Cluster ensemble extraction for knowledge reuse framework

Ebrahim Akbari,Halina Mohamed Dahlan,Roliana Ibrahim
Universiti Teknologi Malaysia
Faculty of Computing
81310 Johor Bahru
Malaysia
ebrahimakbari30@yahoo.com,akbari@iausari.ac.ir,{halina,roliana}@utm.my

Abstract: Cluster ensemble framework attempts to find stable and robust results through composing calculated clusterings obtained from basic clustering algorithms without accessing the features or algorithms that determine these clusterings. Diversity of clusterings is a important factor for improving cluster ensemble performance, where an ensemble of small size of identical clusterings dose not improve the quality and robustness of solution. Concerning limited access to the raw data, how new clusterings with more diversity and size can be created using a few base clusterings. This paper proposes a new approach, cluster ensemble extraction, as a knowledge reuse framework to create a new diversity without accessing the raw data. This approach creates a new set of clusterings from the existing clusterings, which have more diversity and size compared to base clusterings. To evaluate the performance of the proposed approach, several experiments were conducted on several real data sets and the results were compered to the results obtained from executing of cluster ensemble on base clusterings. The comparison results showed the superiority of the proposed approach over the cluster ensemble approach in terms of quality.

Key–Words: Clustering, Knowledge reuse, Diversity, Cluster ensemble extraction

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 [1, 2]. Wide usage of clustering algorithms proves their usefulness in exploratory data analysis [3, 4]. 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 [5]. Therefore, it is difficult to choose a suitable algorithm for a given dataset. Based on the Kleinberg theorem [6], there is no one best single clustering algorithm.

Clustering ensemble (CE) is considered as combining multiple clustering results (clusterings) into final clusters without accessing the features or algorithms. Combining the clusterings are used by a consensus algorithm. Since CE only needs access to the base clusterings (BC) instead of the data itself, it provides a convenient approach to privacy preservation and knowledge reuse. Through composing the BC, the CE approach can achieve some characteristics such as novelty, robustness, stability, and scalability [7, 8].

The accuracy of consensus solution obtained by a consensus algorithm is affected by both quality and diversity of BC. Thus, an ensemble has not acceptable performance on BC obtained from identical single clustering algorithms [9]. Usually a subset of all available clusterings may have more quality and diversity compared to all available clusterings [9, 10]. The main objective of cluster ensemble selection (CES) approach is choosing a subset from a large library of clustering solutions (clusterings) in order to create a smaller cluster ensemble that can perform as appropriately as or better than the set of all available clustering solutions [10, 11, 12]. However, the CE and CES approaches require a large library of BC. In addition, diversity and quality in CES is very related to BC.

This paper proposes a new approach, cluster ensemble extraction (CEE), to improve the consensus solution by extracting the existing clusterings without accessing the raw data. Our contribution in this paper is generating new diversity with different size using a few existing clusterings (BC) without accessing the data. The new diversity is extracted from the BC without using a diversity measure. For generating new diversity from BC, three different methods are proposed: (1) applying different consensus algorithms; (2) using the same consensus algorithm with different parameters such as different number of clusters; and

(3) using different subsets of BC. In our approach, we use all the three methods to generate new diversity by clusterings extraction algorithm (CEA). Effects of different consensus algorithms, here CSPA and HGPA, on BC and EC are experimented. The performance of CSPA and HGPA on EC was compared empirically to those on BC. The evaluation results obtained from different real data sets demonstrated statistically more successful performance of CEE compared to that of CE.

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 cluster ensemble extraction approach in witch clusterings extraction algorithm is presented for generation new diversity. Section 5 presents the experiments carried out on several real datasets and the obtained results. Finally, section 6 concludes the paper and recommends future work.

2 Related work

CE is an approach widely adopted in clustering research to improve the quality and robustness of clustering results. CE 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 [9, 13, 14]. Figure 1 shows the steps of CE and CES approaches. Here, the review of CE and CES methods and some milestone studies conducted on cluster ensemble design are discussed.

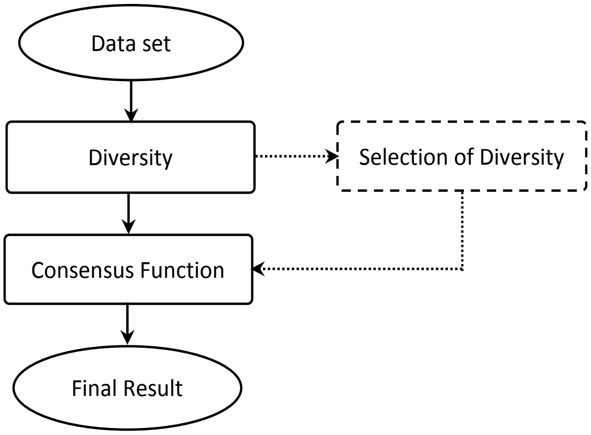


Figure 1: Steps of the clustering ensemble selection approach

In ensemble classifier/clustering techniques, generating diversity is commonly used in supervised and unsupervised combining approaches. Various Methods have been proposed in literature for creating diversity or BC, 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 [15].

2. Different clustering algorithms: a number of different clustering algorithms are used together to generate primary clusterings, which are called heterogeneous ensembles [8, 16].

3. Different subsets of features: features are selected or extracted to create subsets used for the generation of clusterings [8, 15, 17].

4. Different subsets of objects: data are re-sampled with or without replacement for generating clusterings [18, 19].

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 [8, 20].

