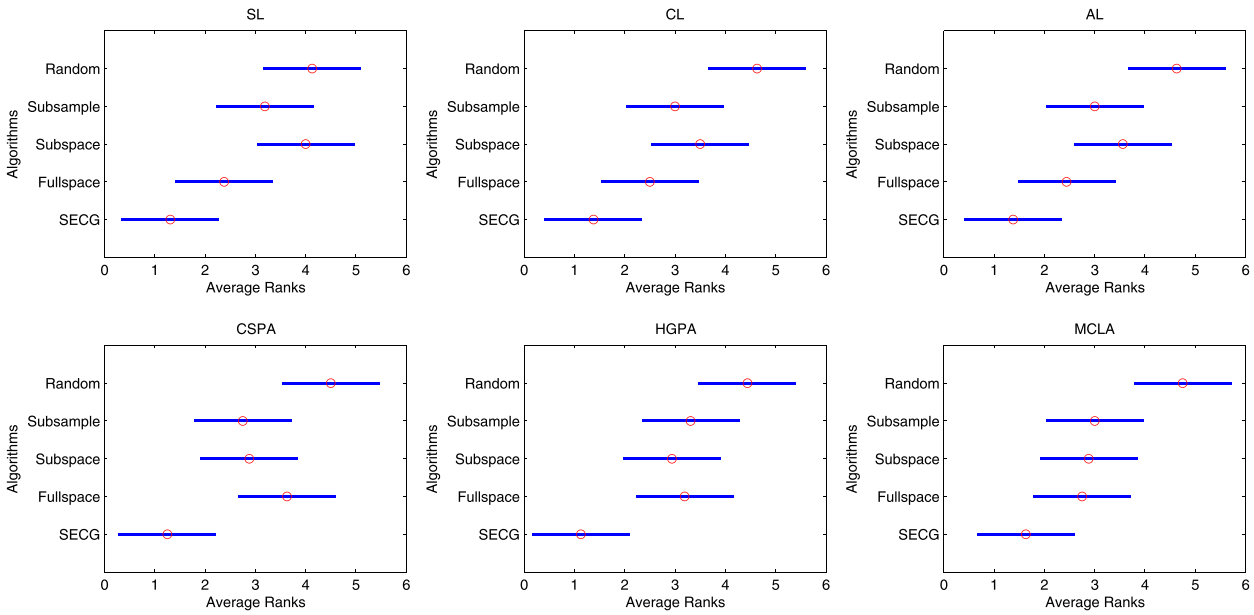

Fig. 3. Nemenyi tests for different base clusterings generation algorithms in terms of ARI. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

where $A = 5$ is the number of algorithms, and $D = 8$ is the number of data sets. We use $\alpha = 0.1$ and get $q_\alpha = 2.459$. Then, the critical difference in our experiment is $CD = 1.9440$.

Figs. 1 –3 show the Nemenyi tests for different base clusterings generation algorithms with each consensus function in terms of CA, NMI and ARI, respectively. The average rank of each algorithm is denoted by a red circle, and a blue bar across the circle shows the critical difference. If the horizontal distance between two circles is larger than the critical difference, then the corresponding two algorithms are significantly different. According to Figs. 2 and 3, the SECG algorithm has significant difference compared with the other four algorithms using CSPA, HGPA and MCLA consensus functions. With SL, CL and AL consensus functions, there exists an overlap between SECG and Fullspace in the horizontal direction, which indicates the proposed SECG algorithm performs as good as or better than using the fullspace based ensemble clusterings generation solutions. In terms of CA index, there exist overlaps among different algorithms with CSPA and MCLA consensus functions, which indicates that the performance of these algorithms is comparable. That is to say, in this case, the proposed SECG algorithm performs as good as the other algorithms.

