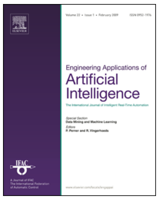
Engineering Applications of Artificial Intelligence 39 (2015) 146–156

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Hierarchical cluster ensemble selection

Ebrahim Akbari [a,b,*], Halina Mohamed Dahlan [a], Roliana Ibrahim [a], Hosein Alizadeh [c]

[a] *Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia*
[b] *Department of Computer Engineering, Islamic Azad University, Sari Branch, Sari, Iran*
[c] *Computer Engineering Department, Iran University of Science and Technology, 1684613114 Narmak, Tehran, Iran*

ARTICLE INFO

Article history:
Received 5 August 2014
Received in revised form
12 November 2014
Accepted 9 December 2014
Available online 31 December 2014

Keywords:
Cluster ensemble selection
Hierarchical clustering
Diversity
Consensus function

ABSTRACT

Clustering ensemble performance is affected by two main factors: diversity and quality. Selection of a subset of available ensemble members based on diversity and quality often leads to a more accurate ensemble solution. However, there is not a certain relationship between diversity and quality in selection of subset of ensemble members. This paper proposes the Hierarchical Cluster Ensemble Selection (HCES) method and diversity measure to explore how diversity and quality affect final results. The HCES uses single-link, average-link, and complete link agglomerative clustering methods for the selection of ensemble members hierarchically. A pair-wise diversity measure from the recent literature and the proposed diversity measure are applied to these agglomerative clustering algorithms. Using the proposed diversity measure in HCES leads to more diverse ensemble members than that of pairwise diversity measure. Cluster-based Similarity Partition Algorithm (CSPA) and Hypergraph-Partitioning Algorithm (HGPA) were employed in HCES method for obtaining the full ensemble and cluster ensemble selection solution. To evaluate the performance of the HCES method, several experiments were conducted on several real data sets and the obtained results were compared to those of full ensembles. The results showed that the HCES method led to a more significant performance improvement compared with full ensembles.

1. Introduction

Clustering is one of the unsupervised rules for searching and analyzing data, which is used in different fields such as statistics, pattern recognition, machine learning, data mining, and bio-informatics (Jain, 2010; Quintana et al., 2003; De Angelis and Dias, 2014; Sun et al., 2012). Wide usage of clustering algorithms proves their usefulness in exploratory data analysis (Jain et al., 2000). The major aim of data clustering is to find groups of patterns (clusters) in such a way that patterns in one cluster can be more similar to each other than to patterns of other clusters. Because of characteristics of dataset, different clustering algorithms obtain different clustering results. It is difficult to choose a suitable algorithm for a given data set. Based on the Kleinberg theorem, there is no the best single clustering algorithm (Kleinberg, 2003).

Clustering ensemble, which is an approach in clustering problem, combines multiple clustering results (clusterings) to achieve final clusters without accessing the features or algorithms that obtain the clusterings. The combination of the clusterings is performed by a consensus algorithm. The clustering ensemble approach attempts to improve the quality and robustness of clustering results (Strehl and Ghosh, 2003; Fred and Jain, 2005; Mimaroglu and Erdil, 2013). Furthermore, clustering ensemble can achieve some properties such as novelty, stability, and scalability (Topchy et al., 2005). There are some applications of clustering ensemble in bio-informatics, image processing, and marketing (Strehl and Ghosh, 2003; Avogadri and Valentini, 2009; Ma et al., 2009; Mimaroglu and Erdil, 2010). Since clustering ensemble only needs to gain access to the base clusterings instead of the data itself, it provides a convenient approach to privacy preservation and knowledge reuse (Strehl and Ghosh, 2003). In many applications, for the objects under consideration, various clusterings may already come to exist. In this condition, these clusterings can be integrated into a single solution. For example, in market basket analysis, assume that a company already has various legacy customer segmentations based on geographical region, credit rating, demographics, and purchasing patterns in their retail stores, and so on. They want to reuse this pre-existing knowledge in order to form a single consolidated clustering. Because the legacy clusterings are provided largely by experts or by other companies by means of proprietary methods, for reusing this knowledge, there is a limited access to original features of raw data and the algorithms that obtain the clusterings (Strehl and Ghosh, 2003).

[*] Corresponding author at: Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia. Tel.: +60 177263810.
E-mail addresses: akbari@iausari.ac.ir, ebrahimakbari30@yahoo.com (E. Akbari).

http://dx.doi.org/10.1016/j.engappai.2014.12.005
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

On the other hand, in contrast to the knowledge reuse, there could be a potential for greater gains when using an ensemble for the purpose of improving clustering quality (Strehl and Ghosh, 2003). Traditionally, a set of large library of clusterings is generated and then the consensus solution is obtained by a consensus function based on all base clusterings. Unlike classification problems where labels of data items are known beforehand, data items in unsupervised clustering problems are unlabeled which may some clustering results unreliability in large library of clusterings. Thus, not all obtained clusterings can truly benefit for the final solution of clustering ensembles (Azimi and Fern, 2009; Hong et al., 2009).

Recently, a subset of diversity is selected rather than all for combining the diversity to obtain the final result (Hadjitodorov et al., 2006). Cluster ensemble selection is mainly aimed to select a subset from a large library of clustering solutions to form a smaller cluster ensemble that performs as properly as or better than the set of all available clustering solutions (Kuncheva and Hadjitodorov, 2004; Fern and Lin, 2008; Azimi and Fern, 2009). Selective ensembles method is also on the basis of the supervised classification area in which it has been recognized that selective classifier ensembles always outperform the conventional ensemble methods in terms of achieving better solutions (Banfield et al., 2005; Zhang et al., 2006). In a straightforward classifiers selection method, the classifiers are ranked based on their individual performance on a held-out test set and the best ones are picked (Caruana et al., 2014). Whereas, in unsupervised clustering area, data items are unlabeled beforehand. As a result, this is not possible to estimate the quality of a single clustering result by computing its quality on the test set.

In ensemble selection, diversity and quality are two important factors that affect ensemble performance (Fern and Brodley, 2003). A few recent studies have investigated heuristically the question how a subset of ensemble members should be selected based on diversity and quality (Minaei-Bidgoli et al., 2014; Alizadeh et al., 2014; Naldi et al., 2013). The most successful method proposed by Fern and Lin (2008) is called the Cluster And Select (CAS) that combines quality and diversity. This, first, partitions the ensemble members into k (the number of clusters) clusters based on their similarities. Then, CAS selects the clusterings with the highest quality from each obtained cluster for the ensemble. They concluded that the use of both quality and diversity in cluster ensemble selection (CES) can make a higher improvement in the results compared to full ensembles. However, the drawback of CAS is that the number of k that can obtain the most appropriate ensemble size is uncertain and the concept of quality and diversity is loosely defined. To address the above problems, a hierarchical diversity selection strategy based on both diversity and quality is proposed to improve the traditional clustering ensemble performance. This strategy also solves the drawback of ensemble selection strategy in the CAS method.

This paper proposes a new combinational method called the Hierarchical Cluster Ensemble Selection (HCES). In the first step of the HCES method, a pairwise matrix of all available clustering members is constructed using two different diversity measures. It then employs three hierarchical methods including single-link, average-link, and complete-link to build a nested tree. An appropriate cut on the obtained tree creates diverse groups of primary partitions that guide us in targeted selection of smaller yet better performing ensemble. Finally, the HCES uses HGPA and CSPA consensus clustering algorithms for obtaining consensus clustering solutions. The HCES method obtains the final solution through the selection of an appropriate layer of the hierarchy. In the HCES method, there is no need to specify the value of k. The HCES empirically is compared to the full ensemble. The evaluation results obtained from different real data sets demonstrate statistically more significant performance improvement compared to the full ensembles. In addition, because of interpretability of the proposed method, the results are improved even with removing only one clustering; the removed clustering is considered as a noise. As a brief, the contributions of the present paper are as follow:

1. Proposing an automatic hierarchical cluster ensemble selection.
2. Proposing a diversity measure and applying to the proposed HCES method.
3. Applying three agglomerative clustering algorithms to the method and showing their effect on the performance of the method.

