Int. J. Nonlinear Anal. Appl. 13 (2022) 2, 1151–1163
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.6392

The impact of diversity on clustering ensemble using Chi^2 criterion

Seyed Saeed Hamidi, Ebrahim Akbari*, Homayun Motameni

Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran

(Communicated by Mohammad Bagher Ghaemi)

Abstract

Clustering ensemble is a technique for improving clustering results' robustness and accuracy. Basically, this technique generates base clusterings and then combines them into a consensus solution whose quality is determined by the diversity of the base clusterings and the consensus function's performance. In order to improve the quality of consensus solutions, it is necessary to generate base clusterings with regard to quality and diversity. Novel techniques were employed in this study to generate diverse base clusterings for both low-dimensional and high-dimensional datasets, as well as new criteria to compute the diversity of base clusterings with respect to quality. The impacts of different levels of diversity on consensus functions were studied. The proposed methods generated diverse base clusterings, according to the findings of the experiments.

Keywords: Base Clusterings, Consensus clustering, Clustering ensemble, Quality, Diversity
2022 MSC: 60E05

1 Introduction

Data clustering, often known as cluster analysis, is a critical issue in data mining. [25]. The objective of data clustering is to understand the structure of unlabeled data by placing data samples into separate groups. Data samples that are more similar to each other are put into the same group (cluster); the ideal state is achieved when data in one group are completely different from data samples in other groups. In other words, clustering algorithms apply a similarity criterion to the process of cluster formation in a way to maximize intra-cluster similarity and, at the same time, minimize the inter-cluster similarity. Cluster analysis is applied into various fields such as bio-informatics, text mining, face detection, health fraud detection, image segmentation, etc. [57, 47, 49, 1, 22]. In the literature, many clustering algorithms have been presented [43, 15, 21], each of which is suitable for some datasets based on their shapes and distributions. However, no clustering algorithm outperforms others. This implies that each clustering algorithm has its unique set of advantages and disadvantages. For a specific dataset, different clustering algorithms obtain various results. Consequently, it is difficult to select the most proper clustering algorithm for a specific dataset. As a result, researchers have introduced the concept of clustering ensemble which refers to aggregating different clusterings into a single clustering result [2, 20, 5, 45].

*Corresponding author

Email addresses: s.hamidi@qaemiau.ac.ir (Seyed Saeed Hamidi), akbari@iausari.ac.ir (Ebrahim Akbari), motameni@iausari.ac.ir (Homayun Motameni)

Received: August 2021 *Accepted:* March 2022

Clustering ensemble or consensus clustering can produce a better and more robust and reliable clustering. First, a set of base clusterings, called ensemble members, is generated. The base clusterings are then aggregated using a consensus function to produce a consensus solution [23, 12]. Producing diverse base clusterings and selecting a proper consensus function yields better consensus solution. Consensus clustering can be applied to different domains such as business, biomedical, security, insurance, recommender system, etc. [31, 18, 33, 56]. The literature comprises a number of methods for generating base clusterings as well as many different consensus functions [54, 44, 37, 17]. The quality of the consensus solution depends on the diversity of base clusterings and the performance of the consensus function. The diversity of base clusterings is the degree to which the base clusterings differ. Higher diversity will result in better quality; on the other hand, between base clusterings, quality is better to be lower.

Different strategies for generating diversity are utilized in this research, and the outcomes are compared. Four methods are used to generate diversity in low-dimensional datasets, which employ the k-means algorithm with different initializations (different values of k, where k is the number of clusters) as well as three methods for high-dimensional datasets. In low-dimensional datasets, employing k-means algorithm with different k values is the best option to generate diverse clusterings [19, 6], whereas in high-dimensional datasets, applying feature selection methods produces base clusterings with higher diversity [54, 50].

To generate diverse base clusterings, most previously-conducted studies employed k-means with different initializations for low-dimensional datasets and selected a subset of features for high-dimensional datasets [29, 54, 50]. To obtain higher diversity, these two general approaches are examined. In the following, the major contributions of the present research are presented:

- Proposing a new diversity generation method for low-dimensional dataset
- Proposing a new diversity generation method for high-dimensional dataset
- Using Chi^2 criterion to compute diversity in terms of quality

The rest of the paper is organized as follows. Section 2 presents the related work. After that, the proposed method will be explained in section 3. Experiments and results will be presented in section 4. Finally, section 5 concludes the paper.

2 Related Work

Clustering ensemble is the process of aggregating multiple clusterings into a single consolidated clustering [48]. In fact, clustering ensemble is expected to acquire more consistent, reliable, and accurate clustering results compared to individual clustering algorithms.

Diversity generation is the first step in clustering ensemble algorithms. In general, a cluster ensemble with higher diversity performs better. Diversity refers to the degree of difference between base clusterings. Applying a suitable generation method is of high importance because the base clusterings obtained at this step affect the final solution. With respect to the literature, base clusterings generation is very important to success of the cluster ensemble solution [36, 29]. Different generation mechanisms have been proposed in the literature for creating diversity or ensemble members, which are discussed as follow.