Consensus function is an algorithm for combining different clusterings (BC) to obtain final clusters [7, 21]. Assume that H is a set of BC, $H = \{h_1, h_2, ..., h_L\}$, where L is size of BC, the consensus function Φ combines all BC of H as $h^* = \Phi(H)$. The value of h^* is result of sharing the most information with the BC. In the CES, the consensus function applies on a subset of BC instead of all. The consensus function for CES is defined as $h_s^* = \Phi(H_s)$, where, $H_s \subset H$. The literature contains several CE 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 [22]. 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 and

Jain [23] have proposed a function called 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 [15]. The graph-based approach includes instance-based, cluster-based, and hybrid approaches. Strehl and Ghosh [7] explore three graph-based consensus algorithms named Cluster-based Similarity Partitioning Algorithm (CSPA), HyperGraph-Partitioning Algorithm (HGPA), and Meta-CLustering Algorithm (MCLA). The CSPA as an instance-based approach constructs a hypergraph in which the number of frequency of two objects which are accrued in the same clusters are considered as weight of each edge. The *k* partitions are obtained using the METIS [24] on the induced similarity graph. On the other hand, MCLA is a famous cluster-based method in which the Jaccard measure is applied as similarity measure between two corresponding clusters. MCLA constructs a metagraph in which clusters are considered as vertices, and the similarity measure between clusters (vertices) are calculated as weight of the edges. 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 [25].

Recently, CES techniques have been proposed to improve the CE performance [9, 10, 12, 26]. These techniques select a subset of BC based on both diversity and quality that are two important factors for improvement of the CE solution [9, 10, 12, 20]. If the generated ensemble members (BC) are different from each other and they also have an acceptable quality, a better CE solution can be achieved [27].

In literature, there are different quality and diversity measures considered for BC [9, 28, 29, 30]. Most of them are based on match index between two partitions. Two diversity measures commonly used in literature are Adjusted Rand Index (ARI) [31] and Normalized Mutual Information (NMI) [7]. These measures are also used for measuring accuracy between two partitions. Hadjitodorov *et al.* [9] used ARI diversity measure on a large number of candidate ensembles (BC) for selection. They constructed four diversity measures based on ARI and found the median of the diversity values for BC and picked the corresponding ensemble. Lu *et al.* [28] introduced a diversity measure based on covariance. Alizadeh *et al.* [30] proposed a CES method in which clusters (instead of clusterings) were selected based on quality and diversity measure. Naldi *et al.* [29] proposed several relative cluster validity indices based on quality and diversity for selection of BC. Using different relative diversity measures, they also investigated the impact of the diversity on BC used for the ensemble. Azimi and Fern [12] proposed adaptive cluster ensemble selection method in which datasets were divided to $stable$ and $non-stable$ based on NMI values. They demonstrated that, for $non-stable$ datasets, the selection of BC with more diversity made an improvement in the solution. Jia *et al.* [13] generalized the selective clustering ensemble algorithm proposed by Azimi and Fern [12] and proposed a novel CES method, namely, Selective Spectral Clustering Ensemble (SELSCE). BC were generated by spectral clustering (SC) that was able to engender diverse committees. The random scaling parameter, Nystrm approximation, and random initialization were used for producing the components (BC) of the ensemble system. After the generation of BC, the bagging technique was used to rank and assess the BC. Based on this ranking, BC were selected for ensemble.

All the pervious works need to large library of BC for ensemble. If a few clusterings with the same quality exist, a consensus algorithm cannot be appropriately performed on BC. Thus, the consensus solution may not accuracy, novelty, and so on. Since diversity between clusterings is a critical factor to improve the consensus solution, the question is that how diversity from the existing clusterings (BC) without accessing the data can be created. In addition, all CES approaches proposed to improve the accuracy use at least one diversity measure. These diversity measures are not deterministic because data is unlabel in clustering.

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) [7] and adjusted rand index (ARI) [31] are commonly employed to measure the diversity or quality of partition(s). The ARI and NMI quality measures are calculated, respectively as follow:

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

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

and

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

where, in both equations, $h_a=\{c_1^a,c_2^a,...,c_{k_a}^a\}$ and $h_b=\{c_1^b,c_2^b,...,c_{k_b}^b\}$ 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.

Both NMI and ARI quality measures are based on label matching between two clusterings. In literature, the value of accuracy of the clustering result is obtained by a quality measure (e.g., NMI) based on a predefined class label. If $\overline{h}$ is a known class label and h^* is consensus result or clustering result, the accuracy value is obtained by $NMI(\overline{h},h^*)$. This value indicates the accuracy of result. In this paper, CE and CEE results are compared using the NMI values.

The quality of clusterings can be measured by $NMI(\overline{h},h_i)$ as an external criteria or $NMI(h^*,h_i)$ as an internal criteria, $i=1,2,...,L$. A high value of these measures indicates that the clusterings have high quality and low diversity. Whereas, a low value of these measures indicates that the clusterings have low quality and high diversity. Thus, these measures are applied to both quality and diversity measures. Since, in the clustering, there is no label, the internal criteria is used for choosing a subset of BC based on quality and diversity in CES methods. The external criteria usually is used for testing final results based on quality and diversity.

4 Cluster Ensemble Extraction

The CEE approach is very simple and efficient when the BC are small, and their qualities are almost the same and raw data is not available. Given a set of BC as an input, first, a new set of clusterings (diversity) is obtained by extracting it from BC; then, a consensus solution is obtained by applying a proper consensus algorithm to the new set of clusterings. Consensus algorithms usually does not have an appropriate performance on identical clusterings [9]. EC often have more diversity than BC; note that the diversity is an important factor to improve the consensus solution. A new diversity can be extracted in three methods: (1) using different consensus algorithms, (2) using various parameters such as different number of clusters, and (3) using different subsets of all available clusterings.

In the first method, comparative performance of different consensus algorithms can vary significantly across BC. For example, in case of most of the data sets, MCLA and CSPA outperformed HGPA in terms of accuracy [15]. Thus, different consensus algorithms obtained different solutions with different accuracy on the same clusterings[7, 8, 15].

