Pattern Recognition 69 (2017) 150–168

Contents lists available at ScienceDirect

Pattern Recognition

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

Clustering ensemble selection for categorical data based on internal validity indices

Xingwang Zhao[a,b], Jiye Liang[a,*], Chuangyin Dang[b]

[a] 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
[b] Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 17 August 2016
Revised 16 March 2017
Accepted 17 April 2017
Available online 18 April 2017

Keywords:
Clustering ensemble selection
Categorical data
Clustering validity indices
Quality
Diversity

ABSTRACT

Clustering ensemble selection is an effective technique for improving the quality of clustering results. However, traditional methods usually measure the quality and diversity based on the cluster labels of base clusterings while missing the information of the original data. To solve this problem, a new clustering ensemble selection algorithm for categorical data is presented. In this algorithm, five popular internal validity indices and the normalized mutual information are utilized to measure the quality and diversity of the base clusterings, respectively. According to the quality measure, the partition with the highest value is firstly selected to participate in the ensemble. Then, the base partitions with the highest clustering quality and diversity with respect to the selected base partitions in previous iterations are iteratively selected, until the size of selected base clusterings is satisfied. The effectiveness and robustness of the proposed algorithm are evaluated in comparison with full ensemble, random selection ensemble and the state-of-the-art ensemble selection algorithms. Experimental results on real categorical data sets show that the proposed algorithm is competitive with the existing ensemble selection algorithms in terms of clustering quality.

1. Introduction

The goal of data clustering, also known as cluster analysis, is to discover the inherent structure from the unlabeled data. A good clustering algorithm will produce high quality clusters where the intra-cluster similarity is maximized while the inter-cluster similarity is minimized. Clustering analysis has been widely used in the fields of pattern recognition, social network analysis, bioinformatics, etc [1].

In the past five decades, many clustering algorithms have been developed for numerical data in the literature [2–4]. However, as the attribute values of categorical data are unordered nominal values rather than numerical ones, most of these algorithms can not be directly used to deal with categorical data. To address this deficiency, some categorical data clustering algorithms have been proposed in the clustering community, including *k*-modes type algorithms [5–10], Squeezer [11], COOLCAT [12], ROCK [13], LIMBO [14], DHCC [15], STIRR [16], CLICKS [17], etc. However, there exist no clustering algorithm that performs best for all the categorical data. That is to say, each algorithm has its own strength and weakness, and can not discover all types of cluster structures presented in the categorical data. For a given categorical 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 clustering algorithm would be a proper alternative for a given data set. To overcome these limitations, clustering ensemble algorithms have recently emerged as a powerful alternative to standard clustering algorithms in the clustering community [18–21]. Their main purpose is to improve the robustness and effectiveness of clustering results by merging different base clusterings according to some criterion.

A typical clustering ensemble framework usually involves two key processes: (1) producing a group of base clustering results, and (2) creating a final clustering using a consensus function. Traditionally, all of the base clustering results are used to create the final consensus clustering in the second process. Unfortunately, not all the base clustering results contribute to create the final clustering [22]. Recently, new methods have been developed to improve the clustering quality by evaluating and selecting a subset of base partitions. These methods are widely known as “clustering ensemble selection” or “'cluster ensemble selection”. Their main objective is to generate clustering results based on a subset of base partitions,

* Corresponding author.
E-mail addresses: zhaoxw84@163.com (X. Zhao), ljy@sxu.edu.cn (J. Liang), mecdang@cityu.edu.hk (C. Dang).

http://dx.doi.org/10.1016/j.patcog.2017.04.019
0031-3203/

which performs as good as or better than using all clustering solutions. Toward this goal, a few studies have been reported in the literature [22–25]. These studies have shown that the quality and diversity of the base clustering results are two critical characteristics for cluster ensemble selection. However, in these methods, the normalized mutual information (NMI) or adjusted rand index (ARI) are only based on the labels of base clusterings without the information of the original data set to measure the diversity or quality in the selection process. Thus the desirable characteristics of clustering results, such as clusters compaction and separation, are missing in the NMI or ARI measures. It is well known that different data types, such as, numerical data or categorical data, require different treatments to measure the quality and diversity of the base partitions. This together with the above analysis motivates our work reported in this paper.

To evaluate the quality of clustering results, various effective clustering validity indices have been developed in the literature. Unfortunately, each validity index is only related to some particular features of a clustering result. That is to say, in clustering analysis, no single validity index can capture all different aspects [26]. For our purpose, we will select several suitable clustering validity indices to assess the base partitions in the clustering ensemble selection process. Intuitively, an ensemble should work the best when its clustering solutions are of good quality and at the same time differ from one another significantly. The trade off between quality and diversity is the key design choice that we need to make for ensemble selection. Based on these insightful observations, a new cluster ensemble selection algorithm for categorical data is developed in this paper. In this algorithm, five popular internal validity indices for categorical data and the normalized mutual information are utilized to measure the quality and diversity of base clustering results, respectively. We firstly select the partition with the highest quality to participate in the ensemble. Then, we iteratively select the base partitions with the highest clustering quality and diversity with respect to the set of base partitions selected in previous iterations (selected set), until the desired number of partitions is obtained. Experimental results on several real data sets show that the proposed method is competitive with the existing clustering ensemble selection algorithms in terms of clustering effectiveness and robustness.

The remainder of this paper is organized as follows. Section 2 discusses the related work on categorical data clustering and clustering ensemble selection. In Section 3, five internal validity indices for categorical data are reviewed, and the proposed clustering ensemble selection algorithm is described. A series of experiments to evaluate the performance of the proposed algorithm are conducted in Section 4. Finally, Section 5 provides the conclusions and discussions of future work.

2. Related work

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

2.1. Categorical data clustering

In many fields, a majority of data sets are often described by categorical attributes. Examples of such data sets include statistics data, psychological data, financial records in commercial banks, demographic data, etc. Categorical data clustering is an important task. However, due to the lack of a natural ordering relationship among attribute values of categorical data, this task becomes more challenging. Recently, categorical data clustering has received much attention [5–9,11–17]. These algorithms can be classified as various approaches: partitioning (e.g., *k*-modes), incremental (e.g., Squeezer and COOLCAT), hierarchical (e.g., ROCK, LIMBO, and DHCC), graph-based (e.g., STIRR and CLICKS), etc. Among them, the *k*-modes type algorithms [5,7,8], as a kind of partitional clustering technique, are very popular in different application areas. These algorithms remove the numeric-only limitation of the *k*-means type algorithms [27] and can be used to cluster large categorical data effectively and efficiently. Like the *k*-modes algorithm, the Squeezer algorithm [11] is a one-pass algorithm that is based on summarizing clusters. However, it reads data objects one-by-one. The first object forms a cluster alone. Next objects are either put into an existing cluster or rejected by all to form a new cluster. In order to solve the problem of sensitivity to the initial cluster centers of the *k*-modes algorithm, the COOLCAT algorithm [12] attempts to generate centers based on information entropy. The ROCK algorithm [13] is an agglomerative hierarchical algorithm, which evaluates the similarity between objects of clusters with a link-based similarity measure. In [14], the authors developed a scalable hierarchical categorical clustering algorithm LIMBO, which employs the information bottleneck framework to build a distributional cluster feature. The DHCC algorithm [15] is a divisive hierarchical clustering algorithm, which performs the task of clustering categorical data from an optimization perspective. Some methods apply graph theory to the categorical data algorithm design [16,17]. For example, Gibson et al. [16] first constructed a hypergraph according to the data set, and then clustered the hypergraph using a discrete dynamic system STIRR. The CLICKS algorithm [17] models a categorical data set as a graph, in which vertices are categorical attribute values and edges indicate the co-occurrence relationships of values.

In the above categorical data clustering algorithms, the traditional *k*-modes algorithm [5] is widely used because it is easy to implement, and its efficiency in processing large categorical data sets. Therefore, in the experiments, it is used as the base clustering algorithm in the ensemble clustering.

2.2. Clustering ensemble selection

Clustering ensemble solves a clustering problem in two steps. The first step, known as base clustering, takes a data set as input and generates a set of data partitions. The second step takes the base clustering as input and combines the solutions through a consensus function, and then produces final clustering results.

After generating the initial base clustering results, most of the previous methods use all generated partitions to form final clustering. This may not be the best because some ensemble members are less accurate than others and some may have detrimental effects on the final performance. Thus, one of the key steps is to choose a subset of clustering results with high quality and diversity, which is defined as the clustering ensemble selection by Fern and Lin [22]. Toward this goal, a few clustering ensemble selection methods have been developed to improve the performance of final clustering results. Fern and Lin [22] proposed three heuristics methods for selecting subsets of base clustering that consider both the diversity and quality of the ensemble members. Among these methods, the method named *Cluster And Select* (CAS) was empirically demonstrated to achieve the best overall performance. This method first clusters all ensemble members using spectral clustering algorithm and then selects one solution from each cluster to form the final ensemble. Recently, the authors proposed an adaptive clustering ensemble algorithm which selects a subset of partitions adaptively [23]. Jia et al. [28] generalize the selective clustering ensemble algorithm proposed in [23] and a novel clustering ensemble method was developed based on bagging technique. Hong et al. [24] developed a novel selective clustering ensemble method based on resampling technique. In [25], a hybrid clustering ensemble selection strategy based on the feature selection tech-

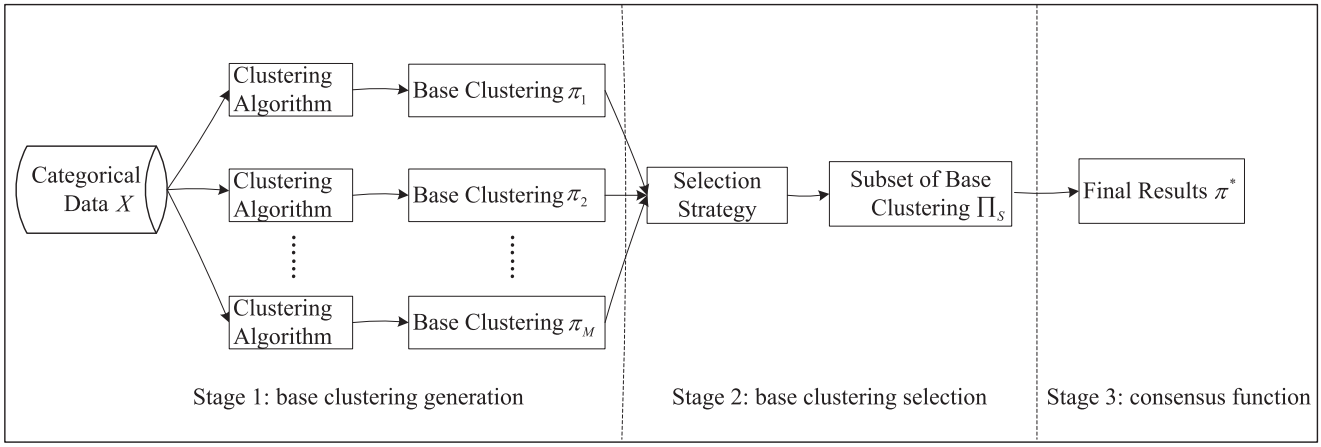

Fig. 1. The framework of the clustering ensemble algorithm.

nique is developed. Instead of selecting a subset of base clustering in the process of clustering ensemble selection, the authors used a subset of clusters to form final clustering results by getting rid of some "bad" clusters [29,30]. Different from the above methods which select a subset of base clusterings to form final clustering results, some efforts have been made to assign varying weights to different base clusterings [31–33]. For example, in [32], several property validity indices (PVIs), namely, variance (VI), connectivity(CI), silhouette width (SI) and Dunn index (DI) are exploited to assign a weight to each base clustering in the ensemble process and a new clustering ensemble method based on kernel functions is developed.

Most of the previous research contributions use the clustering external criteria, i.e., normalized mutual information (NMI) or adjusted rand index (ARI), to evaluate the quality and diversity of base clustering results. These methods have not taken the characteristics of the original data into the process of clustering ensemble selection and deserve further research, which stimulates our current work.

3. Proposed clustering ensemble selection algorithm

In this section, the formal definition of clustering ensemble selection is first given, and then five internal validity indices are reviewed. Based on these five indices, the proposed algorithm of clustering ensemble selection is described in the subsequent subsections.

The clustering ensemble selection problem can be formulated as follows. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be a categorical data set of N objects described by m attributes, and let $\Pi = \{\pi_1, \cdots, \pi_M\}$ be a cluster ensemble with M base clusterings, each of which is also referred as an "ensemble member". Each base clustering consists of a set of clusters $\pi_g = \{C_1^g, C_2^g, \cdots, C_{k_g}^g\} (1 \le g \le M)$, such that $\bigcup_{j=1}^{k_g} C_j^g = X$, where k_g is the number of clusters in the gth base clustering. The problem of clustering ensemble selection is to choose a subset of base clusterings $\Pi_S = \{\pi'_1, \cdots, \pi'_S\} (\Pi_S \subseteq \Pi)$ from the cluster ensemble Π via a certain selection strategy. Then, a final clustering solution $\pi^* = \{C_1, C_2, \cdots, C_k\}$ of the given data set X based on the selected base clusterings Π_S is formed using a consensus function. The process of clustering ensemble framework is depicted in Fig. 1. It is divided into three stages: the generation of base clusterings Π, the selection of the clustering solutions Π_S, and the generation of the final results using the consensus function based on Π_S. This paper mainly focuses on the second stage, whose aim is to select a subset Π_S from the existing base clusterings Π.

3.1. Review of the internal validity indices

Several related articles on clustering ensemble selection have shown that quality and diversity of the base clusterings are crucial for a successful clustering ensemble [22,23]. As is well known, clustering validity measures, including external indices and internal indices, are very useful to evaluate the quality of clustering results. In contrast to external validity indices, which examine the agreement between the cluster labels with the category labels based on a priori information, internal validity indices evaluate the clustering structure of data partitions without any external information. Note that reviewing the literature on clustering validity indices for categorical data is out of the scope of our work. In this paper, we use five different clustering internal validity indices for categorical data to evaluate the quality of base clusterings and select the most promising ones for an ensemble.

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be a categorical data set of N objects with m attributes $A = \{A_1, A_2, \cdots, A_m\}$. V_r^X is the value domain of attribute $A_r \in A$ for categorical data X, i.e., $V_r^X = \{v_r^{(1)}, v_r^{(2)}, \cdots, v_r^{(n_r)}\}$, where n_r is the number of categories of attribute A_r, for $1 \le r \le m$. Similarity, $V_r^{C_i}$ is the value domain of attribute $A_r \in A$ for the data in the cluster C_i. An object $\mathbf{x}_i \in X$ can be represented as a vector $[x_{i1}, x_{i2}, \ldots, x_{im}]$, where $x_{ir} \in V_r^X$, for $1 \le r \le m$. Suppose that one of the base clusterings for categorical data X is $\pi = \{C_1, C_2, \cdots, C_k\}$, where k is the number of clusters. In the following, the five internal indices are briefly described.

- **Category utility function (CU)**
  The category utility function introduced by Gluck and Corter [34], is a measure of "category goodness", which has been applied in some categorical data clustering algorithms [35,36] and consensus clustering [38]. It can be described as follows:

$$CU(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{|C_i|}{N}\sum_{r=1}^{m}\sum_{q=1}^{n_r}\left[P(A_r = v_r^{(q)}|C_i)^2 - P(A_r = v_r^{(q)})^2\right], \tag{1}$$

  where $P(A_r = v_r^{(q)}|C_i)$ is the probability of $v_r^{(q)}$ for the rth attribute in the cluster C_i, i.e., $P(A_r = v_r^{(q)}|C_i) = \frac{|\{\mathbf{x}_i | x_{ir} = v_r^{(q)}, \mathbf{x}_i \in C_i\}|}{|C_i|}$; $P(A_r = v_r^{(q)})$ is the probability of $v_r^{(q)}$ for the rth attribute in the categorical data set X, i.e., $P(A_r = v_r^{(q)}) = \frac{|\{\mathbf{x}_i | x_{ir} = v_r^{(q)}, \mathbf{x}_i \in X\}|}{N}$. According to Eq. (1), one can find that the CU index attempts to maximize the probability that two data objects in the same cluster obtain the same attribute values, which looks different from more traditional clustering criteria adhering to similarities

and dissimilarities between objects [37,39]. And it is clear that the higher the value of CU index, the better the clustering results.

- **Categorical data clustering with subjective factors (CDCS)**
The CDCS index [40] is defined based on intra-cluster cohesion and inter-cluster similarity for a clustering result, which is given as follows:

$$CDCS(\pi) = \frac{intra(\pi)}{inter(\pi)}. \tag{2}$$

First, the intra-cluster cohesion for a clustering result is defined as the weighted cohesion of each cluster, where the cohesion for a cluster C_i is the summation of the highest probability for each attribute as shown below:

$$intra(\pi) = \sum_{i=1}^{k} \frac{|C_i|}{N} \sum_{r=1}^{m} \frac{1}{m} \left(\max_{q=1}^{n_r} P(A_r = v_r^{(q)}) \right)^3. \tag{3}$$

And, the inter-cluster similarity for a clustering result is the summation of cluster similarity for all cluster pairs, weighted by the cluster size. The $inter(\pi)$ is described as follows:

$$inter(\pi) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} Sim(C_i, C_j)^{1/m} \cdot |C_i \bigcup C_j|}{(k-1) \cdot N}, \tag{4}$$

where the similarity between two clusters C_i and C_j, is computed as, $Sim(C_i, C_j) = \prod_{r=1}^{m} \left[\sum_{q=1}^{n_r} \min\{P(A_r = v_r^{(q)}|C_i), P(A_r = v_r^{(q)}|C_j)\} + \epsilon \right]$. The idea behind this definition is that the more the clusters intersect, the more similar they are. If the distribution of attribute values for two clusters is similar, they will have a higher similarity score. The exponent $1/m$ is used for normalization since there are m component multiplications when computing $Sim(C_i, C_j)$. According to this equation, the best partition should be indicated by larger values of $Intra(\pi)$ and lower values of $Inter(\pi)$. That is to say, good partitions are distinguished by large values of CDCS.