- Different clustering algorithms: Base clusterings are generated by employing various clustering algorithms, which are called heterogeneous ensembles [8, 30, 34]. This mechanism assumes that different clustering algorithms analyze the dataset in different ways; thus, diverse base clusterings are obtained.
- Different initialization of a clustering algorithm: Base clusterings are produced by employing a single clustering algorithm by repeated runs of the algorithm with different initializations, which are known as homogeneous ensembles [7, 37]. The clustering algorithm most commonly used to produce base clusterings is the k-means algorithm due to its simplicity and low time-complexity. Many studies have used k-means algorithm with different inputs [29, 16, 51].
- Different subsets of instances: Random samples of instances are picked out to create base clusterings [44, 32, 40]. This mechanism is suitable for big data. Diverse base clusterings are produced by using several subsets of instances.

- Different subsets of features: Random samples of features are extracted to generate base clusterings [54, 50]. Differences in features lead to the production of different base clusterings. This mechanism is suitable to high-dimensional datasets.

The second and fourth mechanisms are mostly used for low-dimensional and high-dimensional datasets, respectively. Therefore, this paper focuses on these two mechanisms.

The aggregation phase is called clustering ensemble or consensus clustering, which is performed by applying a consensus function to ensemble members. The consensus function aggregates ensemble members into a consensus result. Co-association approaches, Graph based methods, Relabeling methods, and Feature-based methods are different types of consensus functions proposed in the literature. This paper focuses on the diversity generation phase. In the following, a review of recent and related work on diversity generation is presented.

Some researchers have used methods frequently used in the literature [19, 6, 41, 37]. These generally-used methods produce base clusterings by employing the k-means algorithm with different k values selected from $[2, \sqrt{n}]$, where n is the number of instances. Hamidi et al. [20] employed k-means algorithm with different k values to engender base clusterings. They demonstrated that, to obtain base clusterings with higher diversity, the value of k should be larger than the true number of clusters.

In [3] Alizadeh et al. applied the k-means algorithm with different values of k in the range $[k, 2k]$ to engendering base clusterings. Pividori et al. in [38] used two schemes to create base clusterings, and in both schemes, the k-means clustering algorithm was used; in Scheme 1, k was a fixed value, while in Scheme 2, k was selected randomly from the range $[2, 20]$. Nazari et al. [35] selected the value of k from the range $[2, 4k]$ and applied the k-means clustering algorithm with the chosen k as an initial parameter.

Iam-On & Boongoen [24] used two approaches for base clusterings generation and employed the k-means clustering algorithm. In the first approach, the value of k was fixed to $\sqrt{n}$, whereas in the second one, k was selected randomly from the range $[2, \sqrt{n}]$. Jia et al. [27] generated diverse base clusterings by employing three methods: 1) Random sampling in the Nyström approximation, 2) Random scaling parameter in spectral clustering, and 3) using the k-means algorithm with different initialization.

Fern et al. [11] studied three various diversity generations for high-dimensional datasets based on the random projection and principal component analysis. They showed that the random projection technique generates diverse base clusterings. Wu et al. used three different generative mechanisms to produce base clusterings: 1) running k-means algorithm with fixed k values and various cluster centers, 2) selecting different subsets of instances and then applying k-means to the new sets, and 3) selecting different subsets of features by employing ReliefF algorithm [28] and then applying k-means to these new datasets. Finally, they showed that the best mechanism is using the k-means algorithm with different initializations to generate diversity.

Yang et al. [52] used several traditional generation methods and suggested a new method to produce base clusterings. Traditional methods are 1) employing the k-means algorithm by randomly choosing k from the range $[2, 10]$, 2) selecting random subsets of features and then applying the k-means algorithm, and 3) selecting random subsets of instances based on the nearest centroid followed by applying the k-means algorithm. Their suggested method was selecting random subsets of instances based on the nearest neighbor followed by applying the k-means algorithm. The fourth generation method, which is suggested in the present paper, was empirically found capable of obtaining higher diversity.

3 Proposed method

In this section, the proposed method is described in detail. The notations used in this paper are presented in Table 1. At first, an ensemble is generated based on the second and fourth diversity generation mechanisms for low-dimensional and high-dimensional datasets, respectively. The k-means clustering algorihtm with different values of k was employed to generate ensemble members for low-dimensional datasets. In all suggested methods, the k-means clustering algorithm was used to generate diversity with different parameters. Algorithm 1 shows the procedure of diversity generation for low-dimensional datasets.

For high-dimensional datasets, different feature selection methods like random-based, wrapper-based, and filter-based methods are employed. By applying these methods, a subset of features is selected. Then, the k-means clustering algorithm is applied to the new dataset with minimal features subset to generate diversity.

Four methods are introduced for low-dimensional datasets ($M1$, $M2$, $M3$, and $M4$) and three methods are suggested for high-dimensional datasets ($M5$, $M6$, and $M7$), which are described as follow:

Table 1: Notations used in this paper

Symbol	Description
X	Set of instances
x_i	i-th instance of X
n	Number of instances
Π	Set of base clusterings
π_i	i-th clustering of Π
M	Number of base clusterings
C_i^j	j-th cluster of π_i
k_i	Number of clusters in π_i
Φ	Consensus function
π^*	Consensus clustering

- $M1$: The value of k was fixed to the value of true number of classes of the dataset (The true numbers of classes for all datasets are shown in the last column of Table 3).
- $M2$: The value of k was selected between $[2, \text{true number of classes}]$.
- $M3$: The value of k was chosen between $[2, \sqrt{n}]$.
- $M4$: The value of k was selected using Eq. 3.1.
- $M5$: Random-based feature selection (Algorithms 2).
- $M6$: Wrapper-based feature selection (Algorithms 3).
- $M7$: Filter-based feature selection (Algorithms 4).