The rest of the paper is organized as follows. Section 2 gives an overview of related work. Section 3 introduces different diversity and quality measures. Section 4 presents the hierarchical ensemble selection method. Section 5 presents the experiments carried out on several real data sets and the obtained results. Finally, Section 6 concludes the paper and recommends future work.

2. Related work

Clustering ensemble is an approach that is widely adopted in clustering research to improve the quality and robustness of clustering results. Clustering ensemble includes two main parts: diversity (creating multiple clusterings) and consensus function (combining multiple clusterings). Recently, researchers have suggested the selection of diversity to improve the ensemble performance (Hadjitodorov et al., 2006; Jia et al., 2011; Hong et al., 2009). Fig. 1 shows the steps of clustering ensemble selection approach. In this section, some clustering ensemble methods and recent studies conducted on cluster ensemble design are reviewed.

2.1. Diversity generation

In ensemble classifier/clustering techniques, generating diversity is commonly used in supervised and unsupervised combining approaches. Various methods have been proposed in the literature for creating diversity or ensemble members, including

1. *Different parameter initializations*: Primary clusterings are created using repeated runs of a single clustering algorithm with several sets of parameter initializations such as cluster centers of the k-means clustering technique, which are known as homogeneous ensembles (Fred and Jain, 2005).
2. *Different clustering algorithms*: A number of different clustering algorithms are used together to generate primary clusterings,

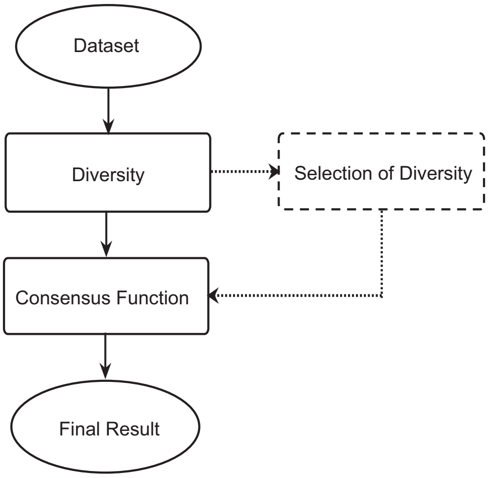


Fig. 1. Steps of the clustering ensemble selection approach.

which are called heterogeneous ensembles (Topchy et al., 2005; Berikov, 2013).

3. *Different subsets of features*: Features are selected or extracted to create subsets that are used for the generation of clusterings (Fred and Jain, 2005; Topchy et al., 2003; Hong et al., 2008).
4. *Different subsets of objects*: Data are re-sampled with or without replacement for generating clusterings (Minaei-Bidgoli et al., 2004; Yu et al., 2012).
5. *Projection to subspace*: The objects are projected on different subspaces, which include the projection to one dimension and random cut that are applied to the production of clusterings (Topchy et al., 2003; Fern and Brodley, 2003).

Typically, the first and second methods are suitable for low-dimensional data, the third and fifth methods are suitable for high-dimensional data, and the fourth method is suitable for large dataset.

2.2. Selection of diversity

Recently, cluster ensemble selection (CES) techniques have been proposed to improve the ensemble performance (Hadjitodorov et al., 2006; Fern and Lin, 2008; Azimi and Fern, 2009; Wang et al., 2013). These techniques select a subset of ensemble members based on both diversity and quality that are two important factors for improvement of the ensemble solution (Fern and Brodley, 2003; Hadjitodorov et al., 2006; Fern and Lin, 2008; Azimi and Fern, 2009). If the generated ensemble members (clusterings) are different from each other and they also have an acceptable quality, a better ensemble solution can be achieved (Yang et al., 2014).

In the literature, there are different quality and diversity measures that are considered for ensemble members (Hadjitodorov et al., 2006; Lu et al., 2013; Naldi et al., 2013; Alizadeh et al., 2014). Most of them are based on match index between two partitions. Two diversity measures commonly used in the literature are Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and Normalized Mutual Information (NMI) (Strehl and Ghosh, 2003). These measures also are used for measuring quality between two partitions. Hadjitodorov et al. (2006) used ARI diversity measure on a large number of cluster ensembles as candidate ensembles for selection. They constructed four diversity measures based on ARI and found the median of the diversity values for ensemble members and picked the corresponding ensemble. Lu et al. (2013) introduced a diversity measure based on covariance. Alizadeh et al. (2014) proposed a cluster ensemble selection method in which clusters (instead of clusterings) are selected based on quality and diversity measures.

Naldi et al. (2013) proposed several relative cluster validity indices based on quality and diversity for selection of clusterings. Using different relative diversity measures, they also investigated the impact of the diversity on partitions (clusterings) used for the ensemble. Azimi and Fern (2009) proposed the adaptive cluster ensemble selection method in which data sets were divided to *stable* and *non-stable* based on NMI values. They demonstrated that, for *non-stable* data sets, the selection of clusterings with more diversity made an improvement in the solution. Jia et al. (2011) generalized the selective clustering ensemble algorithm proposed by Azimi and Fern (2009) and they proposed a novel clustering ensemble method, namely SELective Spectral Clustering Ensemble (SELSCE). Ensemble members were generated by spectral clustering (SC) that was able to engender diverse committees. The random scaling parameter, Nyström approximation, and random initialization were used for producing the components of the ensemble system. After the generation of component clusterings, the bagging technique was used to rank and assess the component clustering. Based on this ranking, ensemble members were selected for ensemble. Fern and Lin (2008) investigated a variety of heuristic methods for selecting subsets, which considered both the diversity and the quality of the ensemble members. Among these methods, CAS was empirically demonstrated to achieve the most robust performance. This method first partitions all ensemble members into k clusters and then selected one solution from each cluster to form the final ensemble. However, the k is not a certain value in the CAS method.

2.3. Consensus function

Consensus function is an algorithm for combining different clusterings (ensemble members) to obtain final clusters (Strehl and Ghosh, 2003; Fred and Jain, 2005; Mimaroglu and Erdil, 2013). Assume that H has L ensemble members, $H=\{h_1,h_2,\ldots,h_L\}$, the consensus function Φ combines all ensemble members of H as $h^*=\Phi(h_1,h_2,\ldots,h_L)$. In the cluster ensemble selection, the consensus function affects a subset of ensemble members instead of all. The consensus function for cluster ensemble selection is defined as $h_s^*=\Phi(H_s)$, where $H_s\subset H$. The literature contains several approaches that can be divided into voting, feature-based, pairwise, and graph-based approaches.

The voting approach is also referred to as direct approach or re-labeling approach. Contrary to other approaches in which it is not necessary to solve the correspondence problem between the labels of known and achieved clusters, the voting approach solves the correspondence problem. A re-labeling can be done optimally between two clusterings using the Hungarian algorithm (Kuhn, 1955). After an optimal re-labeling, a simple voting can be used to assign objects to clusters, with which final consensus partitions are identified.

In the feature-based approach, output of each clustering algorithm is considered as a categorical feature. In this approach, L features can be considered as an intermediate feature space on which other clustering algorithms can work. Topchy et al. (2004) have proposed a function called the generalized mutual information. Considering the fact that the objective function equals the total intra-cluster variance of the partition in the transformed space of labels, the k-means algorithm in such space can provide corresponding consensus solutions.

The pairwise approach constructs the co-association matrix in which the similarity between points is the number of times that points are in the same created clusters of clusterings. Usually, hierarchical algorithms such as single-link, average-link, and complete-link are used for combining results by co-association matrix (Fred and Jain, 2005).

The graph-based approach includes instance-based, cluster-based, and hybrid approaches. In instance-based approach, the objects are considered as vertices and a similarity measure between the objects (vertices) in clusters are calculated as weight of the edges. The cluster-based similarity partitioning algorithm (CSPA) as an instance-based approach constructs a hypergraph in which the number of frequencies of two objects which are accrued in the same clusters is considered as weight of each edge. The k partitions are obtained using the METIS (Karypis and Kumar, 1998) on the induced similarity graph (Strehl and Ghosh, 2003). On the other hand, cluster-based approach constructs a meta-graph in which clusters are considered as vertices, and the similarity measure between clusters (vertices) is calculated as weight of the edges. The meta-clustering algorithm (MCLA) is a famous cluster-based method in which the Jaccard measure is applied as a similarity measure between two corresponding clusters (Strehl and Ghosh, 2003). In the hybrid approach, both objects and clusters are considered as vertices, and the similarity measures are calculated simultaneously based on objects and clusters located between two vertices (Fern and Brodley, 2004).