- **Davies-Bouldin (DB)**
The DB index [41] is a function of the ratio of the sum of within-cluster scatter to between-cluster separation, which is firstly used to evaluate the clustering results of numerical data. Here we use it to evaluate the base clustering of categorical data. The scatter within the ith cluster, s_i, is computed as $s_i = \frac{1}{|C_i|} \sum_{\mathbf{x}_i \in C_i} \sum_{r=1}^{m} d(x_{ir}, z_{ir})$, and the distance between clusters C_i and C_j, denoted by d_{ij}, is defined as $d_{ij} = \sum_{r=1}^{m} d(z_{ir}, z_{jr})$. Here, z_{ir} and z_{jr} represent the modes of ith and jth clusters in the rth attribute, respectively. And d is a simple matching dissimilarity measure, i.e., $d(x, y) = \begin{cases} 0, & x = y, \\ 1, & x \neq y. \end{cases}$ Then, the DB index is defined as follows:

$$DB(\pi) = \frac{1}{k} \sum_{i=1}^{k} \max_{j, j \neq i} \{ \frac{s_i + s_j}{d_{ij}} \}. \tag{5}$$

According to this equation, good partitions, composed of compact and separated clusters, should be indicated by small values of DB.

- **Cluster cardinality index (CCI)**
Inspired by set operations, which are often used to describe the property and structure of categorical data, a new cluster validity index for categorical data, named cluster cardinality index (CCI) [42], is defined as follows:

$$CCI(\pi) = \frac{1}{k} \sum_{i=1}^{k} \max_{j, j \neq i} \left\{ \frac{CI(i) + CI(j)}{CI(i, j)} \right\}, \tag{6}$$

where $CI(i) = \frac{1}{m} \sum_{r=1}^{m} \frac{|V_r^{C_i}|}{|C_i|}$, $CI(i, j) = \frac{1}{m} \sum_{r=1}^{m} \frac{|V_r^{C_i} \cup V_r^{C_j}| - |V_r^{C_i} \cap V_r^{C_j}| + 1}{|V_r^{C_i} \cup V_r^{C_j}| + 1}$, $V_r^{C_i}$ and $V_r^{C_j}$ are the value domains of the rth attribute within clusters i and j. In the above equation, $CI(i)$ is the average number of categorical values of cluster i over all attributes and $CI(i, j)$ is the average number of different categorical values between clusters i and j over all attributes. The CCI tries to minimize the average dissimilarity of objects within the same cluster with a smaller number of categorical values in each attribute and maximize the dissimilarity of different clusters with a larger number of different categorical values in all attributes. Hence, the CCI index is small if the clusters are compact and far from each other.

- **Information entropy (IE)**
The IE index [12] uses the information-theoretic principles and the notion of entropy to measure the clustering results. The basic intuition is that groups of similar objects have lower entropy than those of dissimilar ones. The information entropy index for clustering results of categorical data is defined as follows:

$$IE(\pi) = -\frac{1}{k} \sum_{i=1}^{k} \frac{|C_i|}{n} \sum_{j=1}^{m} \sum_{q=1}^{n_j} P(A_j = v_j^{(q)}|C_i) \log P(A_j = v_j^{(q)}|C_i), \tag{7}$$

where $P(A_j = v_j^{(q)}|C_i)$ represents the same meaning as the definition of Eq. (1). It is verify that clustering results composed of good clusters are distinguished by small values of IE.

3.2. Quality and diversity measures

For the problem of clustering ensemble, the quality and diversity measures are considered as two critical factors for the selection of base clusterings to be ensemble. In this subsection, we will explain how we measure the quality and diversity of base clustering results.

The five clustering internal validity indices are adapted to assess the quality of the candidate base clusterings and to select the base partitions. In order to measure the quality easily, a function $CIVI(\pi_i, index_j)$ is used to return the value of clustering internal validity index for the ith base clustering π_i when evaluated by the jth index. For a given data set X, it is noticed that the maximal values of some validity indices (e.g., CU and CDCS) are correspondent with the best clustering performance. Whereas the optimal clustering results will be found when the values of these indices (e.g., DB, CCI, and IE) are minimal. In order to measure the quality consistently, the values of the later three indices (e.g., DB, CCI, and IE) are recalculated by $exp(-1 * CIVI(\pi_i, index_j))$. And the $CIVI$ values of the same validity index for the different base clusterings are normalized to [0, 1] by the following criterion:

$$CIVI(\pi_i, index_j) = \frac{CIVI(\pi_i, index_j)}{\sum_{r=1}^{M} CIVI(\pi_r, index_j)}, \tag{8}$$

where M is the number of base clusterings, $1 \leq i \leq M$, and $j = 1, 2, \ldots, 5$. Therefore, the higher the values of these five indices for a clustering result, the better the clustering performance is. Then, the sum of the individual values is then calculated for each base partition (over the different indices) and the base partitions with highest sums are selected for the ensemble. The quality of each base partition is calculated as follows:

$$quality(\pi_i) = \frac{1}{\nu} \sum_{j=1}^{\nu} CIVI(\pi_i, index_j), \tag{9}$$

where $\nu = 5$ stands for the number of internal validity indices.

About the diversity measure of base clustering results, there are several different measures in the literature of cluster ensembles. The majority of them are based on the matching of labels acquired from two data partitions. Because the normalized mutual information (NMI) has been shown to impact the clustering ensemble performance and it is easy to compute, we use it to measure the diversity in this paper. The lower the NMI value, the higher is the diversity. Note that the diversity measure used in this paper do not limit it. Let $\pi_p = \{C_1^p, C_2^p, \cdots, C_{k_p}^p\}$ and $\pi_q = \{C_1^q, C_2^q, \cdots, C_{k_q}^q\}$ are two base clusterings for the data set X, the NMI between them is given by:

$$NMI(\pi_p, \pi_q) = \frac{\sum_{i=1}^{k_p}\sum_{j=1}^{k_q} N_{ij} \log \frac{N \cdot N_{ij}}{N_i^p \cdot N_j^q}}{\sqrt{\sum_{i=1}^{k_p} N_i^p \log \frac{N_i^p}{N} \sum_{j=1}^{k_q} N_j^q \log \frac{N_j^q}{N}}}, \tag{10}$$

where N is the number of objects of the data set X; N_{ij} is the number of common objects of clusters C_i^p and C_j^q; N_i^p is the number of objects in cluster C_i^p; and N_j^q is the number of objects in cluster C_j^q. Then, $1 - NMI(\pi_i, \pi_j)$ denotes the diversity between base clusterings π_p and π_q. In particular, given a set of base clusterings Π, the average diversity measure between a partition $\pi_p (\pi_p \notin \Pi)$ and the set Π is defined as:

$$diversity(\pi_p, \Pi) = \frac{1}{|\Pi|} \sum_{\pi_q \in \Pi} (1 - NMI(\pi_p, \pi_q)). \tag{11}$$

Intuitively, the higher $diversity(\pi_p, \Pi)$ is, the more diverse of base clustering π_p regarding the set of base partitions Π.

3.3. Algorithm description

Based on the above mentioned formulations and notations, the developed clustering ensemble selection algorithm is shown in Algorithm 1, which is abbreviated as SIVID (Sum of Internal Validity Indices with Diversity). In this algorithm, the sum of the individual internal validity indices is firstly calculated for each base partition, and the base partition with the highest value is selected for the ensemble. Then, the base partitions with the highest clustering quality and diversity with respect to the set of the selected base partitions is iteratively selected, until the desired number of partitions is satisfied.

Algorithm 1 The SIVID algorithm.

Input:
X: a categorical data set;
$\Pi = \{\pi_1, \cdots, \pi_M\}$: a set of candidate base clusterings;
S: the number of base clusterings to be selected for the ensemble.
Output:
$\Pi_S (\Pi_S \subseteq \Pi)$: the selected base clusterings.
Construct the selected base clusterings set $\Pi_S = \varnothing$;
According to Eq. (9), compute the quality for each base clustering;
$\pi_t = \underset{\pi_m \in \Pi}{argmax}\, quality(\pi_m)$;
Update $\Pi_S = \Pi_S \bigcup \{\pi_t\}$ and $\Pi = \Pi - \{\pi_t\}$;
repeat
  According to Eq. (11), compute the diversity of base clusterings $\pi_m \in \Pi$ with respect to the set of the selected base partitions Π_S;
  $\pi_t = \underset{\pi_m \in \Pi}{argmax}[quality(\pi_m) * diversity(\pi_m, \Pi_S)]$;
  Update $\Pi_S = \Pi_S \bigcup \{\pi_t\}$ and $\Pi = \Pi - \{\pi_t\}$;
until $|\Pi_S| = S$;
return Π_S.

The computational complexity of the SIVID algorithm has two factors, the computation of quality and the computation of diversity. Suppose that the most computation requirements for invalidity indices are $O(I_{\max})$, and the computation of the quality of each base partition is $O(\nu MI_{\max})$, where $\nu = 5$ stands for the number of internal validity indices, M is the number of the base partitions. Then, the complexity of diversity computation needs $O(M^2)$. Therefore, the total time complexity of the SIVID algorithm is $O(\nu MI_{\max} + M^2)$.

Table 1
Characteristics of the categorical data sets.

Data sets	# objects	# attributes	# classes
zoo	101	17	7
promoter	106	58	2
hayes	160	5	3
dermatology	366	35	3
vote	435	17	2
balance	625	5	2
breastcancer	699	10	2
krvskp	3196	37	2
mushroom	5644	23	2
nursery	12,960	8	3
basehock	1993	4862	2
pcmac	1943	3289	2

4. Experimental analysis

This section presents the effectiveness and robustness evaluation of the proposed algorithm over 12 real-world data sets in terms of some benchmark evaluation criteria.

4.1. Data sets

The characteristics of the categorical data sets are shown in Table 1. The first ten data sets are downloaded from the UCI machine learning repository [44]. The last two data sets are text data, which are from website[1] and have been used earlier to evaluate feature selection algorithms. Note that all these data sets are labeled and contain supervised class information. However, the class labels were not used in the processes of clustering or ensemble selection 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), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI), which measure the agreement of the clustering results produced by an algorithm and the ground truth.

Suppose that $C = \{c_1, c_2, \cdots, c_k\}$ and $P = \{p_1, p_2, \cdots, p_{k'}\}$ 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

[1] http://featureselection.asu.edu/old/datasets.php

pre-defined class labels. The CA is defined as:

$$CA = \frac{\sum_{i=1}^{k} max_{j=1}^{k'} N_{i,j}}{N}. \quad (12)$$

• *Adjusted Rand Index (ARI).* ARI takes into account the number of objects that exist in the same cluster and different clusters [45]. The ARI is defined as:

$$ARI = \frac{\binom{N}{2}\sum_{i=1}^{k}\sum_{j=1}^{k'}\binom{N_{i,j}}{2} - [\sum_{i=1}^{k}\binom{N_i^c}{2}\sum_{j=1}^{k'}\binom{N_j^p}{2}]}{\frac{1}{2}\binom{N}{2}[\sum_{i=1}^{k}\binom{N_i^c}{2} + \sum_{j=1}^{k'}\binom{N_j^p}{2}] - [\sum_{i=1}^{k}\binom{N_i^c}{2}\sum_{j=1}^{k'}\binom{N_j^p}{2}]}. \quad (13)$$

• *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:

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'} N_{i,j} log\frac{N \cdot N_{i,j}}{N_i^c \cdot N_j^p}}{\sqrt{\sum_{i=1}^{k} N_i^c \cdot log\frac{N_i^c}{N} \cdot \sum_{j=1}^{k'} N_j^p \cdot log\frac{N_j^p}{N}}}. \quad (14)$$

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 SIVID algorithm, it is compared with a number of cluster ensemble algorithms, both the full clustering ensemble algorithm and the state-of-the-art clustering ensemble selection algorithms developed in the literature. Details of these compared algorithms are described in the following.

- Full clustering ensemble algorithm (Full): The final consensus solution is obtained by a consensus function based on all the base clusterings.
- Random selection algorithm (RS): This algorithm randomly selects part of the base clusterings to form the final results.
- Cluster and selection algorithm (CAS) [22]: This algorithm firstly partitions the clustering solutions into similar groups using spectral clustering algorithm, and then selects only one clustering solution with the highest quality from each group to form the ensemble.
- Adaptive cluster ensemble selection algorithm (ACES) [23]: This algorithm firstly generates a diverse set of solutions and combines them into a consensus partition P^*. Based on the diversity between the ensemble members and P^*, a subset of base clusterings is selected and combined to obtain the final clustering results.
- Selective spectral clustering ensemble (SELSCE) [28]: In this algorithm, after the generation of base clusterings, the bagging technique was used to rank and assess the base clusterings. Based on this ranking, ensemble members were selected for ensemble. In this algorithm, we set the Iteration $T = 10$.
- Hybrid clustering solution selection strategy (HCSS) [25]: This algorithm views clustering solution selection as feature selection problem. Based on four different feature selection approaches, a hybrid clustering solution selection strategy is designed to identify suitable clustering solutions.

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

- For generating the base clusterings, the k-modes clustering and simply random partition algorithms are performed. For these two algorithms, the number of clusters k are different, which are selected in the range of $\{2, \sqrt{N}\}$, where N is the number of objects. For the k-modes algorithm, 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 [46] 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) [43] are used in our experiments.
- In all the experiments, we set the size of base clustering $M = 50$. Unless otherwise mentioned, the size of selected base clusterings is set to 25.
- The reported experimental results are the average values with 20 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 two methods for generating the base clusterings, seven clustering ensemble selection algorithms and six consensus functions result in 84 kinds of ensemble clustering results.

4.4. Results on effectiveness analysis

In this subsection, we focus on the clustering performance of different clustering ensemble selection algorithms on above mentioned 12 data sets with three evaluation criteria. In particular, the results of 50% base clusterings subset are presented, and the related statistical tests are carried out.

Tables 2–7, show the experiment results. For each data set, the rank values of different ensemble selection algorithms 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 ensemble selection 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 for different algorithms with different base clusterings generation methods are listed in Tables 2–3. According to the average ranks, we find that the SIVID algorithm always outperforms other algorithms with two different base clusterings generation algorithms. However, the clustering results using the simply random partition algorithm generating base clusterings are poorer than those with the k-modes clustering algorithm generating base clusterings. That is because the quality of base clusterings using the simply random partition algorithm is low. From Table 2, the SIVID algorithm achieves the best performance on nine of the twelve data sets with SL, CL and AL consensus functions. Using the CSPA consensus function, the SIVID algorithm obtains the best performance on eight of the twelve data sets. This superiority is more evident for the mushroom data sets. In addition, these results in-

Table 2
Results of CA values for the compared algorithms using the *k*-modes clustering algorithm generating base clusterings.