In the second method, number of clusters (k) is often not known in advance, while the value of k for many consensus algorithms is given by experts, which can differ for a Consensus algorithm. Therefore, one consensus function obtains different solutions with different k, leading to different accuracy values for the consensus solutions.

In the third method, a subset of BC may have more diversity compared to BC [12, 13]. Based on the new diversity, consensus algorithm obtains a solution whose quality may different compared to the consensus solution based on BC. Thus, a consensus algorithm on different subsets of BC obtains the solutions with different accuracy values.

Using the above-mentioned three methods, extraction of new clusterings from the BC led to new diversity with different qualities. furthermore, using extraction, we obtained new clusterings with sizes different from or equal to that of BC. As an illustrative example, let the following label vectors [7] specify four clusterings of the same set of eight objects:
$h_1=(1,1,2,2,1,1,2,2)$, $h_2=(1,1,1,1,1,2,2,2)$
$h_3=(1,1,2,2,2,2,1,1)$, $h_4=(1,1,1,2,2,2,2,1)$
Using CSPA on the four clusterings, the consensus solution is $h^*=(1,1,1,1,1,2,2,2)$, where the number of clusters in final clustering is 2 ($k=2$). Using the CEE approach, we extract four new clusterings via two consensus algorithms, CSPA and HGPA. Two new clusterings are obtained by CSPA with $k=3,4$, and two other clusterings are obtained by HGPA with $k=3,4$ based on BC. Note that since the number of clusterings in BC is few ($L=4$), for extraction of new clusterings, the whole BC are used. The four EC are represented by label vectors as follow:
$p_1=(1,1,1,2,2,2,2,3)$, $p_2=(1,2,1,3,4,3,4,2)$
$p_3=(2,2,2,3,3,3,1,1)$, $p_4=(1,1,1,2,2,2,3,4)$
The consensus solution using CSPA with $k=2$ on the EC is $p^*=(2,2,2,1,1,1,1,1)$. The values of diversity measure of BC, $h_i, i=1,2,3,4$, based on h^* using NMI ($1-NMI(h^*,h_i)$) are 0.0499, 1.0000, 0.0499, and 0.0499, respectively, while, the values of diversity measure of EC, $p_i, i=1,2,3,4$, based on p^* using NMI ($1-NMI(p^*,h_i)$) are 0.1912, 0.5231, 0.2412, and 0.5734, respectively. This example shows that diversity values of BC are either extremely big or extremely small, whereas those of EC are roughly moderate (around 0.5). In addition, the consensus solution for BC has not novelty where the h^* is the same as h_2.

On the other hand, the consensus solution for EC has novelty where the p^* does not exist in any sets of BC and EC.

General framework of CEE is shown in Figure2.

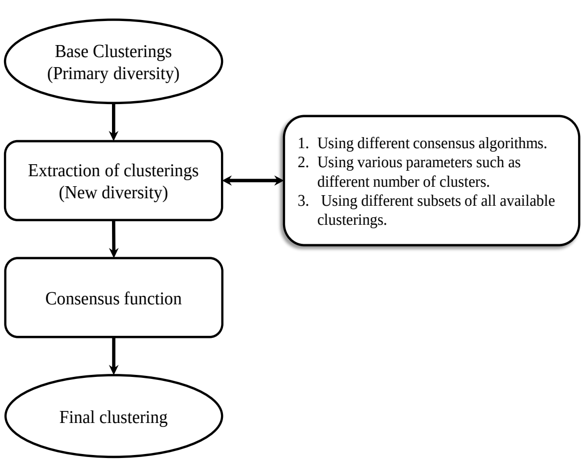

Base Clusterings
(Primary diversity)
Extraction of clusterings
(New diversity)
1. Using different consensus algorithms.
2. Using various parameters such as different number of clusters.
3. Using different subsets of all available clusterings.
Consensus function
Final clustering

Figure 2: Steps of the cluster ensemble extraction approach

There are different methods for extraction of clusterings as explained earlier. Clusterings Extraction Algorithm (CEA) extracts a new clusterings from BC.

Algorithm 1 Clusterings Extraction Algorithm

```
Input: $H = \{h_1, h_2, ..., h_L\}$, set of $L$ base clusterings
Output: $P = \{p_1, p_2, ..., p_M\}$, set of $M$
        extracted clusterings
Initialization:
 Let $P = \emptyset$ denote the empty set
  for r=1 to $M$
      Let $C = \emptyset$ denote the empty set
      $C = \{$ a subset of base clusterings randomly
             with replacement $\}$,
             where, the size of $C$ can be $\lceil \frac{L}{2} \rceil$
      Choose a consensus algorithm ($\phi$)
             to apply on set $C$
      Choose a number of final clusters ($k$)
      $p_r = \phi(C, k)$, where $p_r$ is consensus solution
              based on set $C$ and $k$
      $P = P \cup \{p_r\}$
  end for
```

The CEA algorithm uses three diversity generation mechanisms simultaneously for extraction of new diversity.

In the CEE approach, consensus solution is obtained by a consensus algorithm on EC (set P in Algorithm 1); whereas, in the CE approach, the consensus algorithm obtains the solution based on BC (set H in Algorithm 1).

5 Experimental Results

In our experiments, the CEE solutions were compared with CE solutions based on NMI values. The experiments were conducted with real data sets, where true natural clusters were known. Since our data sets were labeled, we could assess the quality of the clustering solutions using external criteria [7]. Note that although the CEE extracts new clusterings from BC without accessing the data, for generation of BC, we had to gain access to the original features. In this paper, k-means algorithm with different location of initial cluster centers generates BC with almost the same qualities [9]. The external criteria was used to measure the discrepancy between the structure defined by a clustering and the one defined by the class labels. Two consensus algorithms, CSPA and HGPA, were used for obtaining solutions in our experiments.