3. Diversity and quality measures

Two partitions are diverse if one partition's labels are not matched properly with the labels of the other one. The normalized mutual information (NMI) (Strehl and Ghosh, 2003) and adjusted rand index (ARI) (Hubert and Arabie, 1985) are commonly employed to measure the diversity or quality of partition(s). The ARI and NMI quality measures are calculated, respectively, as follows:

$$\mathrm{ARI}(h_a, h_b) = \frac{\sum_{i=1}^{k_a}\sum_{j=1}^{k_b}\binom{n_{ij}}{2} - t_3}{1/2(t_1+t_2)-t_3} \tag{1}$$

where

$$t_1 = \sum_{i=1}^{k_a}\binom{n_{ia}}{2}, \quad t_2 = \sum_{j=1}^{k_b}\binom{n_{bj}}{2}, \quad \text{and} \quad t_3 = \frac{2t_1t_2}{n(n-1)}.$$

and

$$\mathrm{NMI}(h_a, h_b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}\log\left(\frac{n.n_{ij}}{n_{ia}.n_{bj}}\right)}{\sum_{i=1}^{k_a} n_{ia}\log\left(\frac{n_{ia}}{n}\right) + \sum_{j=1}^{k_b} n_{bj}\log\left(\frac{n_{bj}}{n}\right)} \tag{2}$$

where, in both equations, $h_a = \{c_1^a, c_2^a, \ldots, c_{k_a}^a\}$ and $h_b = \{c_1^b, c_2^b, \ldots, c_{k_b}^b\}$ with k_a and k_b clusters, respectively, are two clusterings on dataset D with n samples; n_{ij} signifies the number of common objects in cluster c_i in clustering h_a and in cluster c_j in clustering h_b; n_{ia} denotes the number of objects in cluster c_i in clustering h_a; and n_{bj} stands for the number of objects in cluster c_j in clustering h_b.

Diversity measures can be divided into external and internal diversities. When the class labels are available, the external diversity measure is defined based on a quality measure such as NMI or ARI, as follows:

$$\mathrm{diversity}(\overline{h}, h_i) = 1 - \mathrm{quality}(\overline{h}, h_i) \tag{3}$$

where $\overline{h}$ is the known class label and $h_i, i = 1, 2, \ldots, L$ are clusterings. Note that here NMI is used as a quality measure. The average of diversity is

$$D_e = \frac{1}{L}\sum_{i=1}^{L}\mathrm{diversity}(\overline{h}, h_i).$$

Internal diversity can be divided into pair-wise and non-pair-wise diversities. In pair-wise diversity, each clustering is chosen as a class label implicitly, and other clusterings are measured by the chosen class label. The diversity is calculated as follows:

$$\mathrm{diversity}(h_i, hj) = 1 - \mathrm{quality}(h_i, h_j) \tag{4}$$

where $i \neq j = 1, 2, \ldots, L$. The average of diversity measure is

$$D_p = \frac{1}{L(L-1)}\sum_{i=1}^{L}\sum_{j=1, i\neq j}^{L}\mathrm{diversity}(h_i, h_j).$$

The non pair-wise diversity measure is defined as follows:

$$\mathrm{diversity}(h^*, h_i) = 1 - \mathrm{quality}(h^*, h_i) \tag{5}$$

where $i = 1, 2, \ldots, L$ and h^* is a result obtained by a consensus function. The average of diversity is

$$D_{np} = \frac{1}{L}\sum_{i=1}^{L}\mathrm{diversity}(h^*, h_i).$$

Kuncheva and Hadjitodorov (2004) showed that ensembles with larger spread of individual diversities are generally better than ensembles with a smaller spread. Therefore, based on the h^* obtained by a consensus function and h_i, $i = 1, 2, \ldots, L$ that are ensemble members, in this paper, a new relative diversity measure is proposed as follows:

$$\mathrm{diversity}(h_i, hj) = |\mathrm{quality}(h^*, h_i) - \mathrm{quality}(h^*, h_j)| \tag{6}$$

The relative diversity measure calculates the absolute distance between qualities of h_i and h_j in comparison with the reference consensus partition, h^*. The HCES obtains subsets of clusterings with more diversity using the relative diversity measure compared to that of Eq. (4) (Section 5.2). In our experiment, two diversity measures of Eqs. (4) and (6) were used, and their effects on the performance of the HCES method was compared with that of full ensemble.

4. Cluster ensemble extraction approach

The process of hierarchical clustering methods can be displayed as a dendogram that includes nested partitions (clusterings) of a dataset. In fact, dendogram is a particular type of tree that provides a comprehensive picture of the hierarchical clustering. Each dendogram includes several layers of nodes each of which represents a cluster. One clustering is obtained by cutting the dendogram at the proper layer. The hierarchical clustering methods are grouped into two categories: agglomerative and divisive. In the former, each data is considered as a cluster and, in a bottom-up movement, each pair of clusters are merged recursively with minimum defined distance value until all clusters are merged as one partition (clustering). In the latter, all data is considered as one partition (clustering) and, in a top-down movement, two clusters with maximum distance value are disjoined recursively until each data is constructed as a cluster. Some popular agglomerative clustering methods are single-link, average-link, and complete-link. Some advantages of hierarchical clustering methods are their interpretability and the fact that there is no need to specify the number of clusters (Murtagh, 1983).

In the proposed HCES method, each clustering solution is considered as an entity (node in the dendogram). Using a pair-wise diversity measure on each couple of ensemble members, the pair-wise diversity matrix is constructed. In our experiment, two diversity measures of Eqs. (4) and (6) are used in the HCES method. The clustering solutions (ensemble members), h_i, $i = 1, 2, \ldots, L$, are partitioned by a hierarchical clustering algorithm using the constructed diversity matrix. Here, three agglomerative hierarchical clustering methods: single-link, average-link, and complete-link are used. In each layer of dendogram, two clusterings with minimum diversity (i.e., high quality) are merged. One solution then simply is selected from each group to form an ensemble. The solution is selected based on the highest quality between clusterings of the group and full ensemble solution (h^*). This quality is calculated by the formula of $\mathrm{quality}(h^*, h_i) = \mathrm{NMI}(h^*, h_i)$, where h_i is a member of the group. Fig. 2 shows basic strategy of diversity selection used in the HCES method.

The general framework of HCES method based on the basic strategy is shown in Fig. 3.

The main procedure of the HCES method includes four steps. In the first step, clusterings are partitioned into k groups. Each group is included some homologous clusterings. In the second step, one solution is selected with the highest quality based on a quality measure such as NMI quality measure ($h_i^j \in c_i^j$, $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, L$). In the third step, consensus solution is obtained using a consensus function in each layer ($h_j^* = \Phi(h_1^j, h_2^j, \ldots, h_k^j)$, $j = 1, 2, \ldots, L-1$). At last, final solution is one consensus solution with the highest quality value among the consensus solutions ($h^* = \max_{quality}\{h_i^*,\ i = 2, 3, \ldots,\ L-1\}$). In the first layer, the number of partitions of the clusterings is equal to the number of clusterings (Level 1 in Fig. 3). In this case, ensemble solution (h_1^*) is equivalent to the result of the full ensemble. In the last layer, all clusterings

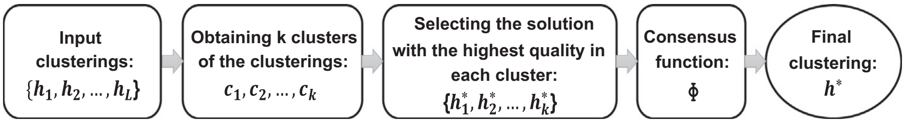

Input clusterings: $\{h_1, h_2, \ldots, h_L\}$
Obtaining k clusters of the clusterings: $c_1, c_2, \ldots, c_k$
Selecting the solution with the highest quality in each cluster: $\{h_1^*, h_2^*, \ldots, h_k^*\}$
Consensus function: Φ
Final clustering: h^*

Fig. 2. The basic strategy of diversity selection.