Data sets	SL							CL						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	**SIVID**	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.8936(1)**	0.8327(3.5)	0.8287(6)	0.8317(5)	0.8327(3.5)	0.8153(7)	0.8351(2)	0.8955(3)	**0.9054(1.5)**	0.8911(5)	0.8871(6)	**0.9054(1.5)**	0.8441(7)	0.8916(4)
promoter	0.5108(2)	0.5104(3.5)	0.5094(5.5)	0.5104(3.5)	**0.5132(1)**	0.5038(7)	0.5094(5.5)	0.5904(5)	**0.6505(1)**	0.5962(4)	0.5986(3)	0.5863(6)	0.5198(7)	0.6104(2)
hayes	**0.4263(1)**	0.4253(2)	0.4244(4)	0.4216(5)	0.4247(3)	0.4203(6)	0.4200(7)	**0.4456(1)**	0.4319(7)	0.4444(2)	0.4434(3)	0.4413(4)	0.4350(6)	0.4391(5)
dermatology	**0.5275(1)**	0.5046(2.5)	0.5029(4)	0.4602(5)	0.5046(2.5)	0.4038(7)	0.4499(6)	**0.7846(1)**	0.7743(2.5)	0.6623(7)	0.7142(4)	0.7743(2.5)	0.6847(5)	0.6839(6)
vote	**0.6386(1)**	0.6144(6)	0.6151(2)	0.6147(4)	0.6143(7)	0.6146(5)	0.6148(3)	**0.8307(1)**	0.7379(2)	0.7253(4)	0.6423(5)	0.7285(3)	0.6346(7)	0.6416(6)
balance	**0.466(1)**	0.4634(4)	0.4634(4)	0.4636(2)	0.4632(6.5)	0.4632(6.5)	0.4634(4)	0.5044(2)	**0.5126(1)**	0.5006(5)	0.4862(7)	0.4962(6)	0.5009(4)	0.5034(3)
breastcancer	**0.6849(1)**	0.6568(2)	0.6565(5.5)	0.6567(3)	0.6566(4)	0.6563(7)	0.6565(5.5)	**0.8448(1)**	0.8301(2)	0.7374(4)	0.7986(3)	0.7300(6)	0.7366(5)	0.7271(7)
krvskp	**0.5224(1)**	0.5222(5.5)	0.5222(5.5)	0.5223(2.5)	0.5222(5.5)	0.5223(2.5)	0.5222(5.5)	**0.5391(1)**	0.5339(2)	0.5243(7)	0.5259(6)	0.5278(5)	0.5329(3)	0.5328(4)
mushroom	**0.7348(1)**	0.6887(3)	0.6584(6)	0.6908(2)	0.6325(7)	0.6753(4)	0.6585(5)	**0.6935(1)**	0.6363(6)	0.6691(2)	0.6625(3)	0.6339(7)	0.6488(4)	0.6411(5)
nursery	**0.3488(1)**	0.3353(3)	0.3335(5.5)	0.3335(5.5)	0.3335(5.5)	0.3335(5.5)	0.3363(2)	**0.3809(1)**	0.3437(3)	0.3385(7)	0.3423(4)	0.3405(6)	0.3453(2)	0.3419(5)
basehock	0.5013(4)	0.5013(4)	0.5013(4)	0.5013(4)	0.5013(4)	0.5013(4)	0.5013(4)	**0.5790(1)**	0.5615(2.5)	0.5615(2.5)	0.5013(7)	0.5063(6)	0.5108(5)	0.5238(4)
pcmac	0.5059(4)	0.5059(4)	0.5059(4)	0.5059(4)	0.5059(4)	0.5059(4)	0.5059(4)	**0.5404(1)**	0.5340(3)	0.5147(6)	0.5129(7)	0.5211(5)	0.5362(2)	0.5239(4)
average ranks	1.58	3.58	4.67	3.79	4.46	5.46	4.46	1.58	2.79	4.63	4.83	4.83	4.75	4.58
Data sets	AL							CSPA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	**SIVID**	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.9010(1)**	0.8847(2.5)	0.8762(6)	0.8827(4)	0.8847(2.5)	0.8510(7)	0.8817(5)	0.8404(2)	0.8213(6.5)	0.8307(5)	**0.8431(1)**	0.8213(6.5)	0.8371(3)	0.8347(4)
promoter	**0.6694(1)**	0.6344(5)	0.5910(6)	0.6637(2)	0.6368(4)	0.5797(7)	0.6425(3)	**0.6755(1)**	0.6057(5)	0.5868(6)	0.6458(2)	0.6377(4)	0.5594(7)	0.6392(3)
hayes	**0.4463(1)**	0.4272(6)	0.4334(5)	0.4347(3.5)	0.4347(3.5)	0.4375(2)	0.4266(7)	**0.4344(1)**	0.4197(6)	0.4184(7)	0.4234(5)	0.4291(2)	0.4263(3)	0.4250(4)
dermatology	**0.7939(1)**	0.7799(4.5)	0.7735(7)	0.7825(3)	0.7799(4.5)	0.7758(6)	0.7896(2)	**0.8169(1)**	0.8094(2.5)	0.8016(6)	0.8056(5)	0.8094(2.5)	0.7813(7)	0.8064(4)
vote	**0.8668(1)**	0.8490(4.5)	0.8490(4.5)	0.8482(6)	0.8474(7)	0.8580(2)	0.8513(3)	0.8526(2)	0.8492(4)	**0.8531(1)**	0.8508(3)	0.8487(5)	0.8256(7)	0.8453(6)
balance	**0.5692(1)**	0.5398(4.5)	0.5398(4.5)	0.5307(7)	0.5418(3)	0.5327(6)	0.5511(2)	0.5694(2)	0.5637(3)	0.5500(7)	0.5510(6)	**0.5719(1)**	0.5512(5)	0.5615(4)
breastcancer	**0.9547(1)**	0.9481(4)	0.9484(3)	0.9489(2)	0.9396(7)	0.9413(6)	0.9476(5)	**0.8328(1)**	0.8247(2)	0.8212(5)	0.8240(4)	0.8166(6)	0.8024(7)	0.8246(3)
krvskp	0.5961(2)	**0.5983(1)**	0.5769(5)	0.5757(6)	0.5646(7)	0.5878(3)	0.5824(4)	**0.5442(1)**	0.5299(6)	0.5396(2)	0.5350(5)	0.5254(7)	0.5364(4)	0.5371(3)
mushroom	**0.8579(1)**	0.8284(6)	0.8521(2)	0.8288(5)	0.8068(7)	0.8340(4)	0.8344(3)	**0.7686(1)**	0.6938(4)	0.6858(5)	0.7231(3)	0.7342(2)	0.6633(7)	0.6752(6)
nursery	**0.5941(1)**	0.5070(3)	0.4685(7)	0.5448(2)	0.4742(6)	0.4827(4)	0.4769(5)	0.5720(2)	**0.6215(1)**	0.4182(7)	0.5192(5)	0.5197(4)	0.5201(3)	0.4338(6)
basehock	0.5013(5)	**0.5314(1.5)**	**0.5314(1.5)**	0.5013(5)	0.5013(5)	0.5013(5)	0.5013(5)	**0.5826(1)**	0.5519(2)	0.5514(3)	0.5444(5)	0.5013(7)	0.5365(6)	0.5507(4)
pcmac	0.5154(2)	0.5118(4)	0.5059(6)	0.5057(7)	**0.5160(1)**	0.5061(5)	0.5142(3)	**0.5733(1)**	0.5548(2)	0.5538(3)	0.5378(4)	0.5319(6)	0.5259(7)	0.5357(5)
average ranks	1.50	3.88	4.79	4.38	4.79	4.75	3.92	1.33	3.67	4.75	4.00	4.42	5.50	4.33
Data sets	HGPA							MCLA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	**SIVID**	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.8636(1)**	0.8604(2.5)	0.851(5)	0.8525(4)	0.8604(2.5)	0.8406(7)	0.8495(6)	**0.8738(1)**	0.8733(2.5)	0.8658(4)	0.8564(5)	0.8733(2.5)	0.8059(7)	0.8475(6)
promoter	0.7476(2)	**0.7575(1)**	0.733(4)	0.7434(3)	0.7042(5)	0.6627(7)	0.6868(6)	**0.7726(1)**	0.7613(2)	0.6995(4)	0.7071(3)	0.6745(6)	0.6156(7)	0.6840(5)
hayes	0.4422(2)	0.4381(5)	0.4388(3)	0.4369(7)	0.4384(4)	**0.4428(1)**	0.4375(6)	0.4300(2)	0.4225(6)	0.4288(3)	0.4247(5)	**0.4378(1)**	0.4275(4)	0.4219(7)
dermatology	0.8308(2)	0.8089(7)	0.8119(6)	**0.8411(1)**	0.8179(4)	0.8128(5)	0.8238(3)	**0.8616(1)**	0.8317(2)	0.8306(4)	0.7988(6)	0.8316(3)	0.7332(7)	0.8270(5)
vote	**0.8851(1)**	0.8759(2)	0.8679(7)	0.8713(5)	0.8757(3)	0.8746(4)	0.8691(6)	**0.8587(1)**	0.8505(7)	0.8538(3)	0.8515(5)	0.8513(6)	0.8531(4)	0.8545(2)
balance	**0.5362(1)**	0.5353(2)	0.5264(7)	0.5273(6)	0.5299(5)	0.5316(4)	0.5340(3)	0.5564(2)	**0.5610(1)**	0.5450(5)	0.5484(4)	0.5530(3)	0.5367(7)	0.5434(6)
breastcancer	0.8655(2)	**0.8682(1)**	0.8499(6)	0.8541(4)	0.8512(5)	0.8642(3)	0.8479(7)	**0.9416(1)**	0.9174(3)	0.9247(2)	0.9092(4.5)	0.8829(6)	0.8663(7)	0.9092(4.5)
krvskp	**0.5527(1)**	0.5229(7)	0.5240(3)	0.5231(6)	0.5236(4)	0.5270(2)	0.5232(5)	**0.5396(1)**	0.5224(7)	0.5255(6)	0.5263(5)	0.5267(4)	0.5353(2)	0.5274(3)
mushroom	0.6848(2)	0.6296(7)	0.6754(4)	**0.6928(1)**	0.6427(6)	0.6823(3)	0.6490(5)	0.6988(5)	0.6960(6)	0.7342(2)	0.7256(3)	0.7020(4)	**0.7458(1)**	0.6095(7)
nursery	0.4638(4)	0.4477(6)	0.4571(5)	0.4093(7)	0.4807(3)	**0.5110(1)**	0.5010(2)	0.4233(5)	0.4815(2)	0.4716(3)	0.3939(7)	0.4150(6)	**0.5225(1)**	0.4352(4)
basehock	0.5677(2)	0.5502(4)	0.5492(6)	0.5585(3)	0.5499(5)	**0.5780(1)**	0.5421(7)	**0.5544(1)**	0.5512(3)	0.5211(5)	0.5519(2)	0.5013(7)	0.5283(4)	0.5178(6)
pcmac	**0.5772(1)**	0.5628(2)	0.5306(5)	0.5247(6)	0.5440(3)	0.5208(7)	0.5407(4)	0.5318(3)	0.5257(6)	0.5299(4)	0.5260(5)	0.5340(2)	**0.5353(1)**	0.5216(7)
average ranks	1.75	3.88	5.08	4.42	4.13	3.75	5.00	2.00	3.96	3.75	4.54	4.21	4.33	5.21

Table 3
Results of CA values for the compared algorithms using the simply random partition algorithm generating base clusterings.

Data sets	SL							CL						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	**SIVID**	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.4792(1)**	0.4356(5)	0.4406(3)	0.4307(6.5)	0.4406(3)	0.4406(3)	0.4307(6.5)	**0.4284(1)**	0.4158(4.5)	0.4208(3)	0.4257(2)	0.4158(4.5)	0.4109(6)	0.4059(7)
promoter	**0.5143(1)**	0.5047(7)	0.5094(4)	0.5094(4)	0.5094(4)	0.5094(4)	0.5094(4)	0.5368(3)	**0.5472(1)**	0.5189(7)	0.5377(2)	0.5330(4.5)	0.5283(6)	0.5330(4.5)
hayes	**0.4188(1)**	0.4125(6)	0.4125(6)	0.4156(3)	0.4125(6)	0.4156(3)	0.4156(3)	0.4406(4.5)	0.4219(6.5)	**0.4656(1)**	0.4531(3)	0.4406(4.5)	0.4219(6.5)	0.4594(2)
dermatology	**0.3177(1)**	0.3169(2.5)	0.3128(7)	0.3156(4.5)	0.3142(6)	0.3156(4.5)	0.3169(2.5)	0.3087(5.5)	0.3115(3.5)	0.306(7)	0.3115(3.5)	**0.3169(1)**	0.3087(5.5)	0.3142(2)
vote	**0.6179(1)**	0.6161(2)	0.6138(6)	0.6138(6)	0.6149(3.5)	0.6138(6)	0.6149(3.5)	0.6138(4)	0.6138(4)	0.6138(4)	0.6138(4)	0.6138(4)	0.6138(4)	0.6138(4)
balance	**0.4667(1)**	0.4624(7)	0.4632(5)	0.4632(5)	0.4632(5)	0.4640(2.5)	0.4640(2.5)	**0.4984(1)**	0.4736(5.5)	0.4728(7)	0.4768(4)	0.4736(5.5)	0.4816(2)	0.4792(3)
breastcancer	0.6558(4)	0.6559(2.5)	0.6559(2.5)	**0.6567(1)**	0.6552(6)	0.6552(6)	0.6552(6)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)
krvskp	**0.5225(1)**	0.5224(3)	0.5224(3)	0.5222(6)	0.5224(3)	0.5222(6)	0.5222(6)	**0.5225(1.5)**	0.5222(5.5)	0.5222(5.5)	0.5222(5.5)	0.5222(5.5)	**0.5225(1.5)**	0.5224(3)
mushroom	**0.6184(1)**	0.6180(5.5)	0.6180(5.5)	0.6182(2)	0.6181(3)	0.6180(5.5)	0.6180(5.5)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)
nursery	**0.3363(1)**	0.3335(4.5)	0.3335(4.5)	0.3335(4.5)	0.3335(4.5)	0.3335(4.5)	0.3335(4.5)	**0.3381(1)**	0.3349(6)	0.3371(2)	0.3344(7)	0.335(5)	0.3355(4)	0.3367(3)
basehock	**0.5118(1)**	0.5015(3)	0.5013(5.5)	0.5013(5.5)	0.5018(2)	0.5013(5.5)	0.5013(5.5)	**0.5103(1)**	0.5043(3)	0.5080(2)	0.5028(4)	0.5023(5.5)	0.5013(7)	0.5023(5.5)
pcmac	0.5057(2.5)	0.5054(6)	0.5057(2.5)	0.5054(6)	0.5054(6)	0.5057(2.5)	0.5057(2.5)	0.5067(4)	0.5054(6.5)	0.5054(6.5)	0.5093(2)	0.5062(5)	0.5080(3)	**0.5134(1)**
average ranks	1.38	4.50	4.54	4.50	4.33	4.42	4.33	2.88	4.50	4.42	3.75	4.42	4.46	3.58
Data sets	AL							CSPA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	**SIVID**	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	0.4389(2)	0.4208(4.5)	0.4208(4.5)	0.4109(6.5)	0.4109(6.5)	0.4356(3)	**0.4406(1)**	0.4109(5)	0.4257(2)	0.4059(6.5)	0.4158(3.5)	0.4158(3.5)	0.4059(6.5)	**0.4307(1)**
promoter	0.5630(2)	0.5283(4)	0.5189(6.5)	0.5236(5)	0.5189(6.5)	**0.5660(1)**	0.5519(3)	**0.5887(1)**	0.5472(4.5)	0.5472(4.5)	0.5708(3)	0.5283(6)	0.5755(2)	0.5142(7)
hayes	0.4438(3)	0.4406(4.5)	0.4188(7)	**0.4563(1.5)**	0.4406(4.5)	**0.4563(1.5)**	0.4219(6)	**0.4500(1)**	0.4469(2.5)	0.4188(7)	0.4375(6)	0.4469(2.5)	0.4438(4)	0.4406(5)
dermatology	0.3187(2)	0.3101(5.5)	0.306(7)	0.3115(4)	0.3101(5.5)	**0.3197(1)**	0.3128(3)	**0.316(1)**	0.3142(2)	0.3115(4)	0.3074(7)	0.3087(6)	0.3115(4)	0.3115(4)
vote	**0.6793(1)**	0.6138(3.5)	0.5614(7)	0.6138(3.5)	0.6133(6)	0.6138(3.5)	0.6138(3.5)	**0.6793(1)**	0.6138(4.5)	0.6138(4.5)	0.6038(7)	0.6138(4.5)	0.6238(2)	0.6138(4.5)
balance	**0.4963(1)**	0.4904(2.5)	0.4784(6)	0.4776(7)	0.4832(5)	0.4848(4)	0.4904(2.5)	**0.4974(1)**	0.4840(4)	0.4800(5)	0.4968(2)	0.4728(7)	0.4944(3)	0.4760(6)
breastcancer	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)
krvskp	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)
mushroom	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)
nursery	**0.3588(1)**	0.3392(4)	0.3405(2)	0.3387(5)	0.3374(6)	0.3401(3)	0.3369(7)	**0.3566(1)**	0.3339(7)	0.3363(3)	0.3343(6)	0.3344(5)	0.3349(4)	0.3377(2)
basehock	0.5148(2)	0.5065(5)	0.5073(4)	0.5048(7)	0.5063(6)	0.5088(3)	**0.5163(1)**	**0.5076(1)**	0.5033(6)	0.5045(5)	0.5023(7)	0.5068(2)	0.5053(4)	0.5063(3)
pcmac	**0.5275(1)**	0.5211(3)	0.5268(2)	0.5144(5)	0.5095(7)	0.5201(4)	0.5103(6)	**0.5278(1)**	0.5178(6)	0.5237(4)	0.5198(5)	0.5108(7)	0.5247(2)	0.5242(3)
average ranks	2.25	4.04	4.83	4.71	5.42	3.00	3.75	2.08	4.21	4.63	4.88	4.63	3.63	3.96
Data sets	HGPA							MCLA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	**SIVID**	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.4356(1.5)**	0.4257(3)	0.4109(5.5)	0.4158(4)	0.4059(7)	**0.4356(1.5)**	0.4109(5.5)	0.4307(2)	0.4109(6.5)	0.4158(4.5)	**0.4356(1)**	0.4257(3)	0.4158(4.5)	0.4109(6.5)
promoter	0.5717(2)	0.5094(7)	0.5472(4)	0.5330(5)	0.5519(3)	**0.5755(1)**	0.5283(6)	0.5755(2)	**0.5802(1)**	0.5613(3)	0.5283(6)	0.5330(5)	0.5377(4)	0.5236(7)
hayes	**0.4731(1)**	0.4313(7)	0.4563(2.5)	0.4469(5)	0.4406(6)	0.4563(2.5)	0.4500(4)	**0.4813(1)**	0.4156(7)	0.4625(2)	0.4531(4.5)	0.4531(4.5)	0.4500(6)	0.4563(3)
dermatology	**0.3208(1)**	0.3169(2)	0.3156(3)	0.3074(6.5)	0.3074(6.5)	0.3128(5)	0.3142(4)	**0.3160(1)**	0.3060(7)	0.3142(3.5)	0.3142(3.5)	0.3128(6)	0.3142(3.5)	0.3142(3.5)
vote	**0.6379(1)**	0.6138(3.5)	0.6138(3.5)	0.6038(7)	0.6138(3.5)	0.6114(6)	0.6138(3.5)	**0.6793(1)**	0.6138(4.5)	0.6138(4.5)	0.5793(7)	0.6138(4.5)	0.6179(2)	0.6138(4.5)
balance	**0.4967(1)**	0.4920(3)	0.4880(4)	0.4776(6)	0.4936(2)	0.4736(7)	0.4800(5)	**0.5736(1)**	0.4672(7)	0.4816(5)	0.4944(4)	0.5016(3)	0.5032(2)	0.4720(6)
breastcancer	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)	0.6552(4)
krvskp	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)	0.5222(4)
mushroom	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)	0.6180(4)
nursery	**0.4691(1)**	0.4622(4)	0.4622(4)	0.4622(4)	0.4219(7)	0.4622(4)	0.4622(4)	**0.3698(1)**	0.3402(2)	0.3394(4)	0.3378(7)	0.3386(5)	0.3385(6)	0.3401(3)
basehock	0.5054(4)	0.5090(3)	0.5093(2)	**0.5153(1)**	0.5013(6.5)	0.5053(5)	0.5013(6.5)	**0.5179(1)**	0.5060(6)	0.5161(3)	0.5143(4)	0.5028(7)	0.5178(2)	0.5073(5)
pcmac	**0.5154(1)**	0.5054(7)	0.5106(3)	0.5106(3)	0.5059(6)	0.5106(3)	0.5098(5)	**0.5144(1)**	0.5054(6)	0.5080(3.5)	0.5139(2)	0.5054(6)	0.5080(3.5)	0.5054(6)
average ranks	2.13	4.29	3.63	4.46	4.96	3.92	4.63	1.92	4.92	3.75	4.25	4.67	3.79	4.71

Table 4
Results of ARI values for the compared algorithms using the *k*-modes clustering algorithm generating base clusterings.