Note that using the true number of classes in the $M1$ and $M2$ methods is only for experimental analysis because, in fact, the true number of classes is not reachable in clustering algorithms. The $M3$ and $M4$ methods are used to generate diversity.

$$k = \begin{cases} [\sqrt{n} - l : \sqrt{n}] & \text{if} \sqrt{n} > l, \\ [\frac{\sqrt{n}}{2} : \frac{\sqrt{n}}{2} + l] & \text{if} \sqrt{n} \leq l, \end{cases} \tag{3.1}$$

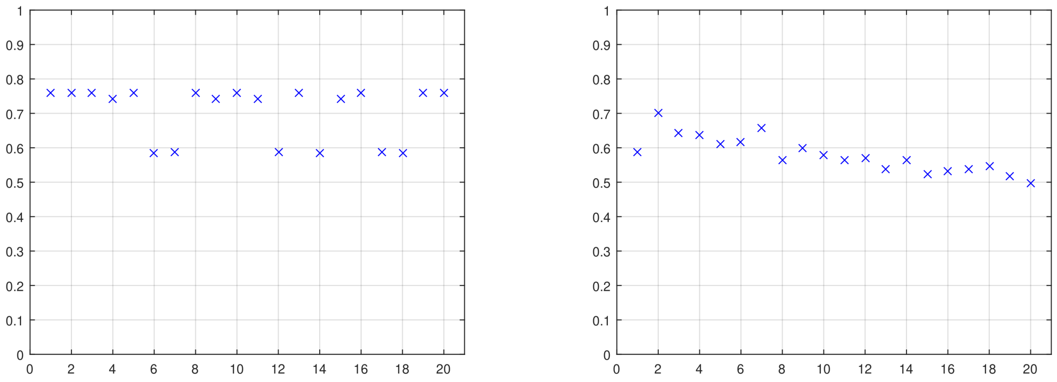
where l signifies the number of runs for different k, and n denotes the number of instances. The k-means algorithm is run h times for each value of k. Finally, $l * h$ base clusterings were generated.

Algorithm 1 : Diversity Generation for low-dimensional datasets using k-means ($M1$, $M2$, $M3$, and $M4$)

```
input: $D$, the dataset;   $M$, the number of ensemble members;
       $k_{min}$, the minimum value of $k$;   $k_{max}$, the maximum value of $k$
  //the value of $k_{min}$ and $k_{max}$ are changed based on diversity generation method ($M1$, $M2$, $M3$, or $M4$)
output: $\{BC_1, BC_2, \cdots, BC_M\}$, ensemble members
$k = k_{min}$
for $i = 1$ to $M$ do
    $BC_i$ = k-means ( $D$, $k$ )
    $k$ = new value from range $[k_{min}, k_{max}]$
end
```

Algorithm 1 illustrated the diversity generation method for low-dimensional datasets. In this study, the k-means algorithm with various values of k is employed to generate M base clusterings. The value of k starts at k_{min} and continues to k_{max}. The value of k_{min} and k_{max} vary based on diversity generation method ($M1$, $M2$, $M3$, and $M4$).

The quality of each ensemble member was calculated using the NMI criterion. The NMI values between $[0, 0.2]$ have lower quality and higher diversity, while the values in the range $[0.8, 1.0]$ have higher quality and lower diversity.



(a)

(b)

Figure 1: The quality of base clusterings in terms of NMI (a) first method, (b) second method.

Therefore, according to [19], lower and higher NMI values were removed. Then, the remaining NMI values were divided into ten equal parts. The number of values in each part was counted. To obtain higher diversity, the best case is that NMI values be distributed in all parts. In other words, the expected counts in each part is equal with others. To calculate the diversity and investigate the relation between quality and diversity, Chi^2 criterion was employed in this paper, which is defined as follow:

$$Chi^2 = \sum_{i=1}^{10} \frac{(O(i) - E(i))^2}{E(i)} \tag{3.2}$$

where $O(i) = n_{p_i}$, and $E(i)$ signifies the expected number of NMI values in the i-th part (p_i). The value of $E(i)$ in all parts is a constant value that is calculated as follows:

$$E(i) = \frac{M}{10}, \quad i = 1, 2, \cdots, 10 \tag{3.3}$$

The Chi-squared (Chi^2) is a number that determines the difference between the observed counts and expected ones. If the observed and expected counts are equal, the Chi^2 is equal to zero. Therefore, a low-value for Chi^2 is better. In comparison between the NMI values distributed in all parts and those distributed in some parts, the former case yields lower value of Chi^2.