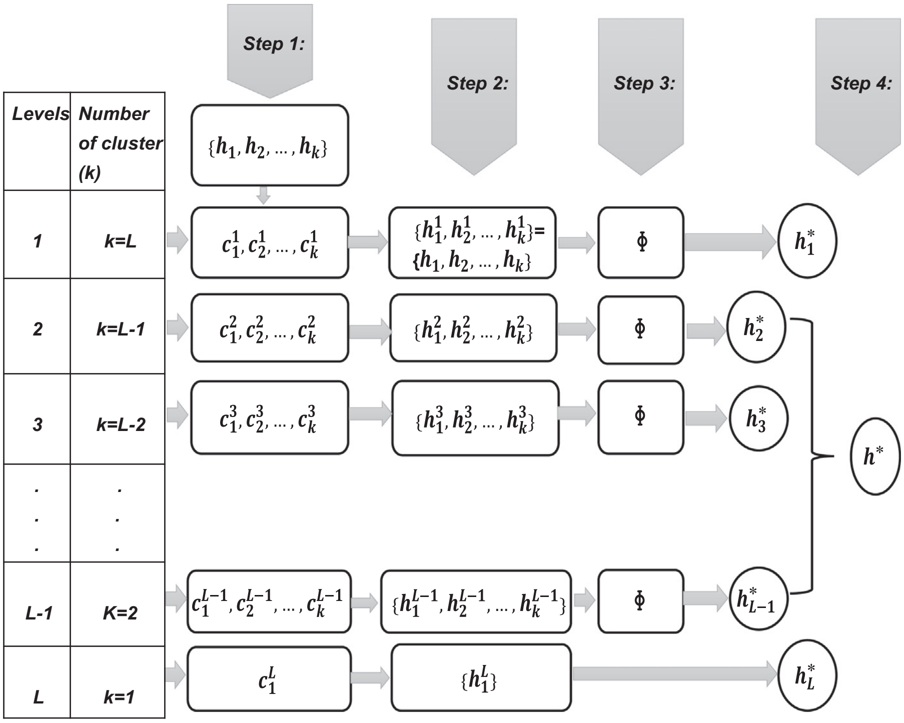

Step 1:
Step 2:
Step 3:
Step 4:
Levels
Number of cluster (k)
$\{h_1, h_2, \ldots, h_k\}$
1
k=L
$c_1^1, c_2^1, \ldots, c_k^1$
$\{h_1^1, h_2^1, \ldots, h_k^1\}= \{h_1, h_2, \ldots, h_k\}$
Φ
h_1^*
2
k=L-1
$c_1^2, c_2^2, \ldots, c_k^2$
$\{h_1^2, h_2^2, \ldots, h_k^2\}$
Φ
h_2^*
3
k=L-2
$c_1^3, c_2^3, \ldots, c_k^3$
$\{h_1^3, h_2^3, \ldots, h_k^3\}$
Φ
h_3^*
h^*
L-1
K=2
$c_1^{L-1}, c_2^{L-1}, \ldots, c_k^{L-1}$
$\{h_1^{L-1}, h_2^{L-1}, \ldots, h_k^{L-1}\}$
Φ
h_{L-1}^*
L
k=1
c_1^L
$\{h_1^L\}$
h_L^*

Fig. 3. HCES framework.

are included in one group (Level L in Fig. 3). In this condition, the final result is equal to choose the solution with the highest quality (h_L^*). Apart from layer 1 and L, different subsets of clusterings are chosen based on diversity/quality in HCES (Levels 2 until $L-1$ in Fig. 3). In each subset, the ensemble solution is obtained by a consensus algorithm. Finally, the best subset is obtained with the highest quality NMI value.

Given a dataset, the framework of the HCES method can be explained based on Figs. 2 and 3 as follows:

1. Generating different clusterings.
2. Obtaining consensus clustering solution h^* by applying a consensus function.
3. Computing pair-wise diversity measure matrix in which each element of matrix is diversity measure between two clusterings.
4. Using a hierarchical clustering algorithm on the diversity measure matrix, all clusterings are partitioned as a dendogram implicitly.
5. Choosing the one solution from each group with highest quality (in each layer of dendogram) using NMI quality measure for finding a new subset of clusterings.
6. Obtaining an ensemble solution by a consensus function on the new subset.
7. Choosing the best ensemble solution among ensemble results based on their quality.

5. Evaluating validity of the HCES solutions

The experiments were conducted with real data sets, where true natural clusters were known. Because our data sets were labeled, the quality of the clustering solutions could be assessed using external criteria (Strehl and Ghosh, 2003). The external criteria were used to measure the discrepancy between the structure defined by a clustering and the one defined by the class labels. In this paper, the NMI measure was used to evaluate the final results obtained by the HCES method. Remind that, if the NMI value is zero, two partitions are completely different. While, they are identical if the value is one. The CAS method is a special case of our method, which is the basic strategy in HCES (Fig. 2). Moreover, Section 5.2 shows that HCES method improves the performance of adaptive clustering selection method proposed by Azimi and Fern (2009). Thus, the performance of the HCES method was compared to that of the traditional ensemble or full ensemble. Both popular graph-based consensus clustering algorithms, CSPA and HGPA, were used in the experiments for finding the consensus solution. Time complexity of CSPA is $O(n^2kr)$ and the HGPA is $O(nkr)$ where n is the number of samples, r signifies the number of clusterings, and k denotes the sum of the clusters that exist in all clusterings. In HCES, the nodes in each layer of dendogram are the clusterings. Since the complexity of the hierarchical methods is at least $O(r^2)$ (Murtagh, 1983). Both CSPA and HGPA that are linear based on r are applied in HCES. As HGPA was suitable for large data sets, in our experiment, Satimage data set was run only by HGPA rather than CSPA.

5.1. Generating cluster ensembles

For generation of ensemble members (clusterings), the k-means algorithm with different parameter k values and different initializations is used. Different k values were selected between k_{min} and k_{max} randomly without replacement. For each selected value k, k-means

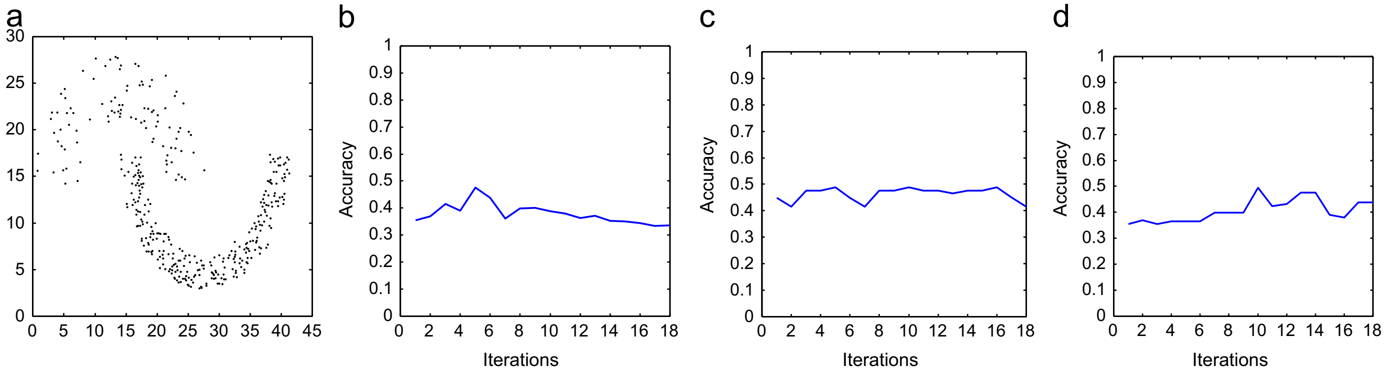


Fig. 4. Quality values of different clusterings based on different k values on 2-half rings data set.

was run r times. This approach created different diversities with different qualities. It should be noted that diversity and quality are two important factors that have impact on quality of the final solution (Fern and Brodley, 2004). Fig. 4 shows an example that why different k and different initializations were used for diversity generation. For different k values and initializations, the k-means algorithm was executed on 2-half rings data set (Jain and Law, 2005) shown in Fig. 4(a). Based on clustering quality, Fig. 4(b–d) shows the impact of different k values (between 2 and 19), one k and different iterations (k=6 and 18 iterations), and different k and different iterations for each k (k is between 2 and 8, and 3 iterations), respectively.