The whole experiments were run for 10 times and their results were averaged on each dataset. The performance of CEE was evaluated using seven real data sets. The real data sets were extracted from the UCI data sets (available at:`http://www.ics.uci/mlearn/MLRespository.html`).
The details of these data sets are presented in Table 1.

Table 1: Distribution of datasets

Number	Data set	(n)	(d)	(k)
1	Symbion (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	10	7
7	Breast cancer	699	9	2

BC were obtained by k-means with one k value that was randomly chosen between $[2, \sqrt{n}]$ and 30 iterations (number of BC is $L = 30$). Our experiments included two parts; in the first one, the EC of the size M ($M = L = 30$) were extracted by ECA algorithm. In the second part, different EC were extracted by ECA algorithm, where the number of the EC was from 10 to 100 with incremental step 10.

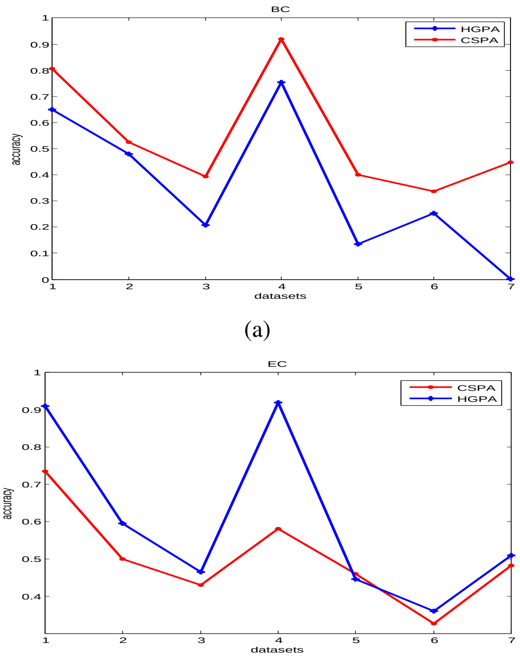
In the first part, clustering performances of CEE and those of traditional CE approaches were compared using CSPA and HGPA consensus algorithms. The CEE solutions were obtained by executing CSPA and HGPA algorithms on EC separately, and CE solutions were obtained by executing CSPA and HGPA

Table 2: Comparing results between CSPA and HGPA for base clusterings

Number	Data sets	CSPA	HGPA
1	Soymbean	0.8072	0.6501
2	Ecoli	0.5239	0.4786
3	Breast tissue	0.3928	0.2070
4	Iris	0.9192	0.7543
5	Wine	0.3995	0.1341
6	Glass	0.3365	0.2536
7	Breast canser	0.4480	0.0014

Table 3: Comparing results between CSPA and HGPA for Extracted clusterings

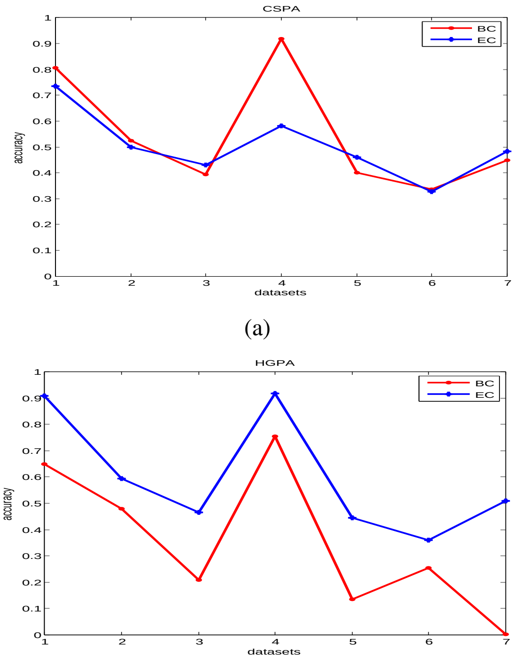
Number	Data sets	CSPA	HGPA
1	Soymbean	0.7345	0.9098
2	Ecoli	0.4999	0.5954
3	Breast tissue	0.4301	0.4653
4	Iris	0.5813	0.9192
5	Wine	0.4604	0.4448
6	Glass	0.3259	0.3592
7	Breast canser	0.4826	0.5092



(a)

(b)

Figure 3: Comparison of CSPA and HGPA for BC



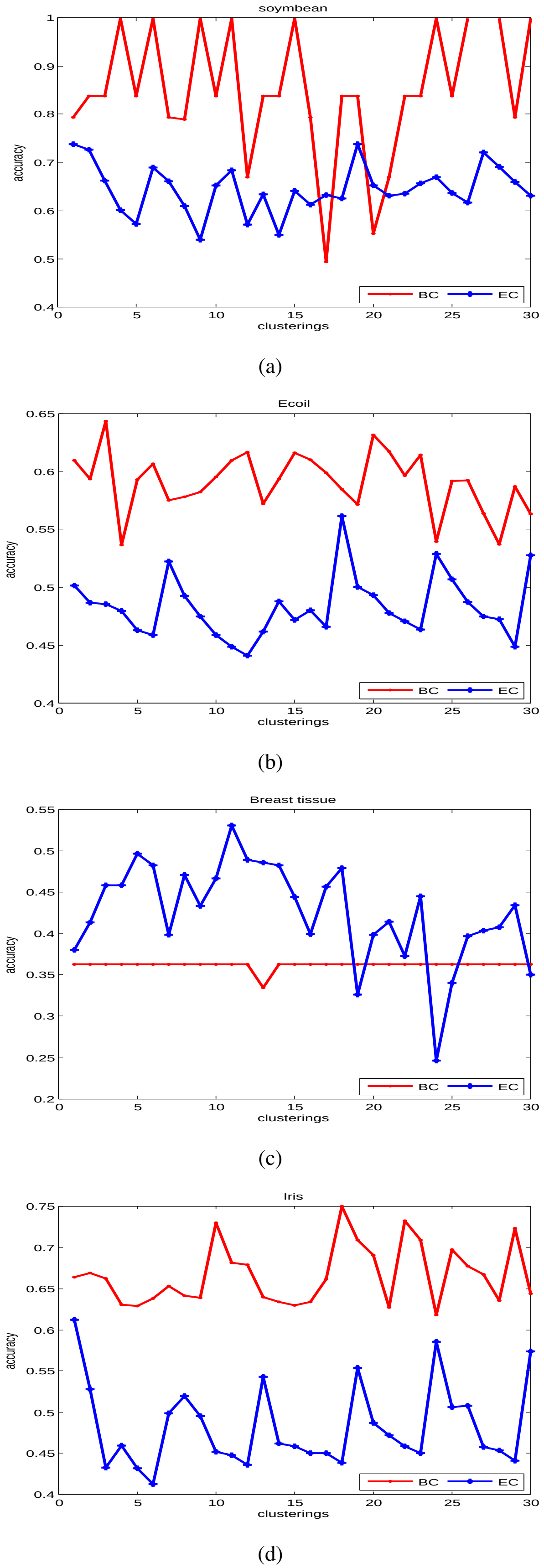
(a)