Table 2: The values of Chi^2 criterion and consensus solution's quality corresponding to Fig. 1

	NMI value distribution	Chi^2	CSPA	MCLA	HGPA
First method	$n_{p_1} = 0, n_{p_2} = 0, n_{p_3} = 0, n_{p_4} = 0,$ $n_{p_5} = 0,\ n_{p_6} = 0,\ n_{p_7} = 6,$ $n_{p_8} = 0,\ n_{p_9} = 4,\ n_{p_{10}} = 10$	56	0.70	0.76	0.50
Second method	$n_{p_1} = 0, n_{p_2} = 0, n_{p_3} = 0, n_{p_4} = 0,$ $n_{p_5} = 1,\ n_{p_6} = 6,\ n_{p_7} = 9,$ $n_{p_8} = 3,\ n_{p_9} = 1,\ n_{p_{10}} = 0$	44	0.88	0.85	0.90

As an example, assume that $X = \{x_1, x_2, \cdots, x_{10}\}$ denotes a toy dataset with 10 objects, and $\Pi = \{\pi_1, \pi_2, \cdots, \pi_{20}\}$ is an ensemble with 20 base clusterings on dataset X. The k-means clustering algorithm is employed to generate ensemble members with two methods. In the first method, the value of k in the k-means algorithm is fixed, whereas in the second, the value of k is increased with incremental step of 1 in each run. The value of NMI criterion is computed for each base clustering according to the dataset class labels. Fig. 1 demonstrates the NMI values of the base clusterings for both methods.

After calculating the NMI values of the ensemble members, the following steps are taken into action:

1. Dividing the NMI values into 10 equal parts $\{p_1, p_2, \cdots, p_{10}\}$
2. Counting the number of NMI values in each part $\{n_{p_1}, n_{p_2}, \cdots, n_{p_{10}}\}$
3. Calculating the Chi^2 measure

The Chi^2 measure is used to compute the diversity between the base clusterings. The value of this measure is calculated based on Eqs. (3.2).

The calculated values for the above-mentioned example as well as the best case and worst case are shown in Table 2. When the values are evenly distributed in all parts, the best-case occurs. In contrast, when all the values are in one part, the worst-case occurs. As shown in Table 2, the NMI values distribution in the second method is performed better than the first one. On the other hand, the second method values are more diverse than those of the first method. The lower values of Chi^2 denotes better diversity, which can be realized with respect to Table 2.

The diversity generation method for high-dimensional datasets based on the random feature selection is depicted in Algorithm 2. First, a subset of dataset features is randomly selected. Then, the k-means algorithm is employed to generate a clustering with minimal features subset. The feature selection phase and the use of the k-means algorithm are repeated for M times to generate base clusterings.

Algorithm 2 : Diversity Generation for high-dimensional datasets based on the random-based feature selction ($M5$)

```
input: $D$, the dataset;  $M$, the number of ensemble members; $k$, the number of clusters
output: $\{BC_1, BC_2, \cdots, BC_M\}$, ensemble members
1. for $i = 1$ to $M$ do
2.     $F$ = a subset of features of D selected randomly (50 features)
3.     $BC_i$ = k-means ( $D$, $F$,$k$ )
4. end
```

The wrapper-based feature selection was employed in Algorithm 3 to produce ensemble members. First, a subset of dataset features was selected using the Butterfly Optimization Algorithm [53]. Then, the k-means algorithm assumed the selected features subset as input for generating a clustering. The feature selection phase and employing the k-means algorithm were repeated M times to generate base clusterings.

Algorithm 3 : Diversity Generation for high-dimensional datasets based on the wrapper-based feature selction ($M6$)

```
input: $D$, the dataset;  $M$, the number of ensemble members; $k$, the number of clusters
output: $\{BC_1, BC_2, \cdots, BC_M\}$, ensemble members
1. for $i = 1$ to $M$ do
2.     $F$ = a subset of features of D selected using BOA algorithm
   //Butterfly Optimization Algorithm (BOA) is an algorithm for wrapper-based feature selection [53]
3.     $BC_i$ = k-means ( $D$, $F$,$k$ )
4. end
```

Algorithm 4 : Diversity Generation for high-dimensional datasets based on the filter-based feature selction ($M7$)

```
input: $D$, the dataset;  $M$, the number of ensemble members; $k$, the number of clusters
output: $\{BC_1, BC_2, \cdots, BC_M\}$, ensemble members
1. $F$ = rank features of D using reliefF algorithm
   //reliefF algorithm is an algorithm for filter-based feature selection [42]
2. for $i = 1$ to $M$ do
3.     $BC_i$ = k-means ( $D$, $[f_1, f_{i+1}]$, $k$ )
4. end
```

Algorithm 4 illustrated the diversity generation method for high-dimensional datasets based on the filter-based feature selection. First, the dataset features were ranked using the reliefF algorithm [42]. Then, the k-means algorithm was repeated M times with different features subset to generate ensemble members. Assume that $F = \{f_1, f_2, \cdots, f_d\}$ denotes the ranked features of a dataset. The first selected features subset is $\{f_1, f_2\}$, the next subset is $\{f_1, f_2, f_3\}$, and the final subset is $\{f_1, f_2, \cdots, f_{M+1}\}$.

4 Experimental analysis

In this section, some experiments were conducted to evaluate the results. The k-means clustering algorithm was employed to generate base clusterings, and the NMI criterion was applied to computing the quality of clusterings. The value of NMI is between $[0, 1]$. The value of one signifies complete correspondence between two clusterings, and the zero denotes dissimilarity between them. The class label information of datasets was used only to compute the quality, and was not used in the process of consensus clustering. Note that the class label information was removed from datasets before the clustering process. All experiments were run in MATLAB R2015a 64-bit environment on Windows 10 Pro 20H2 64-bit, Intel Core i5-6300U CPU, and 8 GB of RAM workstation.