Data sets	SL							CL						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.8607(1)**	0.6165(2.5)	0.6060(4)	0.6050(5)	0.6165(2.5)	0.5871(6)	0.5802(7)	**0.7692(1)**	0.6767(2.5)	0.6612(4)	0.6604(5)	0.6767(2.5)	0.5802(7)	0.6541(6)
promoter	**0.0006(1)**	0.0001(4)	0.0000(6.5)	0.0001(4)	0.0005(2)	0.0001(4)	0.0000(6.5)	0.0596(2)	**0.1165(1)**	0.0465(4)	0.0453(5)	0.0431(6)	0.0404(7)	0.0541(3)
hayes	**0.0008(1)**	−0.0008(7)	0.0003(3.5)	0.0000(5)	−0.0006(6)	0.0006(2)	0.0003(3.5)	0.0036(2)	−0.0039(7)	0.0006(5)	0.0015(4)	**0.0047(1)**	−0.0003(6)	0.0027(3)
dermatology	**0.2550(1)**	0.2274(3.5)	0.2340(2)	0.1672(5)	0.2274(3.5)	0.0997(7)	0.1607(6)	**0.6852(1)**	0.6329(2.5)	0.4628(7)	0.5435(4)	0.6329(2.5)	0.4893(6)	0.4910(5)
vote	**0.0498(1)**	0.0001(7)	0.0012(4)	0.0011(5)	0.0003(6)	0.0020(2)	0.0014(3)	**0.463(1)**	0.2614(2)	0.2375(3)	0.0478(5)	0.2341(4)	0.0181(7)	0.0469(6)
balance	**0.0005(1)**	−0.0003(5.5)	−0.0001(3)	0.0004(2)	−0.0003(5.5)	−0.0003(5.5)	−0.0003(5.5)	0.0039(4)	**0.0111(1)**	0.0044(3)	0.0022(7)	0.0053(2)	0.0035(6)	0.0036(5)
breastcancer	**0.0792(1)**	0.0028(2)	0.0022(5)	0.0025(3)	0.0022(5)	0.0016(7)	0.0022(5)	**0.4849(1)**	0.4534(2)	0.2069(4)	0.3531(3)	0.1587(7)	0.1770(5)	0.1654(6)
krvskp	−0.0001(4)	−0.0001(4)	−0.0005(7)	−0.0003(6)	−0.0001(4)	**0.0000(1.5)**	**0.0000(1.5)**	**0.0066(1)**	0.0039(2)	−0.0002(7)	0.0004(6)	0.0012(5)	0.0027(4)	0.0037(3)
mushroom	**0.2369(1)**	0.1263(3)	0.0609(6)	0.1372(2)	0.0028(7)	0.0953(4)	0.0610(5)	**0.1467(1)**	0.0213(6)	0.0891(2)	0.0808(3)	0.0206(7)	0.0570(4)	0.0251(5)
nursery	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	**0.0122(1)**	0.0005(3)	0.0001(7)	0.0003(4)	0.0002(5.5)	0.0006(2)	0.0002(5.5)
basehock	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	**0.0149(1)**	0.0146(2.5)	0.0146(2.5)	0.0000(7)	0.0001(6)	0.0006(5)	0.0034(4)
pcmac	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	**0.0095(1)**	0.0042(3)	0.0007(6)	0.0006(7)	0.0025(5)	0.0080(2)	0.0027(4)
average ranks	2.00	4.21	4.42	4.08	4.46	4.25	4.58	1.42	2.88	4.54	5.00	4.46	5.08	4.63
Data sets	AL							CSPA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.8712(1)**	0.6921(3.5)	0.6687(5)	0.7197(2)	0.6921(3.5)	0.6397(7)	0.6455(6)	**0.4782(1)**	0.4596(6.5)	0.4663(5)	0.4746(2)	0.4596(6.5)	0.4686(3)	0.4675(4)
promoter	0.1444(2)	0.1147(4)	0.0542(6)	**0.1515(1)**	0.1000(5)	0.0310(7)	0.1321(3)	**0.1884(1)**	0.0750(5)	0.0467(6)	0.1183(3)	0.1207(2)	0.0138(7)	0.1150(4)
hayes	**−0.0020(1)**	−0.0088(7)	−0.0064(4)	−0.0066(5)	−0.0043(2)	−0.0062(3)	−0.0084(6)	**−0.0051(1)**	−0.0097(6)	−0.0099(7)	−0.0092(5)	−0.0074(2)	−0.0079(3)	−0.0086(4)
dermatology	**0.6779(1)**	0.6659(4.5)	0.6668(3)	0.6633(6)	0.6659(4.5)	0.6448(7)	0.6740(2)	**0.5760(1)**	0.5629(2.5)	0.5516(6)	0.5560(4)	0.5629(2.5)	0.5242(7)	0.5559(5)
vote	**0.5374(1)**	0.4868(5)	0.4876(4)	0.4847(6)	0.4825(7)	0.5129(2)	0.4941(3)	0.4966(4)	0.4868(6)	0.4978(2.5)	0.4913(5)	0.4855(7)	**0.5050(1)**	0.4978(2.5)
balance	**0.0270(1)**	0.0213(5)	0.0248(3)	0.018(7)	0.0251(2)	0.0209(6)	0.0232(4)	0.0363(4)	0.0410(3)	0.0322(6)	0.0352(5)	**0.0486(1)**	0.0320(7)	0.0412(2)
breastcancer	0.8013(3)	0.8010(4)	0.8021(2)	**0.8039(1)**	0.7737(7)	0.7768(6)	0.7991(5)	**0.4424(1)**	0.4215(2.5)	0.4133(5)	0.4203(4)	0.4032(6)	0.3721(7)	0.4215(2.5)
krvskp	0.0277(4)	**0.038(1)**	0.0261(5)	0.0255(6)	0.0194(7)	0.0334(2)	0.0294(3)	**0.0088(1)**	0.0032(6)	0.0073(2)	0.0052(5)	0.0016(7)	0.0056(3)	0.0055(4)
mushroom	0.4384(4)	0.4355(5)	**0.4879(1)**	0.4405(3)	0.3963(7)	0.4182(6)	0.4521(2)	0.2148(3)	0.1671(4)	0.1496(5)	0.2192(2)	**0.2323(1)**	0.1103(7)	0.1309(6)
nursery	**0.1857(1)**	0.1051(3)	0.0699(6)	0.1729(2)	0.0685(7)	0.0735(5)	0.0743(4)	0.2903(2)	**0.3483(1)**	0.0637(7)	0.1981(3)	0.1900(5)	0.1914(4)	0.1033(6)
basehock	**0.0076(1)**	0.0073(2.5)	0.0073(2.5)	0.0000(5.5)	0.0000(5.5)	0.0000(5.5)	0.0000(5.5)	**0.0140(1)**	0.0103(2)	0.0101(3)	0.0074(5)	−0.0005(7)	0.0055(6)	0.0100(4)
pcmac	**0.0077(1)**	0.0004(4)	0.0000(6)	0.0000(6)	0.0011(3)	0.0000(6)	0.0024(2)	**0.0137(1)**	0.0115(2)	0.0111(3)	0.0063(4)	0.0054(6)	0.0042(7)	0.0059(5)
average ranks	1.75	4.04	3.96	4.21	5.04	5.21	3.79	1.75	3.88	4.79	3.92	4.42	5.17	4.08
Data sets	HGPA							MCLA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.5174(1)**	0.5039(2.5)	0.4907(5)	0.4962(4)	0.5039(2.5)	0.4760(7)	0.4890(6)	0.7870(2)	**0.7911(1)**	0.7538(4)	0.7327(5)	0.7638(3)	0.7231(6)	0.7106(7)
promoter	0.2736(2)	**0.3097(1)**	0.2635(4)	0.2676(3)	0.2258(5)	0.1758(7)	0.1970(6)	**0.3395(1)**	0.2929(2)	0.2078(4)	0.2094(3)	0.1535(6)	0.0742(7)	0.1826(5)
hayes	**−0.0003(1)**	−0.0023(5)	−0.0022(4)	−0.002(3)	−0.0025(6)	−0.0004(2)	−0.0027(7)	**−0.0038(1)**	−0.0093(7)	−0.0061(3)	−0.0083(5)	−0.0045(2)	−0.0076(4)	−0.0092(6)
dermatology	0.6057(5)	0.6075(4)	0.6043(6)	**0.6302(1)**	0.6156(2)	0.6042(7)	0.6140(3)	**0.6763(1)**	0.6706(2)	0.6644(4)	0.6335(6)	0.6705(3)	0.5255(7)	0.6548(5)
vote	**0.5739(1)**	0.5642(2)	0.5412(7)	0.5505(5)	0.5638(3)	0.5605(4)	0.5445(6)	**0.5138(1)**	0.4902(7)	0.4997(3)	0.4932(5)	0.4925(6)	0.4977(4)	0.5017(2)
balance	0.0180(3)	**0.0215(1)**	0.0177(4)	0.0159(5)	0.0202(2)	0.0121(7)	0.0123(6)	0.0384(3)	**0.0411(1)**	0.0294(5)	0.0311(4)	0.0388(2)	0.0267(6)	0.0228(7)
breastcancer	**0.5467(1)**	0.5445(2)	0.4899(6)	0.5023(4)	0.4958(5)	0.5315(3)	0.4853(7)	**0.7778(1)**	0.7014(3)	0.7213(2)	0.6727(5)	0.5963(6)	0.5430(7)	0.6762(4)
krvskp	**0.0025(1)**	0.0008(5.5)	0.0008(5.5)	0.0005(7)	0.0010(3.5)	0.0019(2)	0.0010(3.5)	**0.0071(1)**	0.0010(7)	0.0019(6)	0.0021(4.5)	0.0021(4.5)	0.0050(2)	0.0024(3)
mushroom	0.1466(2)	0.0339(7)	0.1304(4)	**0.1591(1)**	0.0583(6)	0.1465(3)	0.0747(5)	0.2313(2)	0.1602(7)	0.2285(3)	0.2144(4)	0.1696(6)	**0.2554(1)**	0.2067(5)
nursery	**0.1598(1)**	0.0525(5)	0.0680(4)	0.0273(6)	0.0236(7)	0.1369(2)	0.1172(3)	0.3406(2)	0.2755(6)	0.3107(4)	**0.3452(1)**	0.3023(5)	0.1561(7)	0.3339(3)
basehock	**0.0319(1)**	0.0115(4)	0.0097(6)	0.0132(3)	0.0095(7)	0.0279(2)	0.0111(5)	**0.0126(1)**	0.0101(3)	0.0025(5)	0.0103(2)	0.0000(7)	0.0042(4)	0.0022(6)
pcmac	0.0131(2)	**0.0159(1)**	0.0050(5)	0.0027(6)	0.0072(4)	0.0020(7)	0.0080(3)	**0.0083(1)**	0.0023(6.5)	0.0033(4)	0.0023(6.5)	0.0059(3)	0.0070(2)	0.0029(5)
average ranks	1.75	3.33	5.04	4.00	4.42	4.42	5.04	1.42	4.38	3.92	4.25	4.46	4.75	4.83

Table 5
Results of ARI values for the compared algorithms using the simply random partition algorithm generating base clusterings.