5. Conclusion

To generate high quality and diversity base clustering results for mixed data in ensemble clustering, this paper proposed a new sequential ensemble clustering generation algorithm dubbed SECG based on the minimization of expected entropy and normalized mutual information. As opposed to other base clusterings generation algorithms, the proposed algorithm considers the correlation between different base clusterings during the process of base clusterings generation. The effectiveness of the proposed algorithm is demonstrated on 8 mixed data sets with three evaluation measures. The experimental results show that the proposed algorithm can effectively extract clustering structures with higher clustering quality in comparison to several state-of-the-art algorithms.

Acknowledgment

The authors are very grateful to the anonymous reviewers and editor. Their many helpful and constructive comments and suggestions helped us significantly improve this work. This work was supported by National Natural Science Foundation of China (Nos. 61603230, 61432011, 61573229, U1435212), the Natural Science Foundation of Shanxi Province, China (No. 201601D202039), Fund Program for the Scientific Activities of Selected Returned Overseas Professionals in Shanxi Province, and the Shanxi Scholarship Council of China (No. 2016–003).

References

[1] J.W. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.
[2] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
[3] R. Xu, D. Wunsch II, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.

[4] A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666.
[5] F.Y. Cao, J.Z. Huang, J.Y. Liang, A fuzzy SV-k-modes algorithm for clustering categorical data with set-valued attributes, Appl. Math. Comput. 295 (2017) 1–15.
[6] Z. He, X. Xu, S. Deng, Scalable algorithms for clustering large datasets with mixed type attributes, Int. J. Intell. Syst. 20 (2005) 1077–1089.
[7] Z.X. Huang, Extensions to the *k*-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Discov. 2 (3) (1998) 283–304.
[8] A. Ahmad, L. Dey, A *k*-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng. 63 (2) (2007) 503–527.
[9] C. Li, G. Biswas, Unsupervised learning with mixed numeric and nominal data, IEEE Trans. Knowl. Data Eng. 14 (4) (2002) 673–690.
[10] C.C. Hsu, C.L. Chen, Y.W. Su, Hierarchical clustering of mixed data based on distance hierarchy, Inf. Sci. 177 (20) (2007) 4474–4492.
[11] J.Y. Liang, X.W. Zhao, D.Y. Li, F.Y. Cao, C.Y. Dang, Determining the number of clusters using information entropy for mixed data, Pattern Recognit. 45 (6) (2012) 2251–2265.
[12] L.I. Kuncheva, S.T. Hadjitodorov, Using diversity in cluster ensembles, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2004, pp. 1214–1219.
[13] J. Ghosh, A. Acharya, Cluster ensembles, WIREs Data Min. Knowl. Discov. 1 (2011) 305–315.
[14] X.W. Zhao, J.Y. Liang, C.Y. Dang, Clustering ensemble selection for categorical data based on internal validity indices, Pattern Recognit. 69 (2017) 150–168.
[15] Z.W. Yu, X.J. Zhu, H.S. Wong, J. You, J. Zhang, G.Q. Han, Distribution-based cluster structure selection, IEEE Trans. Cybern. 47 (11) (2017) 3554–3567.
[16] A.L.N. Fred, A.K. Jain, Combining multiple clusterings using evidence accumulation, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 835–850.
[17] A.P. Topchy, A.K. Jain, W.F. Punch, A mixture model for clustering ensembles, in: Proceedings of the 2004 SIAM International Conference on Data Mining, 2004, pp. 379–390.
[18] L. Kuncheva, D. Vetrov, Evaluation of stability of *k*-means cluster ensembles with respect to random initialization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1798–1808.