Table 3: Datasets used in the experiments

	Dataset	Instances (n)	Features (d)	Classes (k)
Artificial	Jain	373	2	2
	Path based	300	2	3
	Compound	399	2	6
	Flame	240	2	2
Low-dimensional	Iris	150	4	3
	Wine	178	13	3
	Breast tissue	106	9	6
	Seeds	210	7	3
High-dimensional	GLI-85	85	22283	2
	Prostate-GE	102	5966	2
	Leukemia	72	7129	2
	PengLeukEW	72	7070	2

4.1 Datasets

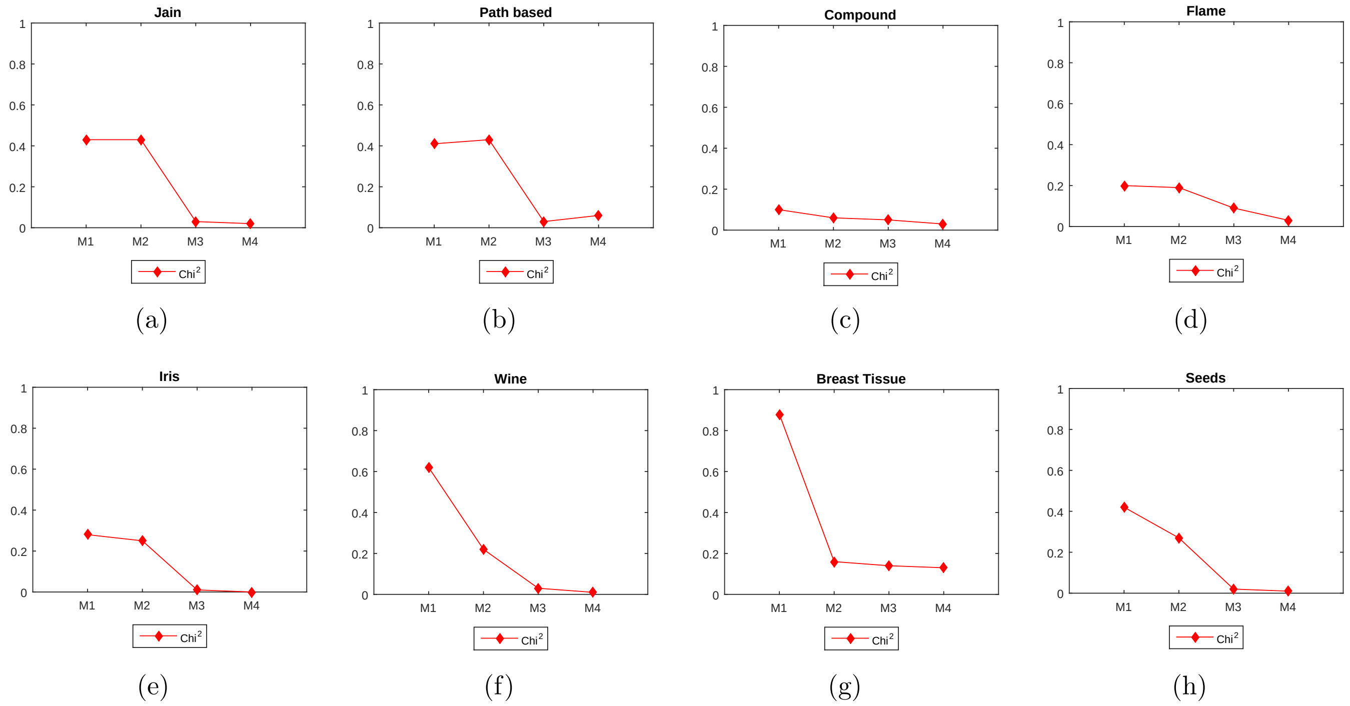
Experiments were executed on four artificial datasets, four real datasets, and four microarray datasets. The details of these datasets are briefly summarized in Table 3 in which n denotes the number of instances, d signifies the number of features, and k is the number of classes. The four artificial datasets came from the literature [26, 9, 55, 14]. The four real datasets were extracted from the UCI machine learning repository [10], while the four microarray datasets were extracted from the literature [4, 39, 46, 13].

4.2 Test results

The first experiment was conducted on the artificial and low-dimensional datasets, while the second experiment was conducted on the high-dimensional datasets. In the first experiment, four methods ($M1$, $M2$, $M3$, and $M4$) were used to generate diversity. For each dataset, $M = 50$ base clusterings were generated, and for each generation method, the algorithm was run 100 times, and the average of the results was recorded. For each method, the diversity was calculated using the introduced criterion (Chi^2), and the results were shown in Fig. 2. The CSPA, MCLA, and HGPA consensus functions were employed to obtain consensus results, which are shown in Fig. 3.

As illustrated in Fig. 2, the diversity values which were obtained by the $M4$ method are better than the values of other methods (Chi^2 values for $M4$ method are lower). It means that the $M4$ method generated better diversity in comparison with the $M1$, $M2$, and $M3$ methods. This can be understood based on the output of the consensus clustering algorithms (Fig. 3). As demonstrated in Fig. 3, the CSPA, HGPA, and MCLA consensus functions obtained higher quality results by the $M4$ method. As a result, the $M4$ method was found the most proper choice to generate diversity in low-dimensional datasets.