The graph shown in Fig. 4(b) demonstrates that the quality of clusterings in $k = 5, 6, 7$ is higher than those in k=2, while for 2-half rings data set, the correct number of clusters is 2. The average value of pair-wise diversity measure for these clusterings is $D_p = 0.2893$. In Fig. 4(c), the quality values are almost monotonic ($D_p = 0.1161$), whereas in Fig. 4(d), the quality values are not monotonic ($D_p = 0.3535$). Accordingly, using only different initializations with one k in k-means algorithm for arbitrary shape data sets as diversity may create the clusterings with low quality and high diversity, or viceversa. Therefore, in our experiments, different k (between $k_{min} = 2$ and $k_{max} = \sqrt{n}$) and for each k, different runs of k-means were selected for generating appropriate diversity in the primary ensembles.

5.2. Test results

The HCES method was evaluated by comparing its performance with that of the full ensembles. Since the HCES method obtained all subsets of full ensemble members (clusterings) in a hierarchical way, the clusterings of size 100 were generated using strategy explained in Section 5.1 on a data set. The HCES method formed different ensemble members based on quality and diversity strategies in different layers. Three agglomerative clustering methods, namely single-link, average-link, and complete-link, were applied to the HCES method. These agglomerative clustering algorithms were run using diversity measure matrix calculated by two diversity measures of Eqs. (4) and (6). Once a set of ensemble members in each layer of dendogram was selected, a consensus function, either CSPA or HGPA, was applied to obtain a consensus clustering solution. In the HCES method, in the first layer of dendogram, all ensemble members were chosen for obtaining full ensemble solution. In the second layer, 99 ensemble members were selected, and in the third layer, 98 ensemble members were selected; this went on to the 99th layer that contained only two ensemble members. All sets of ensemble members that exist in layers 2–99 were subsets of full ensemble members. In this paper, the result of full ensemble was compared to the results obtained from each layers using NMI value computed by the class label information. To evaluate the final performance of the HCES method, the average obtained from 10 NMI values was executed on each data set. The performance of the HCES method was evaluated using eight real data sets. The real data sets were extracted from the UCI data sets (available at: http://www.ics.uci/mlearn/MLRespository.html).

Table 1
Distribution of data sets.

Rank	Data set	Data size (n)	Dimension (d)	No. clusters (K)
1	Soybean (small)	47	16	4
2	Ecoli	336	7	8
3	Breast tissue	106	9	6
4	Iris	150	4	3
5	Wine	178	13	3
6	Glass	214	9	7
7	Breast cancer	699	9	2
8	Satimage	6435	36	7

The details of these data sets are presented in Table 1.

In the first experiment, using three agglomerative algorithms with applying $d1$ diversity measure, the effect of the HCES method on quality of the obtained results was compared to those obtained by the full ensemble. Fig. 5 shows a comparison between full ensemble solution and different cluster ensemble selection solutions. Different clusterings selected from different layers were tested using NMI evaluation and their quality was calculated. The CSPA algorithm was used to obtain the final results. Using CSPA, the HCES method was shown more effective on the quality of results except those of Soymbean data set. Since the ensemble size was fixed for the full ensemble, the performance was plotted as flat lines (red line). In each layer (level) of the process of the HCES, two clusterings with minimum diversity measure (i.e., high quality) were merged, then diversity between clusterings was increased until 99th layer that contained only two clusterings. The clusterings selected in the last layers showed relatively more diversity than those of the first layers. Since HCES method partitioned the cluster members based on quality/diversity, the resulting ensembles achieved competitive performances even when the ensemble size was small. In Ecoli, Wine, and Breast tissue data sets, selection of clusterings in the last layers of hierarchy, that contained small ensemble members, improved the consensus solution.

The performance of the HCES method in more than 90% of layers for Iris, Ecoli, Glass, and Soymbean data sets was not shown better than the full ensemble. According to Azimi and Fern (2009), these data sets were *stable* where majority of the values $NMI(h^*, h_i)$ were more than 0.5. However, there were some layers of Iris, Ecoli, and Glass data sets whose solutions were better than the full ensemble solution. On the other hand, the performance of the HCES method in more than 90% of layers for Breast tissue, Breast cancer, and Wine

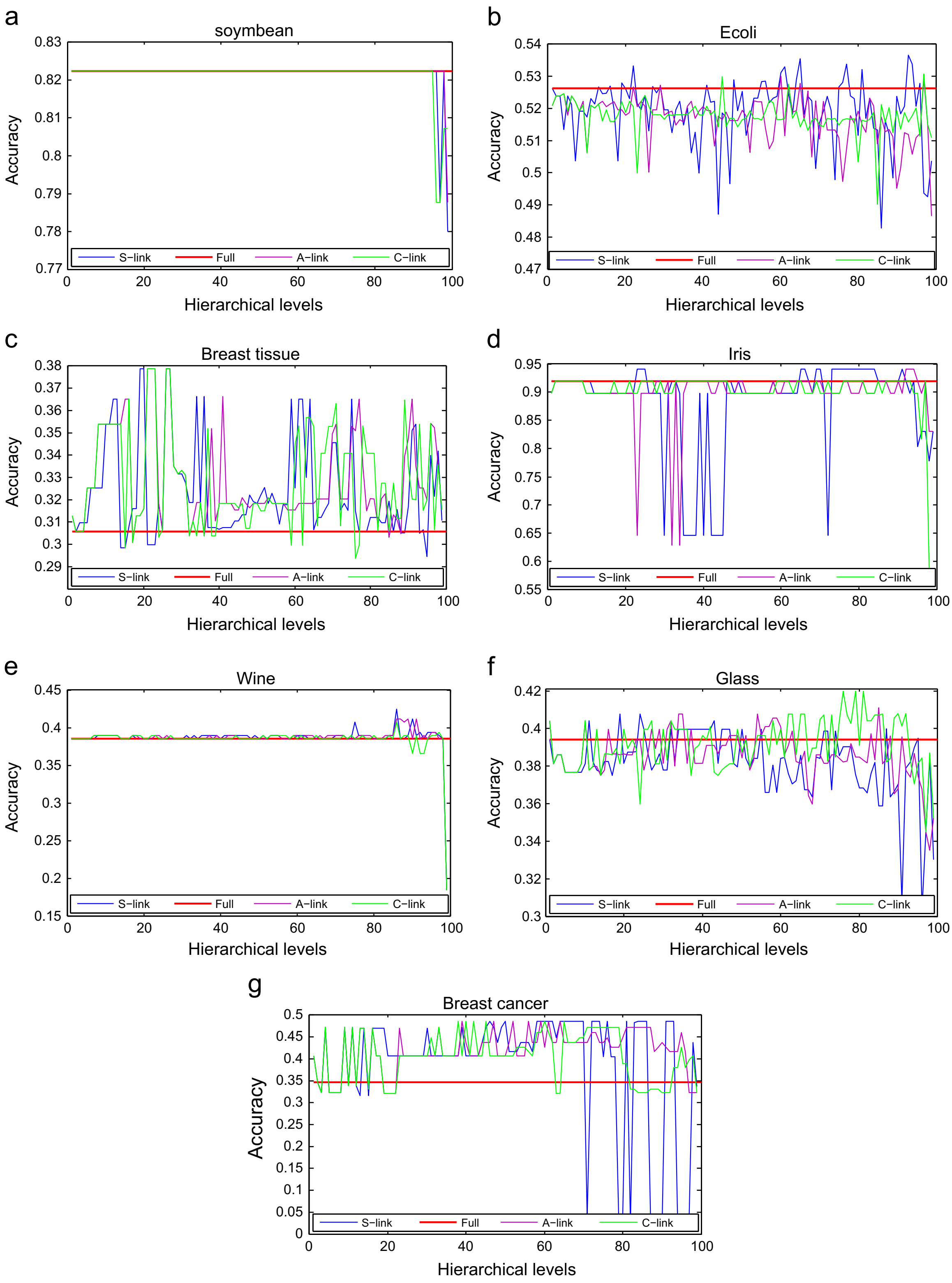


Fig. 5. Performance comparison of results obtained by CSPA for the HCES method that used single-link, average-link, and complete-link and the full ensembles. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

data sets was shown better than the full ensemble. These data sets were *non-stable* where the majority of values $NMI(h^{*}, h_i)$ were less than 0.5. As can be seen, for these data sets, the selected clusterings in the last layers with small size and high diversity improved the consensus solution. Especially, for Wine data set, the most quality was occurred in consensus solution of the small number of selected clusterings that have the highest level of diversity for all three agglomerative clustering algorithms.