[19] A.P. Topchy, A.K. Jain, W.F. Punch, Clustering ensembles: models of consensus and weak partitions, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1866–1881.
[20] A. Strehl, J. Ghosh, Cluster ensembles: a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2002) 583–617.
[21] B. Minaei-Bidgoli, A. Topchy, W. Punch, A comparison of resampling methods for clustering ensembles, in: Proceedings of the International Conference on Machine Learning; Models, Technologies and Application, 2004, pp. 939–945.
[22] A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation, ACM Trans. Knowl. Discov. Data 1 (1) (2007) 1–30.
[23] Z.W. Yu, P.N. Luo, J. You, H.S. Wong, H. Leung, S. Wu, J. Zhang, G.Q. Han, Incremental semi-supervised clustering ensemble for high dimensional data clustering, IEEE Trans. Knowl. Data Eng. 28 (3) (2016) 701–714.
[24] Z.W. Yu, Z.Q. Kuang, J.M. Liu, H.S. Chen, J. Zhang, J. You, H.S. Wong, G.Q. Han, Adaptive ensembling of semi-supervised clustering solutions, IEEE Trans. Knowl. Data Eng. 29 (8) (2017) 1577–1590.
[25] Z.W. Yu, J. You, H.S. Wong, G.Q. Han, From cluster ensemble to structure ensemble, Inf. Sci. 198 (3) (2012) 81–99.
[26] Z.W. Yu, L. Li, H.S. Wong, J. You, G.Q. Han, Y.J. Gao, G.X. Yu, Probabilistic cluster structure ensemble, Inf. Sci. 267 (5) (2014) 16–34.
[27] M. Law, A. Topchy, A.K. Jain, Multiobjective data clustering, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, 2004, pp. 424–430.
[28] Z. Yu, H. Chen, J. You, G. Han, L. Li, Hybrid fuzzy cluster ensemble framework for tumor clustering from bio-molecular data, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (3) (2013) 657–670.
[29] J.C. Gower, A general coefficient of similarity and some of its properties, Biometrics 27 (1971) 857–871.
[30] L. Hunt, M. Jorgensen, Clustering mixed data, WIREs Data Min. Knowl. Discov. 1 (2011) 352–361.
[31] Z. Yu, L. Li, J. Liu, Adaptive noise immune cluster ensemble using affinity propagation, IEEE Trans. Knowl. Data Eng. 27 (19) (2015) 3176–3189.
[32] N. Iam-On, T. Boongoen, S. Garrett, C. Price, A link-based approach to the cluster ensemble problem, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2396–2409.
[33] M. Selim, E. Ertunc, Combining multiple clusterings using similarity graph, Pattern Recognit. 44 (3) (2011) 694–703.
[34] D. Huang, J. Lai, C.D. Wang, Ensemble clustering using factor graph, Pattern Recognit. 50 (2016) 131–142.
[35] Z. Zhou, W. Tang, Clusterer ensemble, Knowl. Based Syst. 19 (1) (2006) 77–83.
[36] D. Barbara, J. Couto, Y. Li, COOLCAT: an entropy-based algorithm for categorical clustering, in: Proceedings of the Eleventh International Conference of Information Knowledge Management, USA, 2002, pp. 582–589.
[37] P. Andritsos, V. Tzerpos, Information-theoretic software clustering, IEEE Trans. Softw. Eng. 31 (2) (2005) 150–165.
[38] K. Chen, L. Liu, Best k: the critical clustering structures in categorical data, Knowl. Inf. Syst. 20 (1) (2009) 1–33.
[39] K. Torkkola, Feature extraction by non-parametric mutual information maximization, J. Mach. Learn. Res. 3 (2003) 1415–1438.
[40] L. Faivishevsky, J. Goldberger, ICA based on a smooth estimation of the differential entropy, Adv. Neural Inf. Process. Syst. 21 (2009).
[41] J.Y. Liang, K.S. Chin, C.Y. Dang, C.M.Y. Richard, A new method for measuring uncertainly and fuzziness in rough set theory, Int. J. Gen. Syst. 31 (4) (2002) 331–342.
[42] UCI machine learning repository, 2012, http://www.ics.uci.edu/mlearn/MLRepository.html.
[43] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.
[44] A.L.N. Fred, A.K. Jain, Data clustering using evidence accumulation, in: Proceedings of the Sixteenth international Conference on Pattern Recognition, 4, 2002, pp. 276–280.
[45] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 3 (2006) 1–30.