(b)

Figure 4: Comparison of CEE and CE approaches using CSPA and HGPA

algorithms on BC. Table 2, showing accuracy values of CSPA and HGPA methods for BC, demonstrates that CSPA method achieves better solutions for all tested data sets compared to HGPA method. For example, the accuracy values for Iris and Wine data sets using CSPA method are 0.9192 and 0.3995, respectively; whereas these values in case of HGPA method are 0.7543 and 0.1341, respectively. Figure 3(a) shows the effect of the two consensus algorithms on BC based on accuracy values obtained from Table 2. Table 3 shows accuracy values of CSPA and HGPA for EC. Unlike the results of Table 2, Table 3 demonstrates that HGPA method achieves better solutions for all tested data sets except for Wine dataset compared to CSPA method. For example, the accuracy values for Iris and Soymbean data sets using HGPA method are 0.9192 and 0.9098, respectively; while these values in case of CSAP method are 0.5813 0.7345, respectively. Figure 3(b) shows the effect of the two consensus algorithms on EC based on accuracy values obtained from Table 3. Figure 4, based on values presented in Tables 2 and 3, compares CEE approach with the traditional CE approach using CSPA (Figure4(a)) and HGPA (Figure4(b)). Figure4 shows that, executing on the tested data sets, CEE achieves comparative or better solutions compared to the traditional CE. Furthermore, the performance of CEE closely depends on the consensus algorithm. Using HGPA, the CEE achieves better solutions for all data sets compared to CE; whereas, using CSPA, the CEE obtains the solutions that have less accuracy for some data sets such as Soymbean, Ecoli, Iris, and Glass data sets. Figure 5 is plotted based on the values of accuracies of BC and EC using NMI with a known class label for each dataset. Note that the higher the value of NMI indicates the lower diversity and viceversa. The accuracy/diversity for EC and BC is calculated based on HGPA. Figure 5 shows that EC have more diversity compared to BC for all data sets except for the Breast tissue. Moreover, the BC for Wine and Breast tissue data sets have monotonic quality. As Figures 4(a) and 5 show, the use of CSPA in CEE dose not lead to an obvious improvement, even it may be worse than CSPA in CE. BC have higher quality (less diversity) compared to EC that have less quality (more quality). On the other hand, in Wine and Breast tissue data sets, BC have monotonic qualities while EC have non-monotonic qualities. In Figures 4(b) and 5, it can be seen that the use of HGPA causes a significant improvement in CEE approach for all data sets. Finally, according to Figures 4 and 5, HGPA obtains more accurate results when ensemble members (BC or EC) have more diversity, while CSPA obtains more accuracy results when ensemble members (BC or EC) have more quality.

In the second part, CEE generates different EC the size between 10 an 100 with incremental step 10. CSPA and HGPA are applied to different EC for each dataset . CSPA and HGPA are also applied to BC. Figure 6 makes a comparison between the performance of CEE and CE approaches with varying EC sizes on seven data sets using CSPA as consensus algorithm. The horizontal axis represents the EC size, and the vertical one is the NMI value between the final consensus solution and the real class label for each dataset. Note that each point in the graph is obtained by averaging on ten runs. In the following, we discuss the performance of CEE based on the results shown in Figure 6. Comparing with CE, CEE achieved comparable or improved performance in most of the data sets. In particular, it achieved statistically significant improvement for the Ecoli, Wine, Breast tissue, and Glass data sets. Figure 7 compares the performance of CEE and CE approaches with varying EC sizes on seven data sets using HGPA as consensus algorithm. Compared to CE, CEE improves the performance for all data sets in all EC sizes except for Iris dataset. Interestingly, the solutions of CEE using HGPA as consensus algorithm have more significant accuracy for all data sets compared to CE solutions. Moreover, the solutions of CEE using HGPA have more accuracy compared to the solutions of CEE using CSPA for most of the data sets. The final result presented in Figures 6 and 7 indicates that the performance of CEE depends closely on the applied consensus algorithms.