Data sets	SL							CL						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	−**0.0011(1.5)**	−0.0043(3.5)	−0.0111(6)	−0.0221(7)	−0.0043(3.5)	−**0.0011(1.5)**	−0.0104(5)	−0.0028(5)	−0.005(6)	0.0007(4)	**0.0119(1)**	0.0028(2)	0.0021(3)	−0.0244(7)
promoter	**0.0003(1)**	−0.0004(7)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	−0.0015(5)	**0.0051(1)**	−0.0040(6)	0.0005(2)	0.0002(3)	−0.0044(7)	−0.0007(4)
hayes	**0.0063(1)**	0.0049(4)	−0.0033(7)	0.005(2.5)	0.0008(6)	0.0041(5)	0.0050(2.5)	0.0030(4)	−0.0021(5)	**0.0113(1)**	−0.0035(6)	0.004(3)	−0.0106(7)	0.0062(2)
dermatology	**0.0022(1)**	0.0010(2)	−0.0030(7)	−0.0004(6)	0.0004(3)	0.0000(4)	−0.0003(5)	0.0069(2)	−0.0031(7)	−0.0011(5)	0.0023(3)	**0.0076(1)**	0.0011(4)	−0.0025(6)
vote	**0.0049(1)**	0.0027(2)	−0.0004(5.5)	−0.0017(7)	0.0005(3.5)	−0.0004(5.5)	0.0005(3.5)	0.0022(3)	**0.006(1)**	−0.0123(7)	−0.0049(5)	−0.0105(6)	0.0026(2)	−0.0032(4)
balance	**0.0041(1)**	−0.0007(4.5)	−0.0007(4.5)	0.0035(2)	−0.0019(7)	−0.0007(4.5)	−0.0007(4.5)	0.0048(2)	**0.0060(1)**	−0.0030(6)	−0.0044(7)	0.0008(4)	0.0044(3)	−0.0021(5)
breastcancer	**0.0053(1)**	0.0006(3)	−0.0001(5)	0.0026(2)	−0.0020(7)	−0.0014(6)	0.0005(4)	0.0035(2)	−0.0070(6)	−0.0062(5)	−0.0013(4)	**0.0082(1)**	−0.0002(3)	−0.0082(7)
krvskp	**0.0003(1)**	0.0000(3)	0.0000(3)	−0.0001(6)	0.0000(3)	−0.0001(6)	−0.0001(6)	**0.0003(1)**	0.0000(3)	−0.0004(6.5)	−0.0004(6.5)	−0.0001(5)	0.0000(3)	0.0000(3)
mushroom	**0.0011(1)**	−0.0001(4)	−0.0003(7)	0.0002(2)	−0.0001(4)	−0.0001(4)	−0.0002(6)	**0.0003(1)**	−0.0004(2)	−0.0019(6)	−0.001(4)	−0.0008(3)	−0.0021(7)	−0.0017(5)
nursery	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)
basehock	**0.0003(1)**	0.0000(4.5)	0.0000(4.5)	0.0000(4.5)	0.0000(4.5)	0.0000(4.5)	0.0000(4.5)	**0.0003(1.5)**	0.0000(4.5)	**0.0003(1.5)**	0.0000(4.5)	0.0000(4.5)	−0.0002(7)	0.0000(4.5)
pcmac	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	**0.0006(1)**	−0.0001(6.5)	−0.0001(6.5)	0.0002(3)	0.0000(4.5)	0.0000(4.5)	0.0005(2)
average ranks	1.54	3.79	5.13	4.25	4.46	4.42	4.42	2.63	3.92	4.88	4.17	3.42	4.54	4.46
Data sets	AL							CSPA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.0247(1)**	0.0100(3)	−0.0094(7)	−0.0049(6)	−0.0010(5)	0.0136(2)	−0.0001(4)	0.0069(4)	0.0030(7)	0.0072(3)	0.0077(2)	0.0044(6)	0.0049(5)	**0.0085(1)**
promoter	0.0060(3)	−0.0056(4)	−0.0077(7)	−0.0063(5)	−0.0073(6)	0.0092(2)	**0.0116(1)**	**0.0235(1)**	−0.0003(5)	0.0026(4)	0.0114(3)	−0.0060(6)	0.0148(2)	−0.0087(7)
hayes	−0.0038(5)	0.0009(3)	−0.0039(6)	0.0066(2)	−0.0013(4)	**0.0153(1)**	−0.0078(7)	−0.0012(3)	−0.0031(4)	−0.0088(7)	−0.0046(6)	−**0.0003(1)**	−0.0005(2)	−0.0036(5)
dermatology	0.0004(2)	−0.0013(5)	−0.0046(7)	−0.0009(4)	**0.0036(1)**	0.0003(3)	−0.0017(6)	**0.0017(1.5)**	0.0014(3)	**0.0017(1.5)**	−0.0019(7)	−0.0006(5)	−0.0010(6)	−0.0003(4)
vote	−0.0022(4.5)	−0.0021(3)	−0.0111(7)	−0.0022(4.5)	−0.0017(2)	**0.0011(1)**	−0.0041(6)	−0.0018(6)	−0.0017(4.5)	−0.0017(4.5)	−0.0016(3)	**0.0019(1)**	−0.0001(2)	−0.0019(7)
balance	0.0031(4)	**0.0037(1.5)**	0.0001(5)	0.0032(3)	−0.0006(6)	**0.0037(1.5)**	−0.0007(7)	−0.0016(5.5)	0.0002(3)	−0.0007(4)	**0.0023(1)**	−0.0019(7)	0.0014(2)	−0.0016(5.5)
breastcancer	**0.0064(1)**	−0.0005(6)	−0.0016(7)	0.0001(3.5)	0.0001(3.5)	0.0000(5)	0.0062(2)	**0.0042(1)**	−0.0006(5)	−0.0013(7)	−0.0010(6)	−0.0002(2)	−0.0004(3)	−0.0005(4)
krvskp	0.0002(2.5)	**0.0003(1)**	−0.0003(6)	0.0000(5)	−0.0005(7)	0.0002(2.5)	0.0001(4)	−0.0001(5)	−0.0003(7)	0.0001(2.5)	0.0000(4)	**0.0005(1)**	−0.0002(6)	0.0001(2.5)
mushroom	0.0016(2)	0.0004(5)	0.0011(3)	−0.0017(6)	−0.0032(7)	0.0007(4)	**0.0019(1)**	0.0001(2)	−0.0001(5.5)	−0.0001(5.5)	−0.0001(5.5)	**0.0002(1)**	0.0000(3)	−0.0001(5.5)
nursery	0.0000(4.5)	0.0000(4.5)	**0.0001(1)**	0.0000(4.5)	0.0000(4.5)	0.0000(4.5)	0.0000(4.5)	**0.0001(1.5)**	0.0000(5)	0.0000(5)	0.0000(5)	0.0000(5)	0.0000(5)	**0.0001(1.5)**
basehock	0.0004(2)	−0.0003(5)	−0.0003(5)	−0.0004(7)	−0.0003(5)	−0.0002(3)	**0.0006(1)**	−0.0005(6)	−0.0005(6)	−0.0004(3.5)	−0.0005(6)	−**0.0003(1.5**)	−0.0004(3.5)	−**0.0003(1.5**)
pcmac	**0.0025(1)**	0.0013(3.5)	0.0024(2)	0.0006(5)	0.0000(6)	0.0013(3.5)	−0.0001(7)	0.0001(6)	0.0009(5)	0.0018(2.5)	0.0014(4)	0.0000(7)	**0.0019(1)**	0.0018(2.5)
average ranks	2.71	3.71	5.25	4.63	4.75	2.75	4.21	3.54	5.00	4.17	4.38	3.63	3.38	3.92
Data sets	HGPA							MCLA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	0.0110(2)	−0.0032(5)	0.0031(4)	0.0034(3)	−0.0133(7)	**0.0193(1)**	−0.0090(6)	0.0149(2)	0.0006(5)	−0.0092(7)	**0.0248(1)**	−0.0076(6)	0.0078(3)	0.0029(4)
promoter	0.0101(2)	−0.0092(7)	−0.0006(4)	−0.0051(5)	0.0035(3)	**0.0137(1)**	−0.0064(6)	0.0138(2)	**0.0237(1)**	0.0057(3)	−0.0049(5)	−0.005(6)	−0.0034(4)	−0.0065(7)
hayes	0.0044(2)	−0.0047(7)	0.0034(5)	0.0038(4)	**0.0048(1)**	0.0007(6)	0.0042(3)	**0.009(1)**	−0.0097(7)	0.0051(4)	0.0064(2)	−0.0008(6)	0.0052(3)	0.0027(5)
dermatology	0.0007(2)	−0.0006(4)	−0.0023(5)	−0.0041(6)	−0.0042(7)	−0.0002(3)	**0.0018(1)**	0.0027(3)	−0.0042(7)	0.0000(4)	−0.0040(6)	0.0030(2)	**0.0035(1)**	−0.0021(5)
vote	**0.0072(1)**	0.0034(2)	−0.0015(7)	−0.0013(6)	−0.0010(4.5)	0.0008(3)	−0.0010(4.5)	0.0015(3)	−0.0021(6.5)	−0.0017(5)	**0.0042(1)**	0.0037(2)	0.0011(4)	−0.0021(6.5)
balance	−0.0002(4)	**0.0025(1)**	0.0006(3)	−0.0009(6)	0.0023(2)	−0.0015(7)	−0.0004(5)	**0.0138(1)**	0.0021(3)	0.0006(4)	−0.0001(5)	−0.0002(6)	0.0132(2)	−0.0036(7)
breastcancer	0.0041(2)	−0.0005(6)	−0.0004(5)	0.0003(3.5)	−0.0013(7)	0.0003(3.5)	**0.0065(1)**	**0.0024(1)**	0.0021(2)	−0.0013(6)	−0.0026(7)	−0.0012(5)	0.0007(3)	−0.0005(4)
krvskp	−0.0003(5.5)	−0.0002(2.5)	−0.0003(5.5)	**0.0000(1)**	−0.0003(5.5)	−0.0003(5.5)	−0.0002(2.5)	**0.0003(1)**	−0.0003(5)	−0.0002(2.5)	−0.0002(2.5)	−0.0003(5)	−0.0003(5)	−0.0005(7)
mushroom	−0.0002(4)	−0.0002(4)	−0.0002(4)	−0.0002(4)	−0.0002(4)	−0.0002(4)	−0.0002(4)	0.0006(2)	−0.0016(7)	−0.0001(5)	0.0001(3)	**0.0009(1)**	0.0000(4)	−0.0007(6)
nursery	0.0547(2)	**0.0550(1)**	0.0543(6)	0.0546(3.5)	0.0545(5)	0.0546(3.5)	0.0542(7)	**0.0004(1)**	0.0000(4.5)	0.0000(4.5)	−0.0001(7)	0.0000(4.5)	0.0000(4.5)	0.0001(2)
basehock	**0.0005(1)**	0.0000(4)	0.0001(3)	0.0004(2)	−0.0005(6.5)	−0.0004(5)	−0.0005(6.5)	0.0003(3.5)	−0.0003(5.5)	0.0007(2)	0.0003(3.5)	−0.0005(7)	**0.0009(1)**	−0.0003(5.5)
pcmac	**0.0005(1)**	−0.0005(7)	0.0000(2.5)	0.0000(2.5)	−0.0004(6)	−0.0001(4.5)	−0.0001(4.5)	**0.0006(1)**	−0.0005(6)	−0.0003(3.5)	0.0005(2)	−0.0005(6)	−0.0003(3.5)	−0.0005(6)
average ranks	2.38	4.21	4.50	3.88	4.88	3.92	4.25	1.79	4.96	4.21	3.75	4.71	3.17	5.42

Table 6
Results of NMI values for the compared algorithms using the *k*-modes clustering algorithm generating base clusterings.

Data sets	SL							CL						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.8665(1)**	0.7703(2.5)	0.7627(6)	0.7652(4)	0.7703(2.5)	0.7514(7)	0.7634(5)	**0.8447(1)**	0.8192(2.5)	0.8081(5)	0.8030(6)	0.8192(2.5)	0.7611(7)	0.8121(4)
promoter	0.0369(2)	0.0360(4)	0.0342(6)	0.0360(4)	**0.0407(1)**	0.0360(4)	0.0334(7)	0.0801(2)	**0.1335(1)**	0.0643(6)	0.0792(3)	0.0675(5)	0.0507(7)	0.0784(4)
hayes	**0.6053(1)**	0.0546(5)	0.0604(2)	0.0549(4)	0.0551(3)	0.0542(6)	0.0525(7)	**0.0165(1)**	0.0067(7)	0.0140(4)	0.0142(3)	0.0145(2)	0.0093(6)	0.0119(5)
dermatology	**0.5273(1)**	0.4910(2.5)	0.4779(4)	0.402(5)	0.4910(2.5)	0.2779(7)	0.3688(6)	**0.7351(1)**	0.7242(2.5)	0.6115(6)	0.6694(4)	0.7242(2.5)	0.5862(7)	0.6192(5)
vote	**0.0567(1)**	0.0115(5)	0.0150(2)	0.0120(4)	0.0100(6)	0.0087(7)	0.0126(3)	**0.4078(1)**	0.2464(2)	0.2189(4)	0.0775(6)	0.2296(3)	0.0732(7)	0.0902(5)
balance	0.0194(2)	0.0184(6.5)	0.0189(5)	**0.0219(1)**	0.0192(3)	0.0184(6.5)	0.0190(4)	**0.0294(1)**	0.0212(5)	0.0250(3)	0.0276(2)	0.0176(6)	0.0161(7)	0.0222(4)
breastcancer	**0.0813(1)**	0.0192(2)	0.0168(6)	0.0182(3)	0.0175(4)	0.0155(7)	0.0172(5)	**0.4435(1)**	0.4089(2)	0.2270(4)	0.3287(3)	0.1991(6)	0.2182(5)	0.1883(7)
krvskp	**0.0057(1)**	0.0046(7)	0.0048(4.5)	0.0052(3)	0.0053(2)	0.0048(4.5)	0.0047(6)	0.0115(5)	**0.0207(1)**	0.0085(6)	0.0068(7)	0.0150(4)	0.0158(2)	0.0155(3)
mushroom	**0.2554(1)**	0.2394(2)	0.1745(6)	0.2229(3)	0.1181(7)	0.2077(4)	0.1771(5)	**0.1657(1)**	0.0746(6)	0.1505(2)	0.1151(3)	0.0614(7)	0.0826(5)	0.0932(4)
nursery	**0.0037(1)**	0.0036(3.5)	0.0036(3.5)	0.0034(6)	0.0036(3.5)	0.0029(7)	0.0036(3.5)	0.0176(2)	**0.0190(1)**	0.0123(7)	0.0155(5)	0.0173(3)	0.0149(6)	0.0172(4)
basehock	**0.0094(1)**	0.0063(4.5)	0.0063(4.5)	0.0063(4.5)	0.0063(4.5)	0.0063(4.5)	0.0063(4.5)	0.0264(2)	0.0111(5.5)	0.0111(5.5)	0.0063(7)	**0.0298(1)**	0.0161(3)	0.0157(4)
pcmac	**0.0095(1)**	0.0066(4.5)	0.0066(4.5)	0.0066(4.5)	0.0066(4.5)	0.0066(4.5)	0.0066(4.5)	**0.0149(1)**	0.0035(5)	0.0009(7)	0.0039(4)	0.0147(2)	0.0110(3)	0.0031(6)
average ranks	1.17	4.08	4.50	3.83	3.63	5.75	5.04	1.58	3.38	4.96	4.42	3.67	5.42	4.58
Data sets	AL							CSPA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.8748(1)**	0.8150(3.5)	0.8072(5)	0.8178(2)	0.8150(3.5)	0.7830(7)	0.8040(6)	**0.7099(1)**	0.6994(5.5)	0.7049(4)	0.7097(2)	0.6994(5.5)	0.6990(7)	0.7052(3)
promoter	**0.1540(1)**	0.1239(4)	0.0612(6)	0.1327(2)	0.0992(5)	0.0416(7)	0.1262(3)	**0.1545(1)**	0.0636(5)	0.0419(6)	0.0964(3)	0.0986(2)	0.0169(7)	0.0939(4)
hayes	**0.0105(1)**	0.0033(7)	0.0057(5)	0.0062(3.5)	0.0074(2)	0.0062(3.5)	0.0039(6)	**0.0066(1)**	0.0022(6)	0.0021(7)	0.0027(5)	0.0043(2)	0.0036(3)	0.0032(4)
dermatology	**0.7894(1)**	0.7769(2.5)	0.7746(4)	0.7644(6)	0.7769(2.5)	0.7333(7)	0.7690(5)	**0.6891(1)**	0.6535(2.5)	0.6394(6)	0.6451(5)	0.6535(2.5)	0.6094(7)	0.6452(4)
vote	**0.4587(1)**	0.4191(7)	0.4226(5)	0.4256(4)	0.4201(6)	0.4428(2)	0.4316(3)	0.4500(2)	0.4388(5)	**0.4512(1)**	0.4437(4)	0.4373(6)	0.4459(3)	0.4226(7)
balance	0.0278(2)	0.0277(3)	0.0268(4.5)	0.0222(7)	**0.0311(1)**	0.0268(4.5)	0.0239(6)	0.0215(5)	0.035(2)	0.0282(4)	0.0306(3)	**0.0415(1)**	0.0211(6)	0.0136(7)
breastcancer	0.6937(2)	0.6927(3)	0.6925(4)	**0.6944(1)**	0.6717(6)	0.6689(7)	0.6900(5)	**0.4446(1)**	0.4161(3)	0.4077(5)	0.416(4)	0.3968(6)	0.3632(7)	0.4169(2)
krvskp	0.0272(4)	**0.0364(1)**	0.0246(5)	0.0279(3)	0.0229(7)	0.0325(2)	0.0241(6)	**0.0066(1)**	0.0025(6)	0.0055(2)	0.0040(5)	0.0014(7)	0.0043(3)	0.0042(4)
mushroom	**0.4833(1)**	0.4323(6)	0.4662(2)	0.4342(5)	0.4120(7)	0.4357(4)	0.4378(3)	0.1975(2)	0.1425(4)	0.1256(5)	0.1872(3)	**0.1991(1)**	0.0920(7)	0.1093(6)
nursery	0.0201(7)	0.1181(3)	0.0822(6)	**0.1807(1)**	0.0854(4)	0.0835(5)	0.1524(2)	0.1931(3)	**0.3654(1)**	0.0597(7)	0.1980(2)	0.1845(5)	0.1864(4)	0.1076(6)
basehock	**0.0094(1)**	0.0087(2.5)	0.0087(2.5)	0.0063(5.5)	0.0063(5.5)	0.0063(5.5)	0.0063(5.5)	0.0075(4)	**0.0078(1)**	0.0076(2.5)	0.0057(5)	0.0000(7)	0.0043(6)	0.0076(2.5)
pcmac	0.0102(2)	0.0038(6.5)	0.0038(6.5)	0.0080(3)	0.0062(5)	0.0065(4)	**0.0107(1)**	**0.0094(1)**	0.0087(2)	0.0084(3)	0.0049(4)	0.0043(6)	0.0034(7)	0.0046(5)
average ranks	2.00	4.08	4.63	3.58	4.54	4.88	4.29	1.92	3.58	4.38	3.75	4.25	5.58	4.54
Data sets	HGPA							MCLA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.7388(1)**	0.7368(2.5)	0.7200(5)	0.7267(4)	0.7368(2.5)	0.7093(7)	0.7198(6)	0.8267(2.5)	0.8267(2.5)	0.8123(5)	0.8055(6)	**0.8274(1)**	0.7733(7)	0.8147(4)
promoter	0.2147(4)	**0.2548(1)**	0.2159(3)	0.2180(2)	0.1858(5)	0.1476(7)	0.1624(6)	**0.2755(1)**	0.2402(2)	0.1721(4)	0.1750(3)	0.1259(6)	0.0636(7)	0.1510(5)
hayes	**0.0150(1)**	0.0142(2)	0.0140(3)	0.0134(6)	0.0122(7)	0.0135(5)	0.0137(4)	0.0041(2)	0.0026(7)	0.0036(4)	0.0037(3)	**0.0065(1)**	0.0034(5)	0.0027(6)
dermatology	0.7170(4)	0.7193(3)	0.7152(5)	**0.7347(1)**	0.7246(2)	0.7020(7)	0.7122(6)	**0.7689(1)**	0.7281(3)	0.7295(2)	0.7025(6)	0.7278(4)	0.6397(7)	0.7244(5)
vote	0.4744(3)	**0.4802(1)**	0.4704(5)	0.4729(4)	0.4800(2)	0.4697(6)	0.4689(7)	**0.4601(1)**	0.4423(7)	0.4506(3)	0.4447(6)	0.4454(5)	0.4501(4)	0.4535(2)
balance	**0.0197(1.5)**	**0.0197(1.5)**	0.0167(5)	0.0152(6)	0.0186(3)	0.0177(4)	0.0121(7)	**0.0376(1)**	0.0354(2)	0.0260(5)	0.0274(4)	0.0342(3)	0.0232(6)	0.0210(7)
breastcancer	0.4699(5)	0.4161(6)	0.4825(3)	0.4858(2)	0.4800(4)	**0.5018(1)**	0.4141(7)	**0.6664(1)**	0.6028(3)	0.6151(2)	0.5747(5)	0.5302(6)	0.4819(7)	0.5839(4)
krvskp	0.0010(2.5)	0.0008(5.5)	0.0008(5.5)	0.0006(7)	0.0009(4)	**0.0016(1)**	0.0010(2.5)	**0.0053(1)**	0.0009(7)	0.0016(6)	0.0018(4.5)	0.0018(4.5)	0.0039(2)	0.0020(3)
mushroom	0.1309(2)	0.0302(7)	0.1223(4)	**0.1430(1)**	0.0509(6)	0.1237(3)	0.0630(5)	0.1500(3)	0.1352(5)	**0.1901(1)**	0.1769(2)	0.1450(4)	0.1222(6)	0.1180(7)
nursery	**0.1554(1)**	0.0476(6)	0.0654(5)	0.0255(7)	0.1083(4)	0.1428(2)	0.1134(3)	**0.3977(1)**	0.3126(6)	0.3385(5)	0.3823(3)	0.3868(2)	0.1674(7)	0.3650(4)
basehock	0.0132(2)	0.0087(4)	0.0074(6)	0.0099(3)	0.0072(7)	**0.0206(1)**	0.0084(5)	**0.0094(1)**	0.0078(3)	0.0032(6)	0.0079(2)	0.0000(7)	0.0040(5)	0.0048(4)
pcmac	**0.0148(1)**	0.0119(2)	0.0040(5)	0.0023(6)	0.0056(4)	0.0018(7)	0.0062(3)	0.0038(4)	0.0022(7)	0.0058(2)	0.0026(6)	0.0055(3)	**0.0070(1)**	0.0032(5)
average ranks	2.33	3.46	4.54	4.08	4.21	4.25	5.13	1.63	4.54	3.75	4.21	3.88	5.33	4.67

Table 7
Results of NMI values for the compared algorithms using the simply random partition algorithm generating base clusterings.