Interestingly, the HCES method is sensitive to even one layer of clusterings. In other words, with elimination of even one clustering

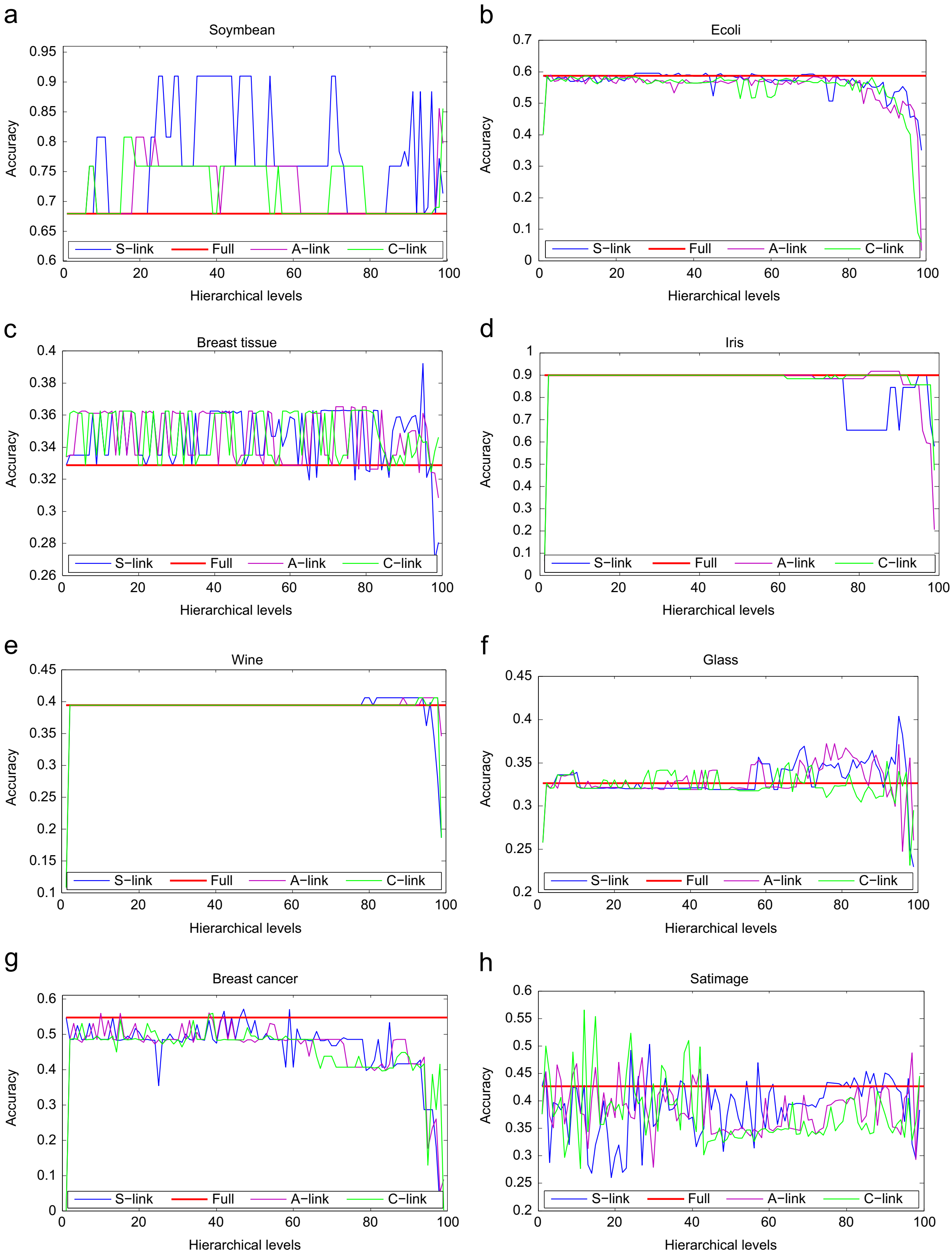


Fig. 6. Performance comparison of results obtained by HGPA for the HCES method that used single-link, average-link, and complete-link and the full ensembles.

as a noise from the clusterings of one previous layer, the quality of consensus solution significantly changes. For example, in the HCES method that uses average-link, the quality value of 23th layer in the Iris data set obtained by CSPA consensus algorithm is nearby 0.64, while the quality value of 24th layer is nearby 0.9.

Fig. 6 compares the results obtained by the full ensemble and the HCES method that uses HGPA consensus clustering algorithm. The performance of the HCES method that used HGPA and that of the HCES method that used CSPA were the same for Iris and Ecoli data sets. However, for Soymbean and Glass data sets, the performance of the HCES method was more successful than the full ensemble. The maximum quality value was obtained in the last layers of dendogram for Breast tissue, Glass, and Wine data sets in which clusterings had more diversity. The HCES method that used HGPA was more capable of improving the performance of Satimage data set compared to the full ensemble, while the HCES method that used CSPA could not be run on the Satimage data set due to restricted memory.

To ease the interpretation of Figs. 5 and 6, the comparison between the performance of three agglomerative algorithms that use two consensus clustering algorithms and the performance of full ensemble is displayed in Tables 2 and 3.

Table 2 presents the results of the comparison made, based on NMI values, between HCES method that uses single-link, average-link, and complete-link, and the full ensemble. In this table, the final solutions were obtained by HGPA algorithm. The NMI value for the obtained results of single-link, average-link, and complete-link algorithms was chosen from the best NMI values of different layers. Using single-link algorithm, the HCES gained the best solutions among other algorithms and full ensemble used by HGPA on many of data sets. In Table 2, the best results are shown by bolded numbers. For all data sets, the results obtained by the HCES method were shown better than those obtained by the full ensemble.

Table 3 shows the results obtained from the comparison made, based on NMI values, between the HCES method that uses single-link, average-link, and complete-link, and the full ensemble. In this table, the final solutions were obtained by CSPA algorithm. Using single-link algorithm, the HCES also gained the best solutions among other algorithms used by CSPA and full ensemble on many of data sets.

Table 2
Cluster quality real data sets using different hierarchical methods and full ensembles with HGPA algorithm.

Data sets	Single-link	Average-link	Complete-link	Full ensembles
Soybean	**0.9098**	0.8554	0.8554	0.6789
Ecoli	**0.5964**	0.5899	0.5881	0.5881
Breast tissue	**0.3921**	0.3652	0.3626	0.3286
Iris	**0.9405**	**0.9405**	0.9192	0.9192
Wine	**0.4063**	**0.4063**	**0.4063**	0.3948
Glass	**0.4038**	0.3722	0.3516	0.3263
Breast cancer	**0.5715**	0.5606	0.5593	0.5474
Satimage	0.5033	0.4875	**0.5651**	0.4262

Table 3
Cluster quality real data sets using different hierarchical methods and full ensembles with CSPA algorithm.