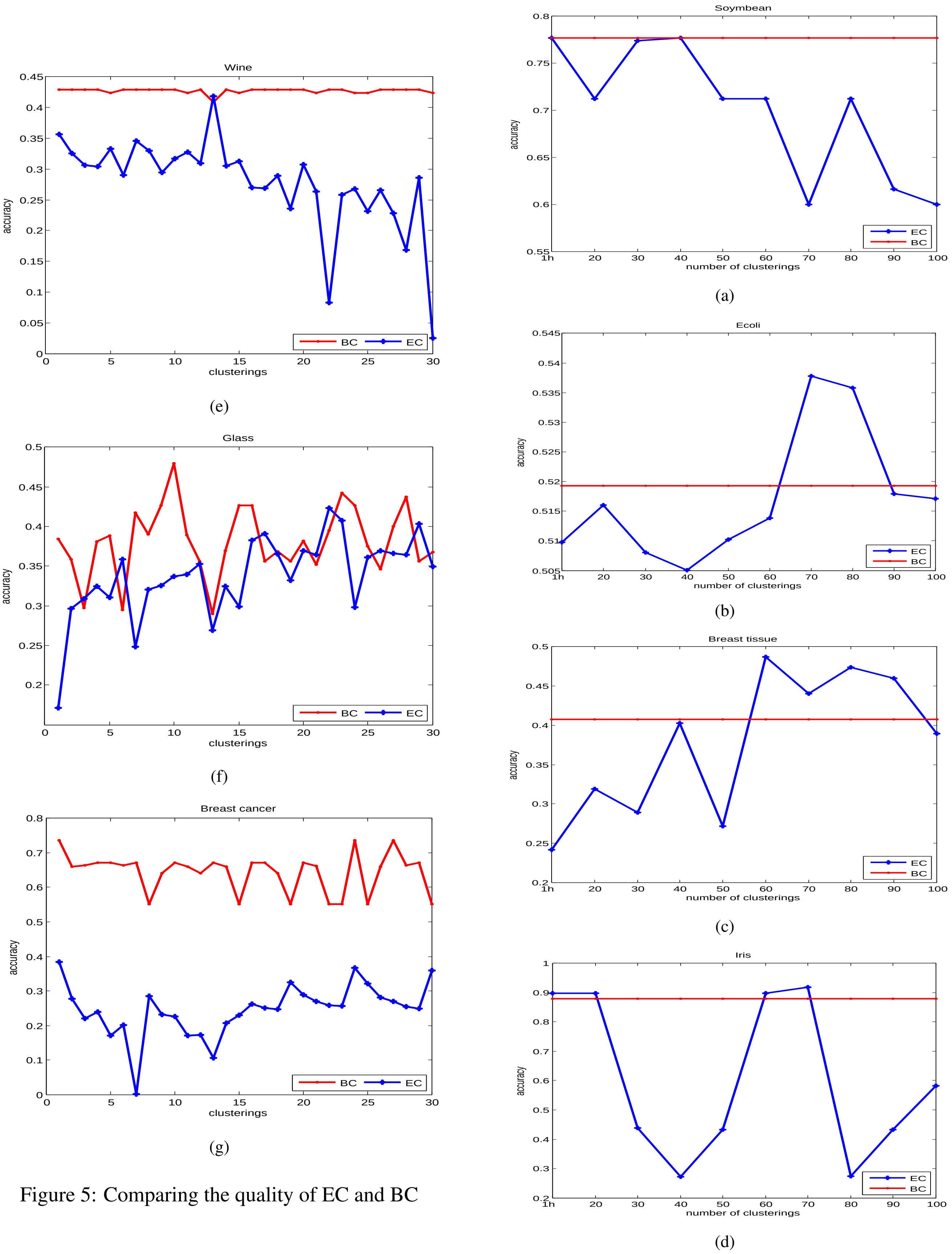


(a)

(b)

(c)

(d)



(a)

(b)

(c)

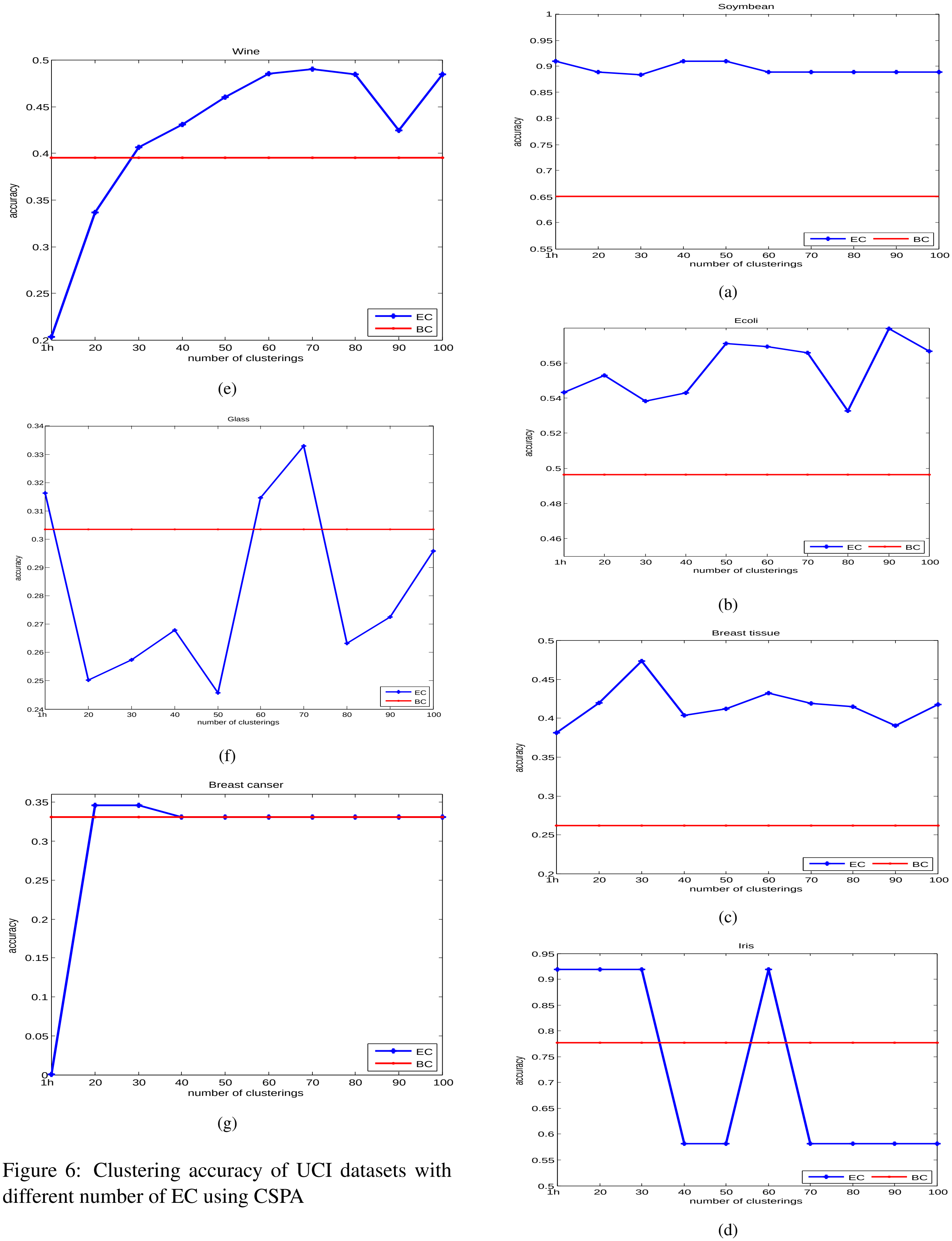
(d)