Data sets	SL							CL						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.1482(1)**	0.1309(4)	0.1245(6)	0.1137(7)	0.1306(5)	0.1439(2)	0.1435(3)	**0.1773(1)**	0.1245(5)	0.1155(6)	0.1400(2)	0.1326(3)	0.1309(4)	0.1012(7)
promoter	**0.0375(1)**	0.0171(7)	0.0342(4)	0.0342(4)	0.0342(4)	0.0342(4)	0.0342(4)	**0.0186(1)**	0.0119(2)	0.0027(7)	0.0090(4)	0.0065(5)	0.0036(6)	0.0095(3)
hayes	**0.066(1)**	0.0565(4)	0.0403(7)	0.0567(2.5)	0.0484(5)	0.048(6)	0.0567(2.5)	0.0125(2)	0.0087(6)	0.0103(5)	0.0108(3)	0.0106(4)	0.0068(7)	**0.0161(1)**
dermatology	0.0617(2)	0.0614(3)	0.0539(7)	0.0597(4.5)	0.0593(6)	0.0597(4.5)	**0.0656(1)**	**0.0288(1.5)**	**0.0288(1.5)**	0.0251(3)	0.0212(5)	0.0250(4)	0.0201(6)	0.0181(7)
vote	0.0146(4)	**0.0210(1)**	0.0058(6.5)	0.0108(5)	0.0159(2.5)	0.0058(6.5)	0.0159(2.5)	0.0031(4)	0.0032(3)	0.0057(2)	0.0021(6)	**0.0084(1)**	0.0028(5)	0.0008(7)
balance	0.0184(2)	0.0168(5.5)	0.0168(5.5)	**0.0291(1)**	0.0134(7)	0.0169(3.5)	0.0169(3.5)	**0.0069(1)**	0.0022(7)	0.0043(3)	0.0064(2)	0.0036(4)	0.0032(6)	0.0033(5)
breastcancer	0.0146(2)	0.0128(4)	0.0145(3)	**0.0183(1)**	0.0090(5)	0.0073(6)	0.0007(7)	**0.0045(1)**	0.0005(6)	0.0043(2)	0.0003(7)	0.0016(4)	0.0011(5)	0.0022(3)
krvskp	**0.0061(1)**	0.0049(4.5)	0.0060(2)	0.0046(6.5)	0.0049(4.5)	0.0057(3)	0.0046(6.5)	**0.0030(1)**	0.0001(5)	0.0029(2)	0.0001(5)	0.0000(7)	0.0002(3)	0.0001(5)
mushroom	**0.0062(1)**	0.0025(6.5)	0.0036(3)	0.0051(2)	0.0030(5)	0.0025(6.5)	0.0031(4)	**0.0011(1)**	0.0002(6)	0.0004(4)	0.0003(5)	0.0009(2)	0.0001(7)	0.0006(3)
nursery	**0.0040(2)**	**0.004(2)**	0.0036(7)	**0.004(2)**	0.0037(5.5)	0.0039(4)	0.0037(5.5)	**0.0012(1.5)**	0.0006(3)	0.0004(4.5)	0.0004(4.5)	**0.0012(1.5)**	0.0003(6.5)	0.0003(6.5)
basehock	**0.0081(1)**	0.0064(2)	0.0063(3.5)	0.0000(6.5)	0.0010(5)	0.0063(3.5)	0.0000(6.5)	0.0007(3)	0.0008(2)	0.0006(4)	0.0003(5)	**0.0011(1)**	0.0000(7)	0.0002(6)
pcmac	0.0065(2.5)	0.0064(6)	0.0065(2.5)	0.0064(6)	0.0064(6)	0.0065(2.5)	0.0065(2.5)	0.0017(3)	0.0003(6)	0.0009(5)	0.0013(4)	0.0001(7)	**0.0037(1)**	0.0032(2)
average ranks	1.71	4.13	4.75	4.00	5.04	4.33	4.04	1.75	4.38	3.96	4.38	3.63	5.29	4.63
Data sets	AL							CSPA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	0.1457(2)	0.1110(7)	0.1240(5)	0.1308(3)	0.1173(6)	**0.1477(1)**	0.1282(4)	0.1402(2)	0.1369(5)	**0.1425(1)**	0.1398(4)	0.1192(7)	0.1400(3)	0.1265(6)
promoter	**0.0244(1)**	0.0028(4)	0.0011(7)	0.0019(5)	0.0012(6)	0.0155(2)	0.0136(3)	0.0094(3)	0.0067(5)	0.0088(4)	0.0151(2)	0.0026(6)	**0.0176(1)**	0.0006(7)
hayes	0.0160(2)	0.0073(5)	0.0060(6)	0.0155(3)	0.0132(4)	**0.0201(1)**	0.0052(7)	0.0078(5)	0.0082(4)	0.0037(7)	0.0072(6)	**0.0115(1)**	0.0108(2)	0.0084(3)
dermatology	**0.0269(1)**	0.0184(5)	0.0192(4)	0.0169(7)	0.0246(2)	0.0221(3)	0.0182(6)	**0.0231(1)**	0.0210(4)	0.0214(2)	0.0196(6)	0.0197(5)	0.0169(7)	0.0212(3)
vote	0.0031(2)	0.0000(7)	**0.0072(1)**	0.0003(6)	0.0013(3)	0.0011(4)	0.0006(5)	0.0003(6)	0.0004(4.5)	0.0004(4.5)	0.0005(3)	**0.0032(1)**	0.0016(2)	0.0002(7)
balance	**0.0055(1)**	0.0049(4)	0.0029(6)	0.0051(2)	0.0026(7)	0.0045(5)	0.0050(3)	0.0012(5)	**0.0044(1)**	0.0024(4)	0.0038(3)	0.0009(7)	0.0039(2)	0.0011(6)
breastcancer	0.0003(6)	0.0005(4.5)	0.0001(7)	0.0005(4.5)	0.0009(2)	0.0006(3)	**0.0036(1)**	**0.0045(1)**	0.0006(5)	0.0000(7)	0.0002(6)	0.0009(2)	0.0007(3.5)	0.0007(3.5)
krvskp	0.0001(5)	**0.0002(2)**	0.0000(7)	**0.0002(2)**	0.0001(5)	0.0001(5)	**0.0002(2)**	0.0004(2)	0.0000(7)	0.0003(3.5)	0.0002(5)	**0.0006(1)**	0.0001(6)	0.0003(3.5)
mushroom	0.0001(4.5)	**0.0002(1.5)**	0.0001(4.5)	0.0001(4.5)	**0.0002(1.5)**	0.0000(7)	0.0001(4.5)	**0.0004(1)**	0.0000(5.5)	0.0000(5.5)	0.0000(5.5)	0.0003(2)	0.0001(3)	0.0000(5.5)
nursery	0.0001(5)	0.0001(5)	**0.0003(1)**	0.0001(5)	0.0001(5)	0.0002(2)	0.0001(5)	**0.0002(1)**	0.0000(5.5)	0.0001(2.5)	0.0000(5.5)	0.0000(5.5)	0.0000(5.5)	0.0001(2.5)
basehock	**0.0009(1)**	0.0002(4)	0.0002(4)	0.0001(6.5)	0.0001(6.5)	0.0002(4)	0.0008(2)	**0.0009(1)**	0.0000(6.5)	0.0001(3.5)	0.0000(6.5)	0.0001(3.5)	0.0001(3.5)	0.0001(3.5)
pcmac	0.0005(5)	0.0013(2.5)	**0.0020(1)**	0.0008(4)	0.0002(7)	0.0013(2.5)	0.0003(6)	**0.0024(1)**	0.0010(6)	0.0016(4)	0.0014(5)	0.0004(7)	0.0018(2)	0.0017(3)
average ranks	2.96	4.29	4.46	4.38	4.58	3.29	4.04	2.42	4.92	4.04	4.79	4.00	3.38	4.46
Data sets	HGPA							MCLA						
	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS	SIVID	Full	RS	CAS	ACES	SELSCE	HCSS
zoo	**0.1473(1)**	0.1062(5)	0.1243(4)	0.1274(3)	0.0966(7)	0.1437(2)	0.1026(6)	0.1353(2)	0.1007(7)	0.1190(5)	**0.1487(1)**	0.1049(6)	0.1214(3)	0.1203(4)
promoter	0.0095(2)	0.0003(7)	0.0064(4)	0.0032(5)	0.0094(3)	**0.0168(1)**	0.0023(6)	0.0169(2)	**0.0242(1)**	0.011(3)	0.0033(5.5)	0.0033(5.5)	0.0044(4)	0.0022(7)
hayes	0.0182(2)	0.0070(7)	0.0145(5)	**0.0207(1)**	0.0158(4)	0.0109(6)	0.0160(3)	0.0151(3)	0.0036(7)	**0.0188(1)**	0.0111(6)	0.0176(2)	0.0123(5)	0.0145(4)
dermatology	**0.0282(1)**	0.0166(6)	0.0163(7)	0.0239(2)	0.0221(4)	0.0231(3)	0.0179(5)	**0.0349(1)**	0.0206(6)	0.0249(5)	0.0345(2)	0.0263(4)	0.0277(3)	0.0168(7)
vote	**0.0073(1)**	0.0043(2)	0.0005(7)	0.0007(6)	0.0009(4.5)	0.0023(3)	0.0009(4.5)	0.0029(2)	0.0009(6)	0.0011(5)	**0.0056(1)**	0.0024(3)	0.0014(4)	0.0001(7)
balance	0.0059(2)	**0.0068(1)**	0.0033(5)	0.0028(6)	0.0051(3)	0.0012(7)	0.0036(4)	**0.0141(1)**	0.0036(5)	0.0027(6)	0.0060(4)	0.0089(2)	0.0088(3)	0.0015(7)
breastcancer	0.0029(2)	0.0006(6)	0.0007(5)	0.0013(3.5)	0.0000(7)	0.0013(3.5)	**0.0064(1)**	0.0011(2.5)	**0.0029(1)**	0.0007(4.5)	0.0007(4.5)	0.0003(7)	0.0011(2.5)	0.0004(6)
krvskp	**0.0003(1)**	0.0001(3.5)	0.0000(6)	0.0002(2)	0.0000(6)	0.0000(6)	0.0001(3.5)	**0.0004(1)**	0.0001(3.5)	0.0001(3.5)	0.0000(6.5)	0.0000(6.5)	0.0001(3.5)	0.0001(3.5)
mushroom	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	0.0000(4)	**0.0008(1)**	0.0001(4)	0.0000(6)	0.0000(6)	0.0007(2)	0.0000(6)	0.0003(3)
nursery	0.0487(2)	**0.0489(1)**	0.0480(6)	0.0482(5)	0.0483(4)	0.0484(3)	0.0479(7)	**0.0004(1)**	0.0002(2.5)	0.0001(5.5)	0.0001(5.5)	0.0001(5.5)	0.0001(5.5)	0.0002(2.5)
basehock	0.0002(4)	0.0003(3)	0.0004(2)	**0.0007(1)**	0.0000(6.5)	0.0001(5)	0.0000(6.5)	**0.0017(1)**	0.0001(5.5)	0.0010(3)	0.0006(4)	0.0000(7)	0.0014(2)	0.0001(5.5)
pcmac	0.0002(5)	0.0000(7)	**0.0004(1.5)**	**0.0004(1.5)**	0.0001(6)	0.0003(3.5)	0.0003(3.5)	**0.0008(1)**	0.0000(6)	0.0001(4)	0.0007(2)	0.0000(6)	0.0002(3)	0.0000(6)
average ranks	2.25	4.38	4.71	3.33	4.92	3.92	4.50	1.54	4.54	4.29	4.00	4.71	3.71	5.21

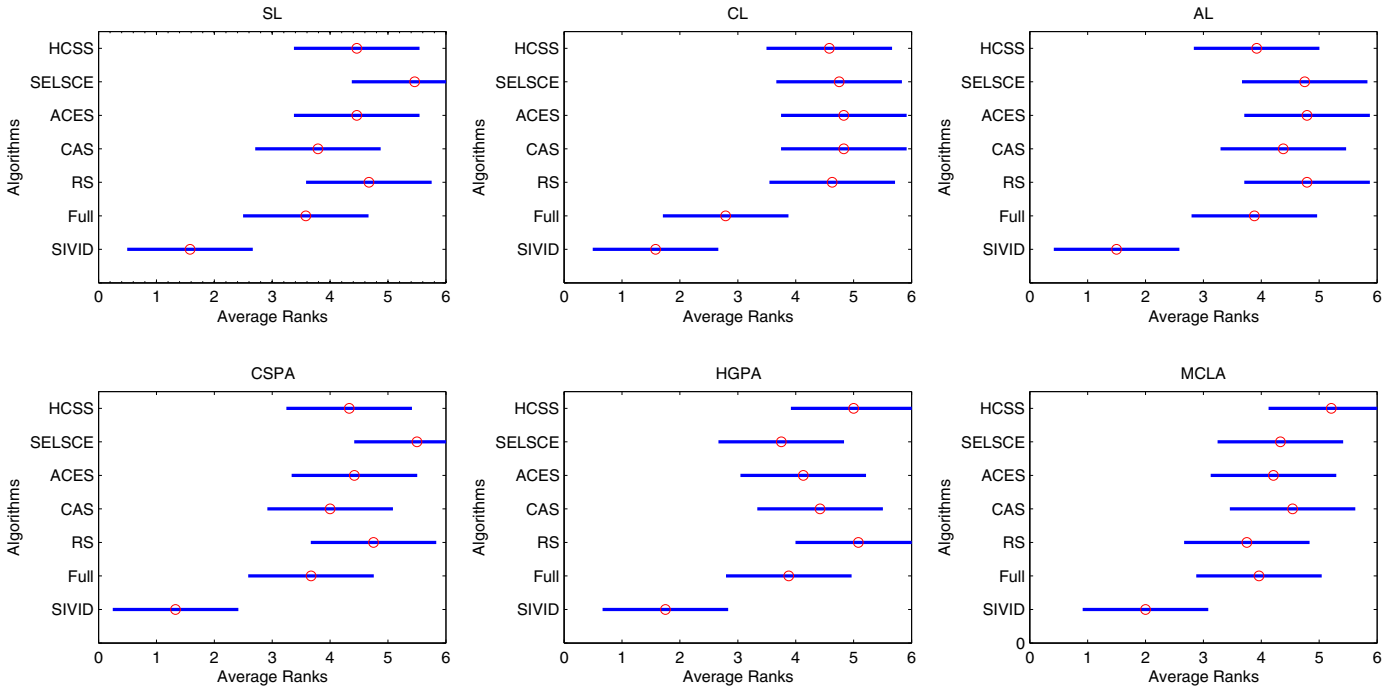

Fig. 2. Nemenyi tests for 50% base clusterings using the k-modes clustering algorithm generating base clusterings in terms of CA.

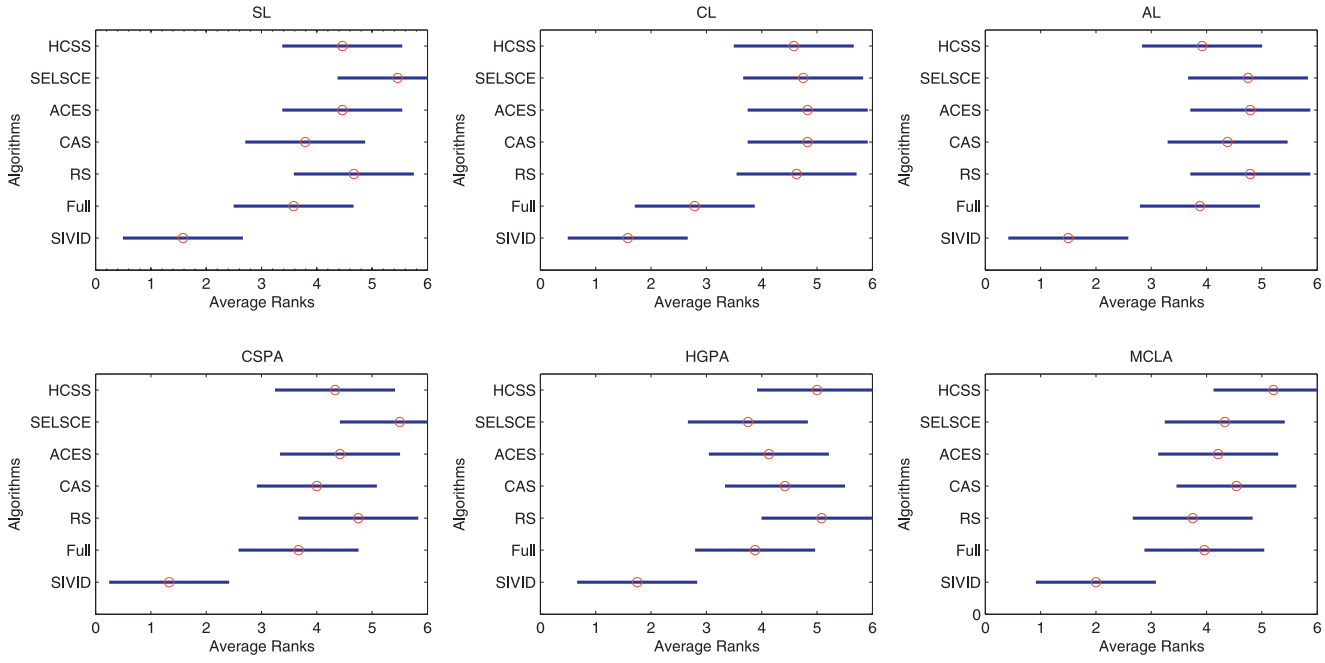

Fig. 3. Nemenyi tests for 50% base clusterings using the k-modes clustering algorithm generating base clusterings in terms of ARI.

dicate that the quality of clustering results produced by the RS algorithm are worse than those obtained by the other ensemble selection algorithms. In general, the SIVID algorithm is the most suitable for performing clustering ensemble selection when compared with most of the existing algorithms. Similar experimental results with these algorithms are observed using ARI and NMI evaluation indices in Tables 4–7.