Data sets	Single-link	Average-link	Complete-link	Full ensembles
Soybean	**0.8224**	**0.8224**	**0.8224**	**0.8224**
Ecoli	**0.5366**	0.5300	0.5092	0.5263
Breast tissue	**0.3787**	**0.3787**	**0.3787**	0.3058
Iris	0.9011	**0.9192**	0.9011	0.9011
Wine	**0.4248**	0.4118	0.4075	0.3856
Glass	0.4078	0.4110	**0.4200**	0.4027
Breast cancer	**0.4854**	**0.4854**	**0.4854**	0.3465

Fig. 7 compares the performance of two consensus clustering algorithms applied to the HCES method with those applied to the full ensemble. Since HCES obtained the good results using single-link algorithm, these results are compared to full ensemble. The performance of the HCES method that used HGPA and the performance of full ensemble (Fig. 7(a)) were compared to each other and the performance of HCES that used CSPA and full ensemble were compared also to each other (Fig. 7(b)). Fig. 7 shows that the HCES method that used both HGPA and CSPA was able to achieve statistically more significant improvement compared to the full ensemble.

In the second experiment, the effect of different diversity measures on the HCES performance was examined. Two diversity measures of Eqs. (4) and (6) were applied to the HCES method. Note that the diversity measure of Eqs. (4) and (6) were shown with d_1 and d_2, respectively. The HCES method was run using only single-link algorithm, and CSPA consensus clustering algorithm was used by the HCES method. For all data sets, the HCES method based on d_2 diversity measure created the clusters of clusterings with more diversity than the created clusters of clusterings based on d_1 diversity measure (Fig. 8).

The HCES method based on d_2 diversity measure was able to achieve more improvement for all data sets except Iris and Breast cancer data sets compared to d_1 diversity measure. Furthermore, for Breast tissue and Wine data sets, there was more improvement in the last layers that had more diversity based on d_2 diversity measure. Specially, the best performance for Wine data set occurred when there was only three clusterings with maximum diversity (Fig. 8). For Iris, Glass, and Soymbean data sets based on both diversity measures, improvement quality appeared in the initial layers; whereas, for Breast tissue and Breast cancer data sets, it appeared in the middle layers. The HCES method based on d_1 diversity measure improved the quality in the middle and last layers monotonically for Breast cancer.

Table 4 demonstrates the results of comparison that was made based on NMI quality values between the HCES method that used two diversity measures d_1 and d_2 and the full ensemble. Based on both diversity measures, the quality value of the HCES method for each data set was chosen from among the best quality values of the layers.

The HCES results obtained based on diversity measure d_2 was shown better than those obtained based on diversity measure d_1. In Table 4, the best results are shown by bolded numbers. For all data sets, the results obtained by the HCES method that used two diversity measures d_1 and d_2 were shown better than those obtained by the full ensemble.

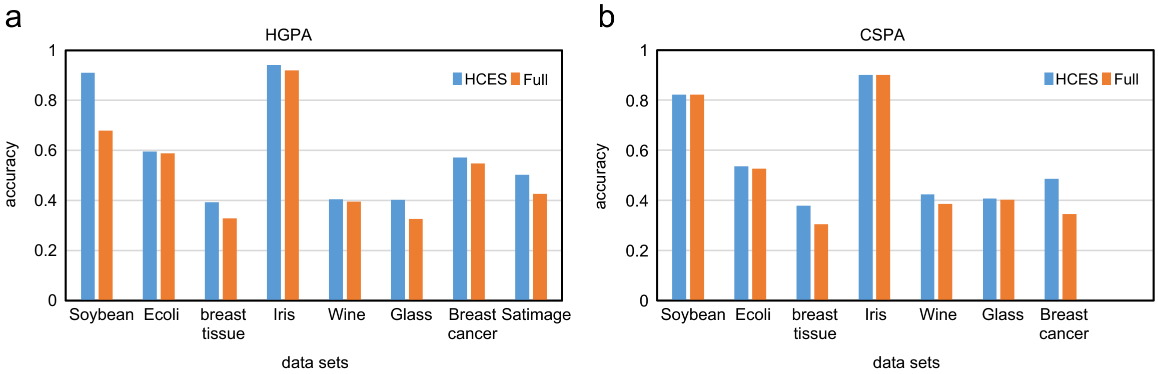


Fig. 7. Performance comparison of the HCES method that used HGPA and CSPA with the full ensemble.

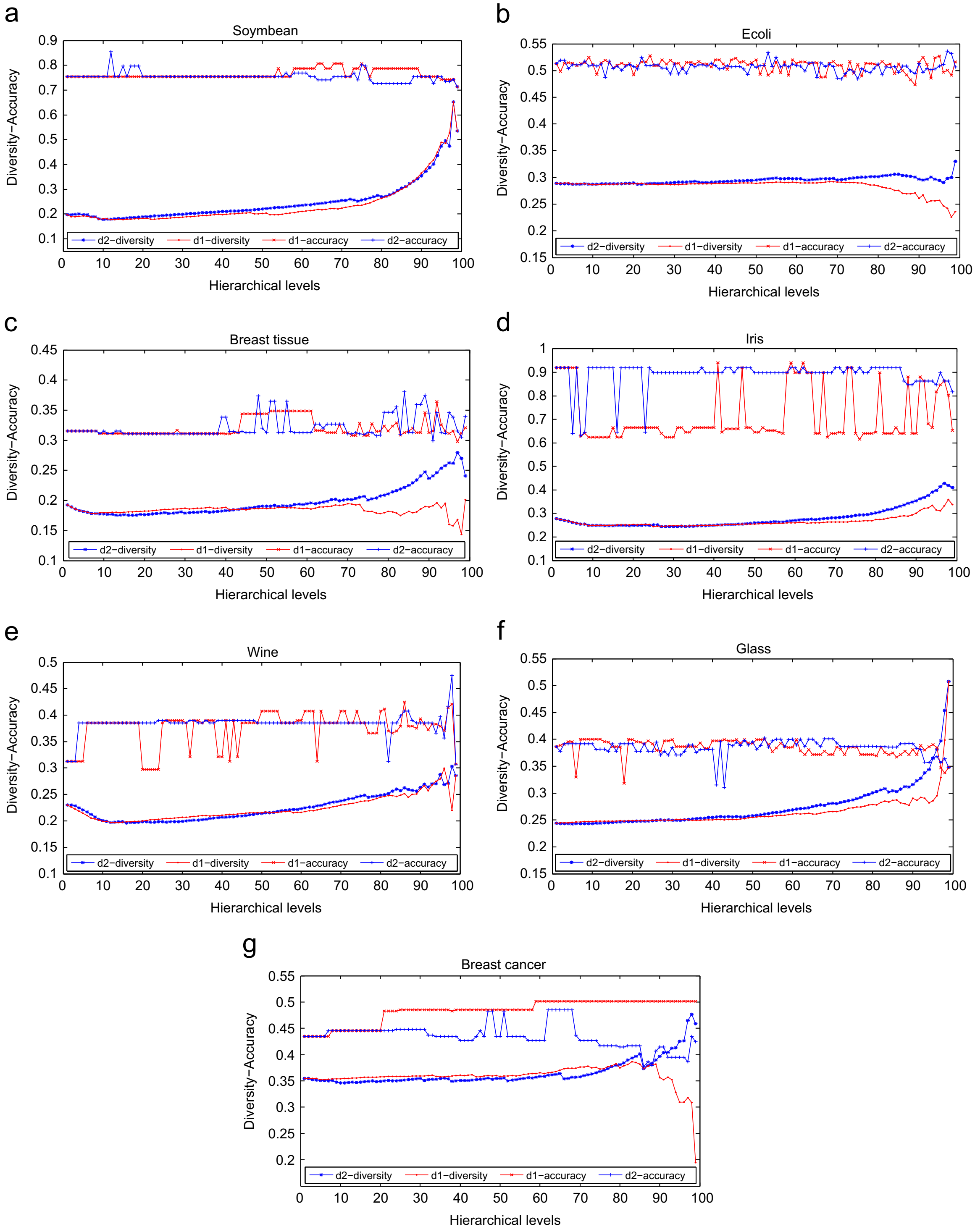


Fig. 8. Comparing the HCES method results obtained using two diversity measures, d_1 and d_2 with applying CSPA, based on diversity measures and quality values.