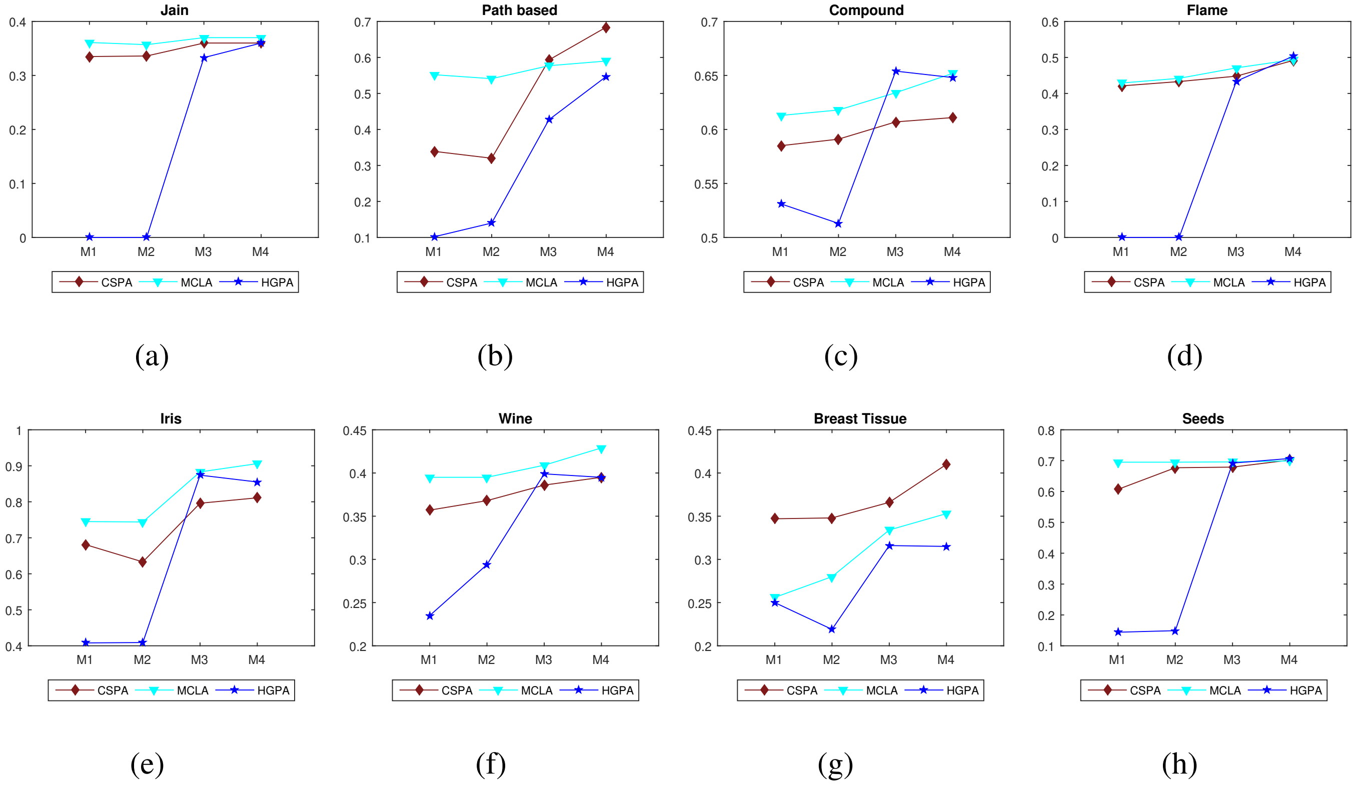
In the third experiment, three methods ($M5$, $M6$, and $M7$) were used to generate diversity for high-dimensional datasets. This experiment was conducted in two phases. The values of k in the k-means algorithm were fixed at 2



(a) (b) (c) (d)

(e) (f) (g) (h)

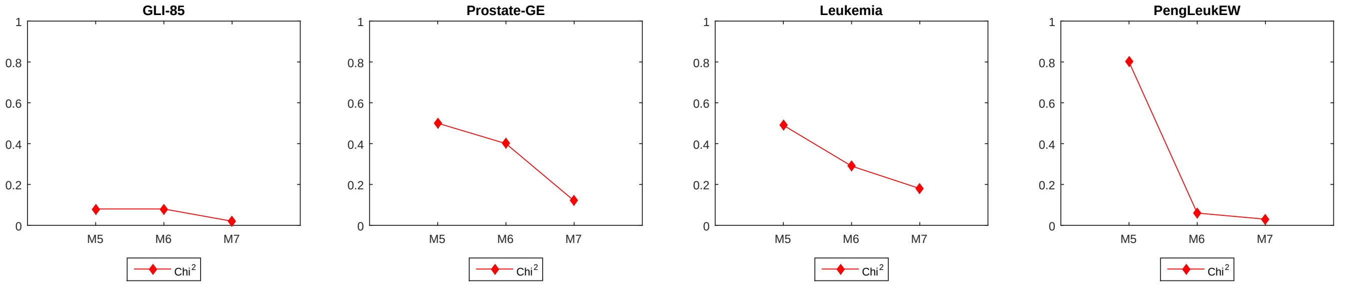
Figure 2: Diversity of base clusterings that is calculated using suggested criterion for artificial and low-dimensional datasets.