In order to give a comprehensive comparison, we further perform the Friedman test and Nemenyi test [47] to analyze the differences between the SIVID algorithm and the other algorithms. For Friedman test, there are $A = 7$ algorithms, $B = 36$ cases (i.e., 2 ensemble generation methods, 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 SIVID algorithm ranks 1 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 cases of the problem considered. Then, the average ranks of the seven algorithms over all 36 cases are calculated to be 1.03, 3.72, 4.96, 3.99, 4.85, 4.61 and 4.85 for SIVID, Full, RS, CAS, ACES, SELSCE and HCSS algorithms, respectively.

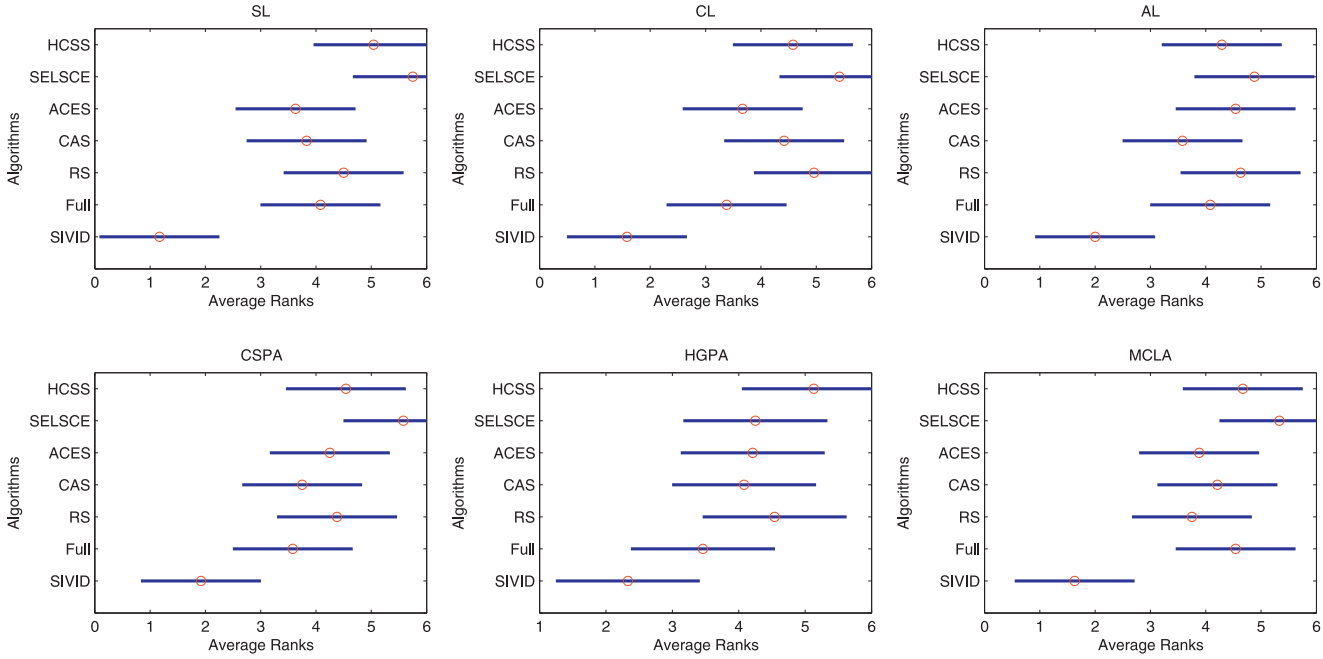

Fig. 4. Nemenyi tests for 50% base clusterings using the k-modes clustering algorithm generating base clusterings in terms of NMI.

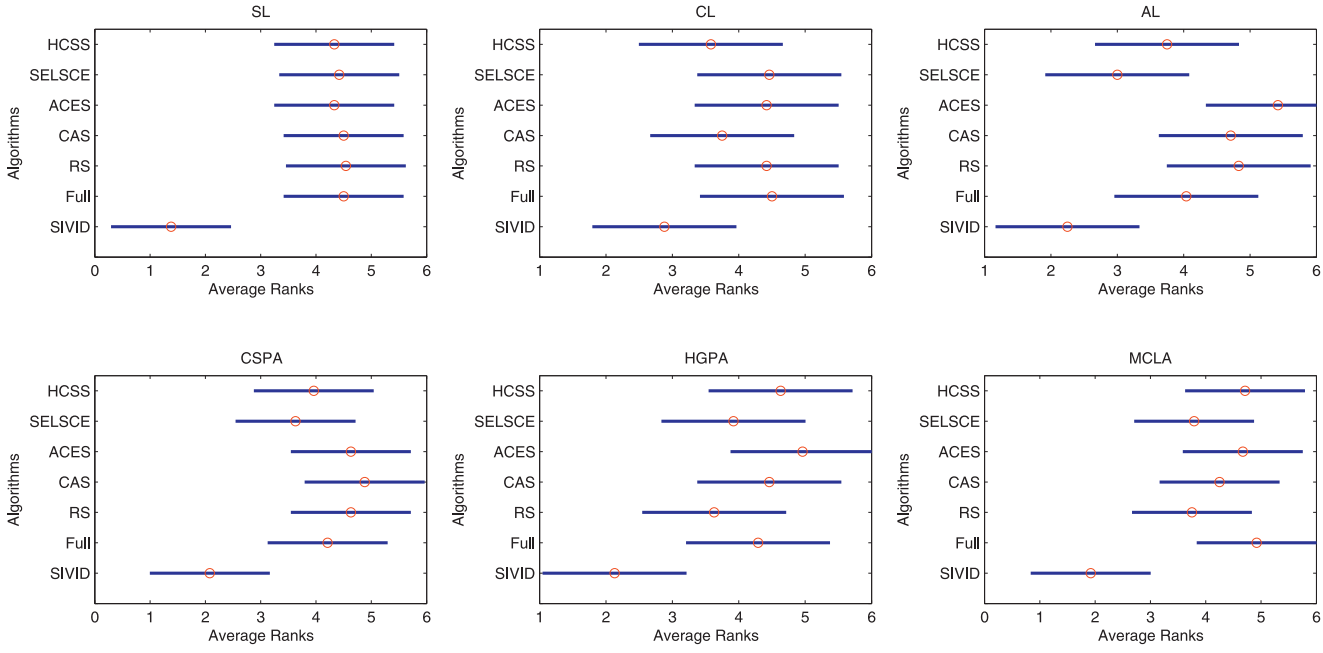

Fig. 5. Nemenyi tests for 50% base clusterings using the simply random partition algorithm generating base clusterings in terms of CA.

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

$$\chi_F^2 = \frac{12B}{A(A+1)} \sum_{j=1}^{A} R_j^2 - 3B(A+1), \tag{15}$$

is distributed according to χ_F^2 with $A-1$ degrees of freedom. According to the Friedman test, a p-value is 2.8291×10^{-17}, which indicates that the null hypothesis can be rejected with high confidence. One can observe that the seven algorithms in comparison 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

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{16}$$

where $k = 7$ is the number of algorithms, and $N = 12$ 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 $CD = 2.1686$.

Figs. 2–7 show the Nemenyi tests for 50% base clusterings with different base clusterings generation algorithms in terms of CA, ARI and NMI, respectively. The average rank of each algorithm is denoted by a red circle, and a blue bar across the circle shows

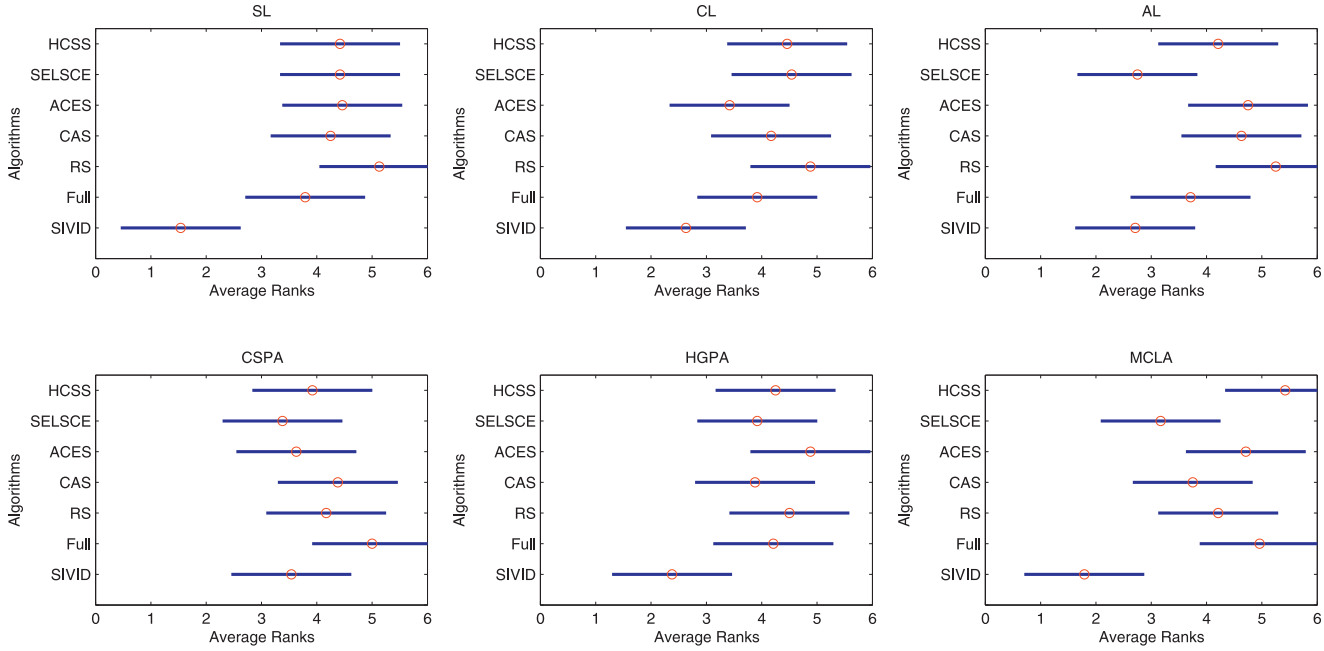

Fig. 6. Nemenyi tests for 50% base clusterings using the simply random partition algorithm generating base clusterings in terms of ARI.

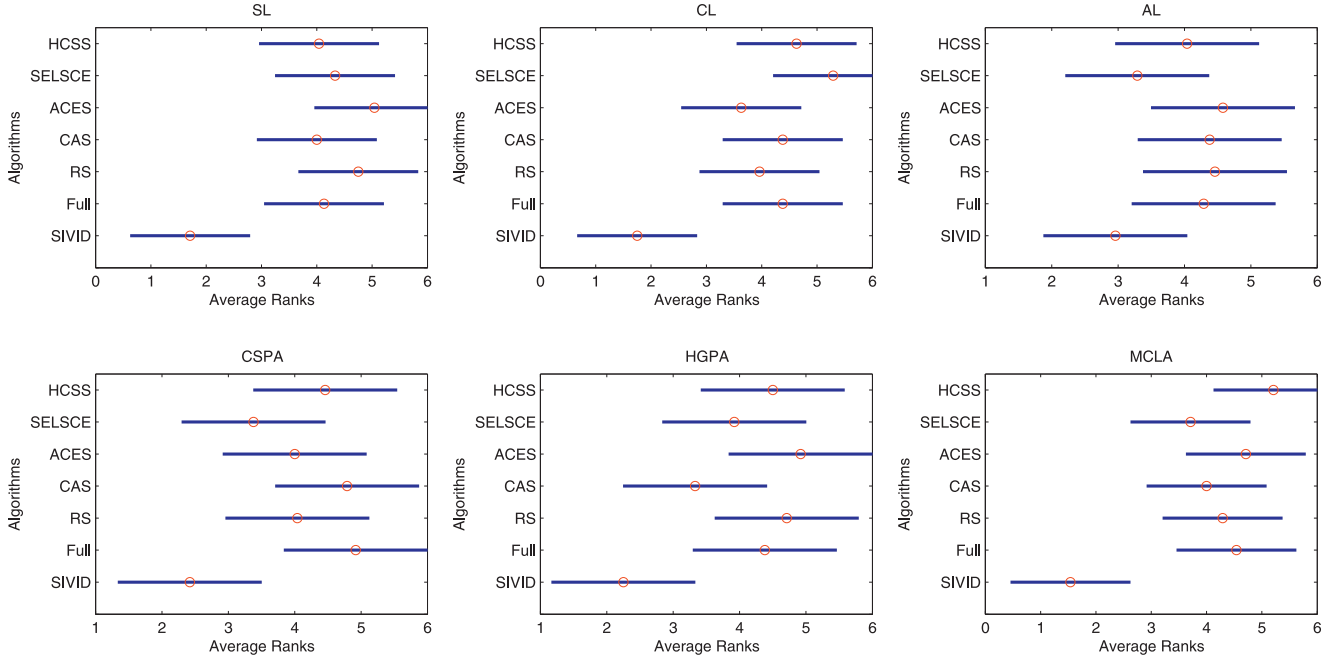

Fig. 7. Nemenyi tests for 50% base clusterings using the simply random partition algorithm generating base clusterings in terms of NMI.

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 Fig. 2, using the *k*-modes clustering algorithm generating base clusterings, the SIVID algorithm has significant difference compared with the other six algorithms in terms of CA. With complete-link (CL) consensus function, there exists an overlap between SIVID and Full in the horizontal direction, which indicates the proposed SIVID algorithm performs as good as or better than using all clustering solutions. Similar experimental results can be seen in terms of ARI and NMI, and the details are shown in Figs. 3–7. Note that in Fig. 6, there exist overlaps among different algorithms with CSPA consensus function, which indicates that the performance of these algorithms is comparable. That is to say, in this case, the proposed SIVID algorithm performs as good as the other algorithms.

4.5. Results on robustness analysis

In this section, we further test the robustness of the proposed algorithm with the ratio of selected base clusterings. For each ratio of selected base clusterings, we run each of the clustering ensemble selection algorithms on every data sets for 20 times with each consensus function and record the average values. In order to make the analysis of the results easier, the averages of validity measures (CA, ARI and NMI) across 12 data sets are depicted as curves in Figs. 8–10. In these figures, the ordinate axis represents the eval-

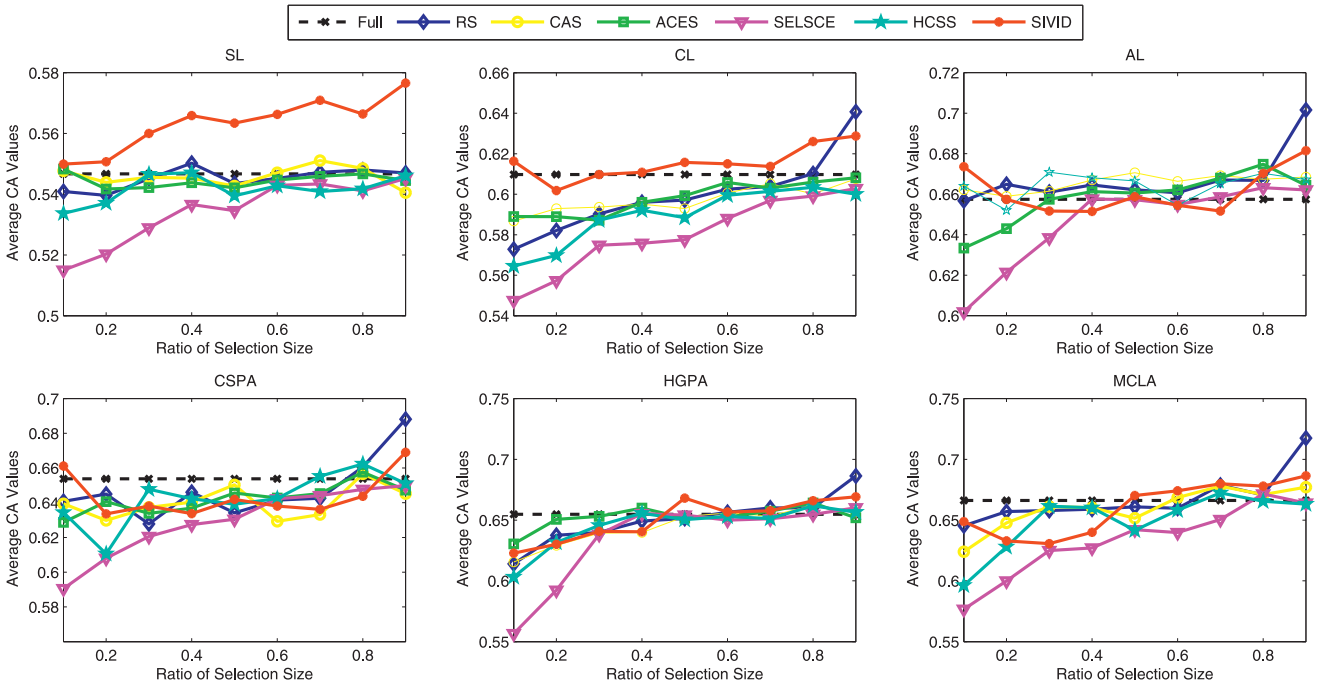

Fig. 8. The average CA values with different consensus functions over the ratio of selected base clusterings.