(e)

(f)

(g)

Figure 5: Comparing the quality of EC and BC



(a)

(b)

(c)

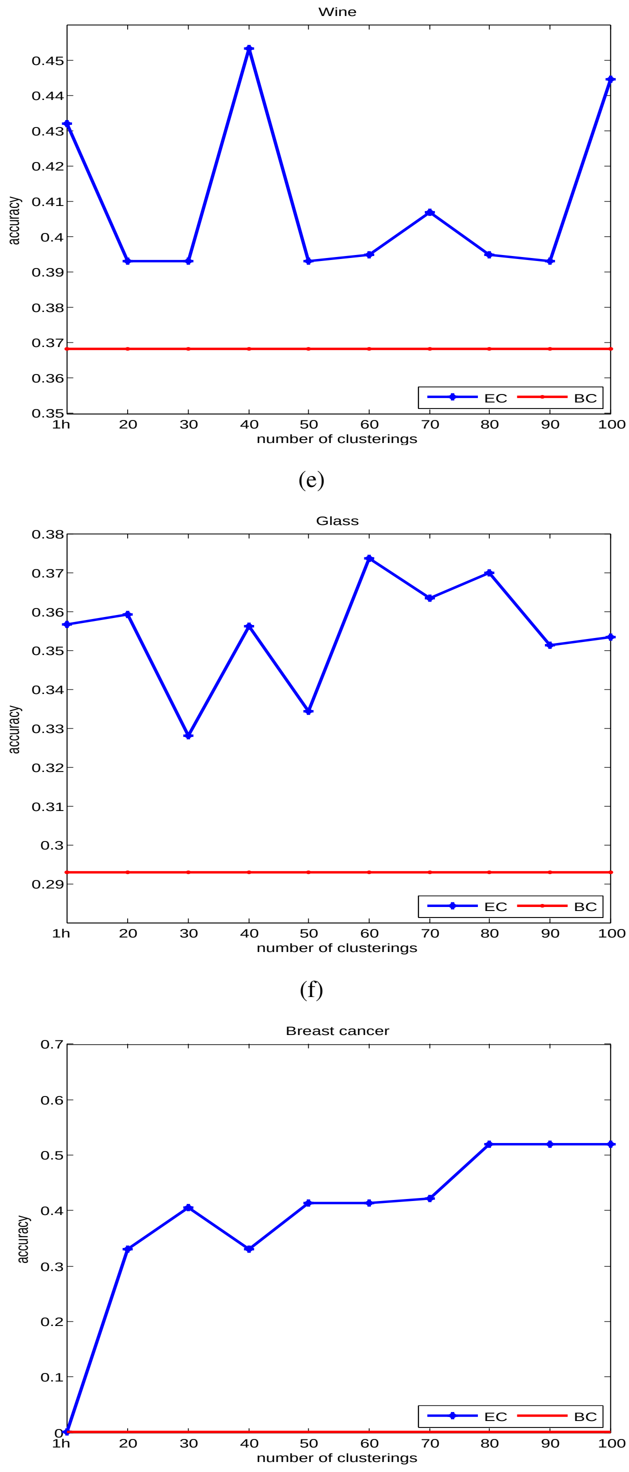
(d)

(e)

(f)

(g)

Figure 6: Clustering accuracy of UCI datasets with different number of EC using CSPA



(e)

(f)

(g)

Figure 7: Clustering accuracy of UCI datasets with different number of EC using HGPA

6 Conclusion

In this paper, the CEE approach was proposed in which the ECA algorithm was used for extraction of new clusterings from base clusterings. Unlike the traditional CE or CES that needed a large library of base clusterings, the CEE could generate a large library of clusterings from few base clusterings using extraction of clusterings. A new diversity was generated without using a diversity measure. From the base clusterings of the size of 30, the ECA generated different new EC with the size of 10 to 100 by incremental step 10. Our experiments showed that the CEE using CSPA and HGPA further improved the results compared to CE using CSPA and HGPA. In addition, we experimentally showed that the CSPA achieved good quality solutions when the clusterings had a high quality. On the other hand, the HGPA obtained the solutions with high quality when the clusterings had more diversity. Generally, CEE achieved statistically significant improvements for all data sets compared to CE. Further study need to perform to generate large library of clusterings from base clusterings of small size by mathematical methods as an optimization problem.