(a) (b) (c) (d)

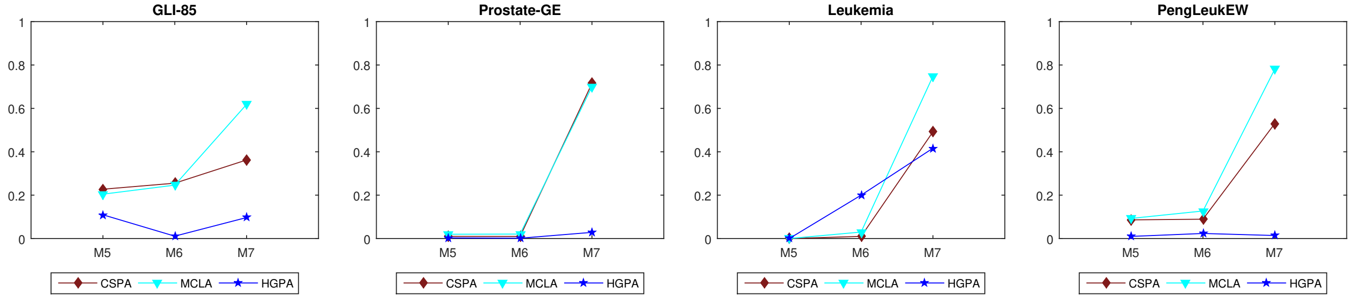
(e) (f) (g) (h)

Figure 3: Quality of the consensus solutions corresponding to Fig. 2.



(a) (b) (c) (d)

Figure 4: Diversity of base clusterings (employing k-means with $k = 2$), which is calculated using the suggested criterion for high-dimensional datasets.



(a) (b) (c) (d)

Figure 5: Quality of the consensus solutions corresponding to Fig. 4.

and 5 in the first and the second phase, respectively. For each dataset, $M = 50$ base clusterings were produced; the computed diversities are illustrated in Fig. 4 for $k = 2$ and Fig. 6 for $k = 5$. In addition, the qualities of consensus solutions were shown in Fig. 5 and Fig. 7.

As illustrated in Figs. 4 and 6, the diversity values obtained by the $M7$ method are better than the values of the $M5$ and $M6$ methods. Additionally, the quality values obtained by the consensus functions for the $M7$ method are higher than others shown in Figs. 5 and 7. Consequently, the $M7$ method was found the most proper choice for producing base clusterings in high-dimensional datasets.

5 Conclusion and future work

New diversity generating methods for low-dimensional and high-dimensional datasets were proposed in this study, which use the k-means clustering algorithm and the feature selection method, respectively. The diversity values were determined using new criterion. Diversity's effect on consensus functions was explored. The findings of the experiments conducted on four artificial, four low-dimensional, and four high-dimensional datasets revealed that for low-dimensional datasets, utilizing the k-means clustering algorithm with a value of k greater than the true number of classes is a good choice. Furthermore, for high-dimensional datasets, the employment of filter-based feature selection methods combined with the k-means clustering algorithm generates diverse base clusterings.

References

[1] N.M. Abdolrazzagh and M. Kherad, *Improved birch clustering by chemical reaction optimization algorithm to health fraud detection*, Iran. J. Electric. Comput. Engin. **17** (2019), no. 2, 153–160.

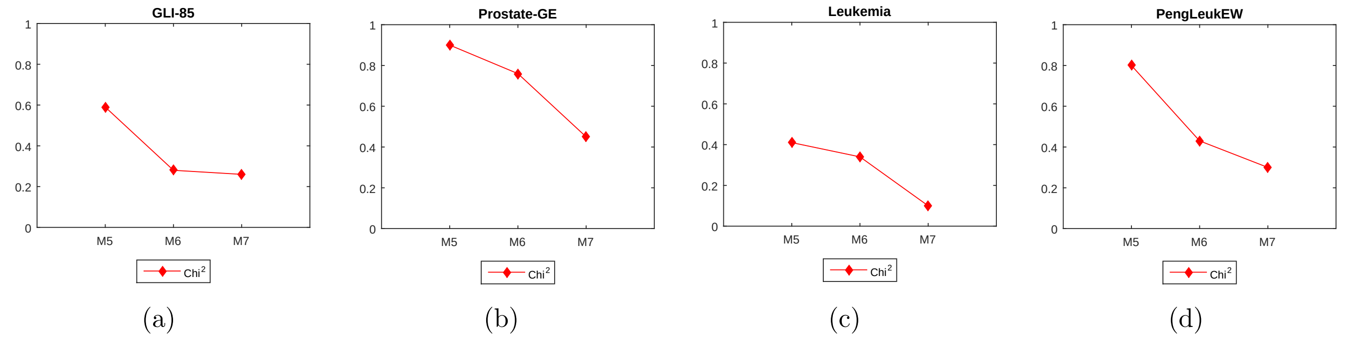


Figure 6: Diversity of base clusterings (employing k-means with $k = 5$), which is calculated using the suggested criterion for high-dimensional datasets.