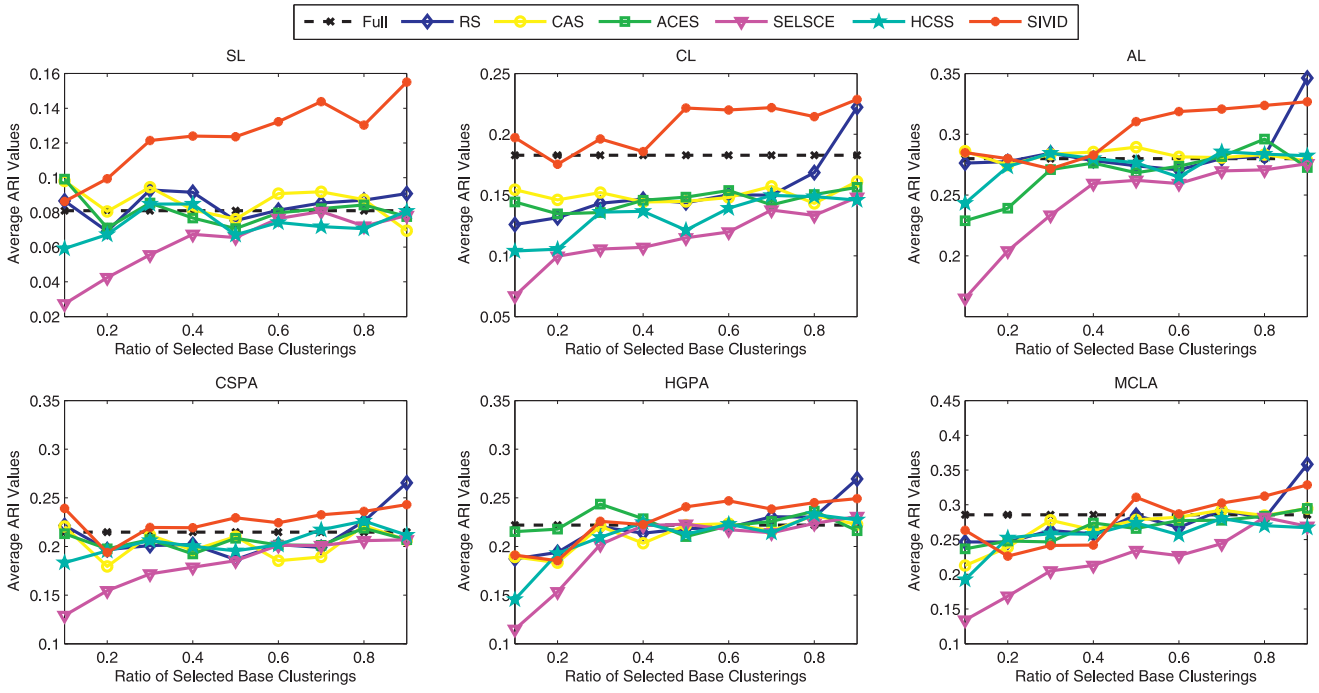

Fig. 9. The average ARI values with different consensus functions over the ratio of selected base clusterings.

uation indices values and the abscissa axis represents the ratio of selected base clustering. In each figure, six sub-fingers present the performance with different consensus functions.

Overall, one important observation is that the performances increase with larger selected base clusterings. However, this relationship is not strictly monotonic. These figures also suggest that the proposed algorithm can obtain consensus partitions with quality higher or equivalent to that of the full ensemble in most cases. Furthermore, the proposed SIVID algorithm performs consistently better than its competitors with all different ratios of selected base clusterings, while the RS algorithm is the least effective. In particular, for the SL, and CL consensus functions, the SIVID algorithm significantly outperforms the other five ensemble selection algorithms. For the other consensus functions, the SIVID algorithm also achieves the best or nearly the best performance among the compared algorithms. On the whole, the proposed SIVID algorithm yields much better and more robust performance than the other clustering ensemble selection algorithms with varying ratios of selected base clusterings on the 12 data sets. Thus, the proposed algorithm is a more robust choice in practical applications.

Beside previous effectiveness and robustness assessments, computational requirements of the proposed algorithm are discussed

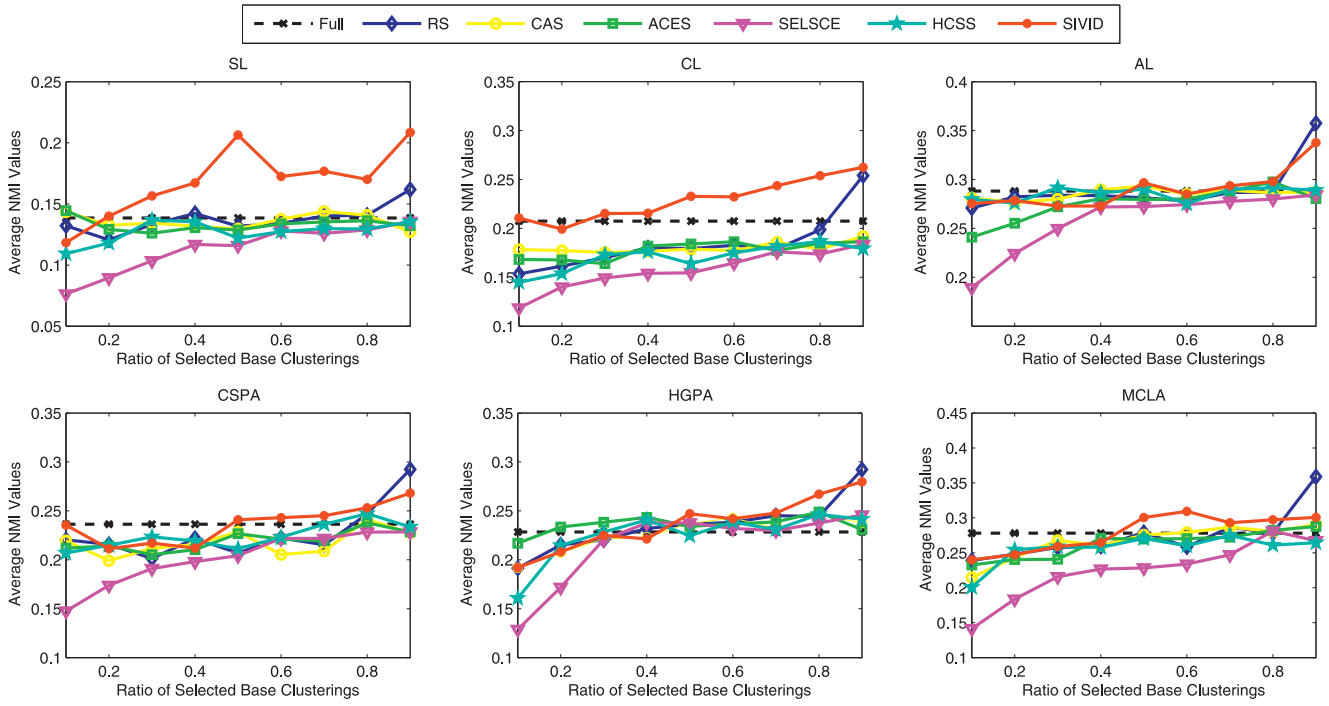

Fig. 10. The average NMI values with different consensus functions over the ratio of selected base clusterings.

here. The proposed SIVID algorithm assess the quality of base partitions via accessing the original data feature space, whereas the other algorithms only use the labels information of base clusterings to measure quality of base clustering. Therefore, the time complexity of the proposed algorithm is relatively higher than its competitors. Fortunately, one advantage of ensemble clustering algorithm is can be implemented in parallel. Thus, this additional computational requirement can be ignored in practical applications. So we did not compare the running time of various algorithms in this section.

5. Conclusion and future work

To improve the performance of the ensemble clustering algorithms for categorical data, a novel selection strategy, namely Sum of Internal Validity Indices with Diversity (SIVID), is presented. In this algorithm, the quality of the base partitions is assessed via five popular internal validity indices. And the diversity measure is based on the normalized mutual information. The effectiveness and robustness of the proposed algorithm is demonstrated on 12 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. As is well known, the performance of the clustering ensemble selection algorithm depends on the quality, diversity of the base clusterings and the number of selected base partitions. Determining the number of selected base partitions automatically is an important problem in this field, which will be part of our future works.

Acknowledgement

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 Fund of China (Nos. 61603230, 61432011, U1435212, 61573229), the National Key Basic Research and Development Program of China (973) (No. 2013CB329404), and the Natural Science Foundation of Shanxi Province, China (No.201601D202039), CityU 11301014 of Hong Kong SAR Government.

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 Networks 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] 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.
[6] Z.X. Huang, M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data, IEEE Trans. Fuzzy Syst. 7 (4) (1999) 446–452.
[7] M.K. Ng, M.J. Li, Z.X. Huang, Z.Y. He, On the impact of dissimilarity measure in k-modes clustering algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 503–507.
[8] L. Bai, J.Y. Liang, C.Y. Dang, F.Y. Cao, The impact of cluster representatives on the convergence of the k-modes type clustering, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1509–1522.
[9] F.Y. Cao, J.Y. Liang, D.Y. Li, X.W. Zhao, A weighting k-modes algorithm for subspace clustering of categorical data, Neurocomputing 108 (2013) 23–30.
[10] L.F. Chen, S.R. Wang, K.J. Wang, J.P. Zhu, Soft subspace clustering of categorical data with probabilistic distance, Pattern Recognit. 51 (2016) 322–332.
[11] Z. He, X. Xu, S. Deng, Squeezer: an efficient algorithm for clustering categorical data, J. Comput. Sci. Technol. 17 (5) (2002) 611–624.
[12] D. Barbara, J. Couto, Y. Li, COOLCAT: an entropy-based algorithm for categorical clustering, in: Proceedings of the 11th International Conference of Information Knowledge Management, USA, 2002, pp. 582–589.
[13] S. Guha, R. Rastogi, K. Shim, ROCK: a robust clustering algorithm for categorical attributes, Inf. Syst. 25 (5) (2000) 345–366.
[14] P. Andritsos, V. Tzerpos, Information-theoretic software clustering, IEEE Trans. Softw. Eng. 31 (2) (2005) 150–165.
[15] T.K. Xiong, S.R. Wang, A. Mayers, E. Monga, DHCC: divisive hierarchical clustering of categorical data, Data Min. Knowl. Discov. 24 (1) (2012) 103–135.
[16] D. Gibson, J. Kleinberg, P. Raghavan, Clustering categorical data: an approach based on dynamical systems, VLDB J. 8 (3) (2000) 222–236.
[17] M.J. Zaki, M. Peters, I. Assent, T. Seidl, Clicks: an effective algorithm for mining subspace clusters in categorical datasets, in: Proceedings of the 11th ACM SIGKDD Conference of Knowledge Discovery Data Mining, Chicago, USA, 2005, pp. 736–742.
[18] J. Ghosh, A. Acharya, Cluster ensembles, WIREs Data Min. Knowl. Discov. 1 (2011) 305–315.

[19] H.G. Ayad, M.S. Kamel, On voting-based consensus of cluster ensembles, Pattern Recognit. 43 (2010) 1943–1953.
[20] D. Huang, J.H. Lai, C.D. Wang, Ensemble clustering using factor graph, Pattern Recognit. 50 (2016) 131–142.
[21] A. Topchy, B. Minaei-Bidgoli, A.K. Jain, W.F. Punch, Adaptive clustering ensembles, in: Proceedings of the 17th International Conference on Pattern Recognition, volume 1, 2004, pp. 272–275.
[22] X.Z. Fern, W. Lin, Cluster ensemble selection, Stat. Anal. Data Min. 1 (3) (2008) 128–141.
[23] J. Azimi, X. Fern, Adaptive cluster ensemble selection, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009, pp. 992–997.
[24] Y. Hong, S. Kwong, H. Wang, Q. Ren, Resampling-based selective clustering ensembles, Pattern Recognit. Lett. 30 (3) (2009) 298–305.
[25] Z.W. Yu, L. Li, Y.J. Gao, J. You, J.M. Liu, H.S. Wong, G.Q. Han, Hybrid clustering solution selection strategy, Pattern Recognit. 47 (10) (2014) 3362–3375.
[26] P.A. Jaskowiak, D. Moulavi, A.C.S. Furtado, R.J.G.B. Campello, A. Zimek, J. Sander, On strategies for building effective ensembles of relative clustering validity criteria, Knowl. Inf. Syst. 47 (2) (2016) 329–354.
[27] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, 1967, pp. 281–297.
[28] J. Jia, X. Xiao, B. Liu, L. Jiao, Bagging-based spectral clustering ensemble selection, Pattern Recognit. Lett. 32 (10) (2011) 1456–1467.
[29] H. Alizadeh, B.M. Bidgoli, H. Parvin, To improve the quality of cluster ensembles by selecting a subset of base clusters, J. Exp. Theor. Artif. Intell. 26 (1) (2014) 127–150.
[30] H. Alizadeh, B.M. Bidgoli, H. Parvin, Cluster ensemble selection based on a new cluster stability measure, Intell. Data Anal. 18 (3) (2014) 389–408.
[31] C. Domeniconi, M. AlRazgan, Weighted cluster ensembles: methods and analysis, ACM Trans. Knowl. Discov. Data 2 (4) (2009) 1–42.
[32] S. Vega-Pons, J. Correa-Morris, J. Ruiz-Shulcloper, Weighted partition consensus via kernels, Pattern Recognit. 43 (8) (2010) 2712–2724.
[33] S. Vega-Pons, J. Ruiz-Shulcloper, A. Guerra-Gandn, Weighted association based methods for the combination of heterogeneous partitions, Pattern Recognit. Lett. 32 (16) (2011) 2163–2170.
[34] M.A. Gluck, J.E. Corter, Information, uncertainty, and the utility of categories, in: Proceeding of the 7th Annual Conference of the Cognitive Science Society, Irvine, CA: Lawrence Erlbaum Associates, 1985, pp. 283–287.
[35] D.H. Fisher, Knowledge acquisition via incremental conceptual clustering, Mach. Learn. 2 (2) (1987) 139–172.
[36] K. McKusick, K. Thompson, 1990, COBWEB/3: A portable implementation, Technical Report FIA-90-6-18-2, NASA Ames Research Center.
[37] L. Bai, J.Y. Liang, Cluster validity functions for categorical data: a solution-space perspective, Data Min. Knowl. Discov. 29 (6) (2015) 1560–1597.
[38] A. Topchy, A.K. Jain, W. Punch, Clustering ensembles: models of consensus and weak partitions, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1866–1881.
[39] 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.
[40] C. Chang, Z. Ding, Categorical data visualization and clustering using subjective factors, Data Knowl. Eng. 53 (3) (2005) 243–263.
[41] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.
[42] C. Gao, W. Pedrycz, D.Q. Miao, Rough subspace-based clustering ensemble for categorical data, Soft Comput. 17 (9) (2013) 1643–1658.
[43] A. Strehl, J. Ghosh, Cluster ensembles: a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2002) 583–617.
[44] 2012, UCI Machine Learning Repository, http://www.ics.uci.edu/mlearn/MLRepository.html.
[45] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.
[46] A.L.N. Fred, A.K. Jain, Data clustering using evidence accumulation, in: Proceedings of 16th international Conference on Pattern Recognition, volume 4, 2002, pp. 276–280.
[47] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 3 (2006) 1–30.

Xingwang Zhao is a Ph.D. candidate in the School of Computer and Information Technology in Shanxi University. He received his M.S. degree from Shanxi University in 2011. His research interests are in the areas of data mining and machine learning.

Jiye Liang received the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1990 and 2001, respectively. He is currently a professor of School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education at Shanxi University. His current research interests include computational intelligence, granular computing, data mining and knowledge discovery. He has published more than 180 journal paper in his research fields.

Chuangyin Dang received a Ph.D. degree in operations research/economics from the University of Tilburg, The Netherlands, in 1991, a M.S. degree in applied mathematics from Xidian University, China, in 1986, and a B.S. degree in computational mathematics from Shanxi University, China, in 1983. He is professor at the City University of Hong Kong. He is best known for the development of the D1-triangulation of the Euclidean space and the simplicial method for integer programming. His current research interests include computational intelligence, optimization theory and techniques, applied general equilibrium modeling and computation. He is a senior member of IEEE and a member of INFORS and MPS.