6. Conclusion

In this paper, a hierarchical cluster ensemble selection (HCES) method was proposed, which was shown both scalable and accurate. The method used three agglomerative clustering algorithms: single-link, average-link, and complete-link. Two consensus functions CSPA and HGPA were used for combining the full ensemble members and combining different subsets of full ensemble members. The HCES method was often more successful in finding the subset of cluster members based on quality and diversity in comparison with full ensemble. Specially, using single-link algorithm, the HCES gained the best solutions among other algorithms and full ensemble used by both CSPA and HGPA. Based on the quality and the diversity, the HCES method clustered all available clusterings hierarchically in which there was no need of the number of clusters. Our experiments were conducted on eight real data sets using two diversity measures and the obtained results showed that the HCES method achieved comparable or better results in comparison with those obtained by full ensemble. Due to the interpretability of the HCES method, applying this method in different domains such as bio-informatics, image processing, and marketing will be part of our future works.

Table 4
Comparing results of the HCES method based on d_1 and d_2 diversity measures with applying CSPA.

Data sets	d_1	d_2	Full
Soybean	0.8072	**0.8553**	0.7546
Ecoli	0.5282	**0.5362**	0.5131
breast tissue	0.3643	**0.3803**	0.3152
Iris	**0.9405**	0.9192	0.9192
Wine	0.4254	**0.4781**	0.3129
Glass	0.4004	**0.4028**	0.3864
Breast cancer	**0.5014**	0.4850	0.4347

References

Alizadeh, H., Minaei-Bidgoli, B., Parvin, H., 2014. To improve the quality of cluster ensembles by selecting a subset of base clusters. J. Exp. Theoret. Artif. Intell. 26 (1), 127–150.

Avogadri, R., Valentini, G., 2009. Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artif. Intell. Med. 45 (2), 173–183.

Azimi, J., Fern, X., 2009. Adaptive cluster ensemble selection. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI), vol. 9, Pasadena, California, pp. 992–997.

Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P., 2005. Ensemble diversity measures and their application to thinning. Inf. Fusion 6 (1), 49–62.

Berikov, V., 2013. Weighted ensemble of algorithms for complex data clustering. Pattern Recognit. Lett. 38 (15), 99–106.

Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A., 2004. Ensemble selection from libraries of models. In: Proceedings of the Twenty-first International Conference on Machine Learning. ACM, p. 18.

De Angelis, L., Dias, J.G., 2014. Mining categorical sequences from data using a hybrid clustering method. Eur. J. Operat. Res. 234 (3), 720–730.

Fern, X.Z., Brodley, C.E., 2003. Random projection for high dimensional data clustering: a cluster ensemble approach. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML), vol. 3, Washington DC, pp. 186–193.

Fern, X.Z., Brodley, C.E., 2004. Solving cluster ensemble problems by bipartite graph partitioning. In: Proceedings of the Twenty-first International Conference on Machine Learning. ACM, p. 36.

Fern, X.Z., Lin, W., 2008. Cluster ensemble selection. Stat. Anal. Data Mining 1 (3), 128–141.

Fred, A.L., Jain, A.K., 2005. Combining multiple clusterings using evidence accumulation. IEEE Trans. Pattern Anal. Mach. Intell. 27 (6), 835–850.

Hadjitodorov, S.T., Kuncheva, L.I., Todorova, L.P., 2006. Moderate diversity for better cluster ensembles. Inf. Fusion 7 (3), 264–275.

Hong, Y., Kwong, S., Chang, Y., Ren, Q., 2008. Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. Pattern Recognit. 41 (9), 2742–2756.

Hong, Y., Kwong, S., Wang, H., Ren, Q., 2009. Resampling-based selective clustering ensembles. Pattern Recognit. Lett. 30 (3), 298–305.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2 (1), 193–218.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31 (8), 651–666.

Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. 22 (1), 4–37.

Jain, A.K., Law, M.H., 2005. Data clustering: a users dilemma. In: Pattern Recognition and Machine Intelligence. Springer, Kolkata, India, pp. 1–10.

Jia, J., Xiao, X., Liu, B., Jiao, L., 2011. Bagging-based spectral clustering ensemble selection. Pattern Recognit. Lett. 32 (10), 1456–1467.

Karypis, G., Kumar, V., 1998. Multilevel k-way partitioning scheme for irregular graphs. J. Parallel Distrib. Comput. 48 (1), 96–129.

Kleinberg, J., 2003. An impossibility theorem for clustering. Adv. Neural Inf. Process. Syst., 463–470.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Naval Res. Logist. Q. 2, 83–97.

Kuncheva, L.I., Hadjitodorov, S.T., 2004. Using diversity in cluster ensembles. In: Proceedings of International Conference on Systems, Man and Cybernetics, vol. 2. IEEE, pp. 1214–1219.

Lu, X., Yang, Y., Wang, H., 2013. Selective clustering ensemble based on covariance. In: Multiple Classifier Systems. Springer, Nanjing, China, pp. 179–189.

Ma, X., Wan, W., Jiao, L., 2009. Spectral clustering ensemble for image segmentation. In: Proceedings of the First ACM/SIGEVO Summit on Genetic and Evolutionary Computation. ACM, Shanghai, China, pp. 415–420.

Mimaroglu, S., Erdil, E., 2010. Obtaining better quality final clustering by merging a collection of clusterings. Bioinformatics 26 (20), 2645–2646.

Mimaroglu, S., Erdil, E., 2013. An efficient and scalable family of algorithms for combining clusterings. Eng. Appl. Artif. Intell. 26 (10), 2525–2539.

Minaei-Bidgoli, B., Parvin, H., Alinejad-Rokny, H., Alizadeh, H., Punch, W.F., 2014. Effects of resampling method and adaptation on clustering ensemble efficacy. Artif. Intell. Rev. 41 (1), 27–48.

Minaei-Bidgoli, B., Topchy, A., Punch, W.F., 2004. Ensembles of partitions via data resampling. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC), vol. 2, IEEE, pp. 188–192.

Murtagh, F., 1983. A survey of recent advances in hierarchical clustering algorithms. Comput. J. 26 (4), 354–359.

Naldi, M., Carvalho, A., Campello, R., 2013. Cluster ensemble selection based on relative validity indexes. Data Mining Knowl. Discov. 27 (2), 259–289.

Quintana, F.J., Getz, G., Hed, G., Domany, E., Cohen, I.R., 2003. Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity. J. Autoimmun. 21 (1), 65–75.

Strehl, A., Ghosh, J., 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617.

Sun, J., Chen, W., Fang, W., Wun, X., Xu, W., 2012. Gene expression data analysis with the clustering method based on an improved quantum-behaved particle swarm optimization. Eng. Appl. Artif. Intell. 25 (2), 376–391.

Topchy, A., Jain, A.K., Punch, W., 2003. Combining multiple weak clusterings. In: Proceedings of the Third International Conference on Data Mining (ICDM). IEEE, Melbourne, Florida, pp. 331–338.

Topchy, A., Jain, A.K., Punch, W., 2004. A mixture model of clustering ensembles. In: Proceedings of the International Conference on Data Mining. SIAM, Florida, pp. 379–390.

Topchy, A., Jain, A.K., Punch, W., 2005. Clustering ensembles: models of consensus and weak partitions. IEEE Trans. Pattern Anal. Mach. Intell. 27 (12), 1866–1881.

Wang, X., Han, D., Han, C., 2013. Rough set based cluster ensemble selection. In: Proceedings of 16th International Conference on Information Fusion (FUSION). IEEE, Istanbul, Turkey, pp. 438–444.

Yang, F., Li, X., Li, Q., Li, T., 2014. Exploring the diversity in cluster ensemble generation: random sampling and random projection. Expert Syst. Appl. 41 (10), 4844–4866.

Yu, Z., Wong, H.-S., You, J., Yu, G., Han, G., 2012. Hybrid cluster ensemble framework based on the random combination of data transformation operators. Pattern Recognit. 45 (5), 1826–1837.

Zhang, Y., Burer, S., Street, W.N., 2006. Ensemble pruning via semi-definite programming. J. Mach. Learn. Res. 7, 1315–1338.