References:

[1] J. Xue, X. Liu, and J. Shandong, A clustering algorithm using DNA computing based on three-dimensional DNA structure and grid tree, *WSEAS Transactions On Information Science And Applications.* 5, 2012(9), pp. 137–146.

[2] X. BAI and X. LIU, Extended Asynchronous SN P Systems for Solving Sentiment Clustering of Customer Reviews in E-commerce Websites, *WSEAS Transactions On Information Science And Applications.* 6, 2013(10), pp. 195–208.

[3] A.K. Jain, R. Duin, and J. Mao, Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 22, 2000(1), pp. 4–37.

[4] A. K. Jain, Data clustering: 50 years beyond kmeans, *Pattern Recognition Letters.* 31, 2010(8), pp. 651–666.

[5] R. Gelbard, O. Goldman, and I. Spiegler, Investigating diversity of clustering methods: An empirical comparison, *Data & Knowledge Engineering.* 63, 2007(1), pp. 155–166.

[6] J. Kleinberg, An impossibility theorem for clustering, *Advances in neural information processing systems.*, 2003, pp. 463–470.

[7] A. Strehl and J. Ghosh, Cluster ensembles–a knowledge reuse framework for combining mul-

tiple partitions, *The Journal of Machine Learning Research.* 3, 2003, pp. 583–617.

[8] A.K. Jain, R. Duin, and J. Mao, Clustering ensembles: Models of consensus and weak partitions, *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 27, 2005(12), pp. 1866– 1881.

[9] S. T. Hadjitodorov, L. I. Kuncheva, L. P. Todorova, Moderate diversity for better cluster ensembles, *Information Fusion.* 7, 2006(3), pp. 264–275.

[10] X. Z. Fern, W. Lin, Cluster ensemble selection, *Statistical Analysis and Data Mining.* 1, 2008(3), pp. 128–141.

[11] L. I. Kuncheva, S. T. Hadjitodorov, Using diversity in cluster ensembles, *Proceedings of Iinternational conference on Systems, man and cybernetics.*, 2004(2), pp. 1214–1219.

[12] J. Azimi, X. Fern, Adaptive cluster ensemble selection, *Proceedings of International Joint Conferences on Artificial Intelligence.*, 2009(9), pp. 992–997.

[13] J. Jia, X. Xiao, B. Liu, L. Jiao, Bagging-based spectral clustering ensemble selection, *Pattern Recognition Letters.* 32, 2011(10), pp. 1456–1467.

[14] Y. Hong, S. Kwong, H. Wang, Q. Ren, Resampling-based selective clustering ensembles, *Pattern recognition letters.* 30, 2009(3), pp. 298–305.

[15] A. L. Fred, A. K. Jain, Combining multiple clusterings using evidence accumulation, *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 27, 2005(6), pp. 835–850.

[16] G. Forestier, P. Gancarski, and C. Wemmert, Collaborative clustering with background knowledge, *Data & Knowledge Engineering.* 69, 2010(2), pp. 211-228.

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

[18] B. Minaei-Bidgoli, A. Topchy, and W. F. Punch, Ensembles of partitions via data resampling, *Proceedings of International Conference on Information Technology: Coding and Computing.*, 2004(2), pp. 188–192.

[19] B. Minaei-Bidgoli, H. Parvin, H. Alinejad-Rokny, H. Alizadeh, and W. F. Punch, Effects of resampling method and adaptation on clustering ensemble efficacy, *Artificial Intelligence Review.* 41, 2014(1), pp. 27–48.

[20] X. Z. Fern and C. E. Brodley, Random projection for high dimensional data clustering: A cluster ensemble approach, *ICML.*, 2003(3), pp. 186–193.

[21] S. Mimaroglu and E. Erdil, An efficient and scalable family of algorithms for combining clusterings, *Engineering Applications of Artificial Intelligence.* 26, 2013(10), pp. 2525–2539.

[22] H. W. Kuhn, The hungarian method for the assignment problem, *Naval research logistics quarterly.* 2, 1955(2), pp. 83–97.

[23] A. P. Topchy, A. K. Jain, and W. F. Punch, A mixture model for clustering ensembles, *Proceedings of International Conference on Data Mining, SIAM.*, 2004, pp. 379–390.

[24] G. Karypis and V. Kumar, Multilevel k-way partitioning scheme for irregular graphs, *Journal of Parallel and Distributed computing.* 48, 1998(1), pp. 96–129.

[25] X. Z. Fern and C. E. Brodley, Solving cluster ensemble problems by bipartite graph partitioning, *Proceedings of International conference on Machine learning, ACM.*, 2004, pp. 36–44.

[26] X.Wang, D. Han, and C. Han, Rough set based cluster ensemble selection, *Proceedings of International Conference on Information Fusion.*, 2013, pp. 438-444.

[27] F. Yang, X. Li, Q. Li, and T. Li, Exploring the diversity in cluster ensemble generation: Random sampling and random projection, *Expert Systems with Applications.* 41, 2014(10), pp. 4844-4866.

[28] X. Lu, Y. Yang, and H. Wang, Selective clustering ensemble based on covariance, *Proceedings of Multiple Classifier Systems, Springer.*, 2013, pp. 79–89.

[29] M. Naldi, A. Carvalho, and R. Campello, Cluster ensemble selection based on relative validity indexes, *Data Mining and Knowledge Discovery.* 27, 2013(2), pp. 259–289.

[30] H. Alizadeh, B. Minaei-Bidgoli, and H. Parvin, To improve the quality of cluster ensembles by selecting a subset of base clusters, *Journal of Experimental & Theoretical Artificial Intelligence.* 26, 2014(1), pp. 127–150.

[31] L. Hubert and P. Arabie, Comparing partitions, *Journal of classification.* 2, 1995(1), pp. 193–218 .