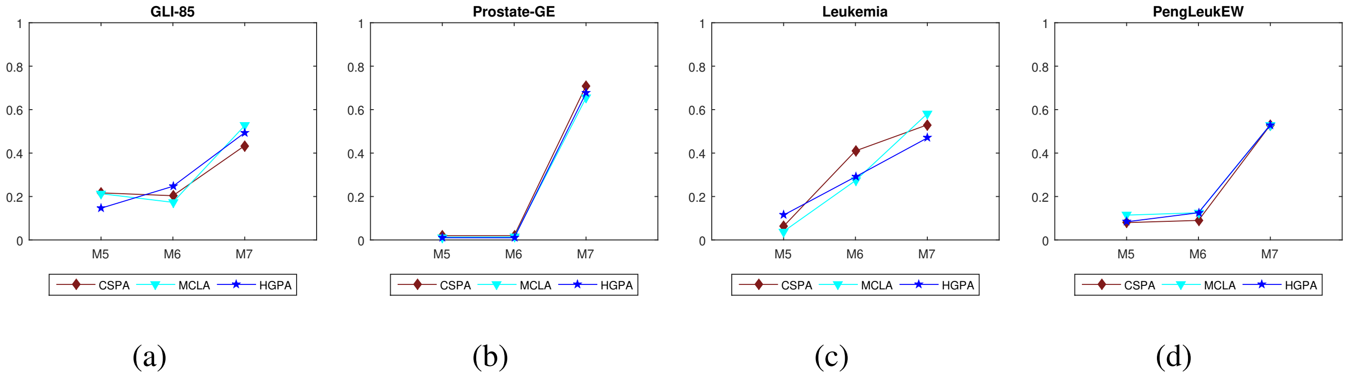


Figure 7: Quality of the consensus solutions corresponding to Fig. 6.

[2] E. Akbari, H. Mohamed Dahlan, R. Ibrahim, and H. Alizadeh, *Hierarchical cluster ensemble selection*, Engin. Appl. Artif. Intell. **39** (2015), 146–156.

[3] H. Alizadeh, B. Minaei-Bidgoli, and H. Parvin, *Cluster ensemble selection based on a new cluster stability measure*, Intell. Data Anal. **18** (2014), no. 3, 389–408.

[4] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proc. Nat. Acad. Sci. **96** (1999), no. 12, 6745–6750.

[5] T. Alqurashi and W. Wang, *Clustering ensemble method*, Int. J. Machine Learn. Cyber. **10** (2019), no. 6, 1227–1246.

[6] J. Azimi and X. Fern, *Adaptive cluster ensemble selection*, Proc. 21st Int. Jont Conf. Artif. Intell., 2009, pp. 992–997.

[7] A. Bagherinia, B. Minaei-Bidgoli, M. Hosseinzadeh, and H. Parvin, *Reliability-based fuzzy clustering ensemble*, Fuzzy Sets Syst. **413** (2021), 1–28.

[8] V. Berikov, *Weighted ensemble of algorithms for complex data clustering*, Pattern Recog. Lett. **38** (2014), 99–106.

[9] H. Chang and D.-Y. Yeung, *Robust path-based spectral clustering*, Pattern Recog. **41** (2008), no. 1, 191–203.

[10] D. Dua and C. Graff, *Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california*, School Inf. Comput. Sci. **25** (2019), 27.

[11] X.Z. Fern and C.E. Brodley, *Cluster ensembles for high dimensional clustering: An empirical study*, (2006).

[12] A. Fiori, A. Mignone, and G. Rospo, *Decoclu: Density consensus clustering approach for public transport data*, Inf. Sci. **328** (2016), 378–388.

[13] W.A. Freije, F.E. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L.M. Liau, P.S. Mischel, and S.F. Nelson, *Gene expression profiling of gliomas strongly predicts survival*, Cancer Res. **64** (2004), no. 18, 6503–6510.

[14] L. Fu and E. Medico, *Flame, a novel fuzzy clustering method for the analysis of dna microarray data*, BMC Bioinf. **8** (2007), no. 1, 1–15.

[15] G. Gan, C. Ma, and J. Wu, *Data clustering: Theory, algorithms, and applications*, SIAM, 2020.

[16] A. Gionis, H. Mannila, and P. Tsaparas, *Clustering aggregation*, Acm Trans. Knowledge Discov. Data **1** (2007), no. 1, 4–es.

[17] K. Golalipour, E. Akbari, S.S. Hamidi, M. Lee, and R. Enayatifar, *From clustering to clustering ensemble selection: A review*, Engin. Appl. Artif. Intell. **104** (2021), 104388.

[18] J. Guan, R.-Y. Li, and J. Wang, *Grace: A graph-based cluster ensemble approach for single-cell rna-seq data clustering*, IEEE Access **8** (2020), 166730–166741.

[19] S.T. Hadjitodorov, L.I. Kuncheva, and L.P. Todorova, *Moderate diversity for better cluster ensembles*, Inf. Fusion **7** (2006), no. 3, 264–275.

[20] S.S. Hamidi, E. Akbari, and H. Motameni, *Consensus clustering algorithm based on the automatic partitioning similarity graph*, Data Knowledge Engin. **124** (2019), 101754.

[21] E. Heidari, H. Motameni, and A. Movaghar, *A meta-heuristic clustering method to reduce energy consumption in internet of things*, Int. J. Nonlinear Anal. Appl. **12** (2021), no. 1, 45–58.

[22] H. Hooda and O.P. Verma, *Fuzzy clustering using gravitational search algorithm for brain image segmentation*, Multimedia Tools Appl. (2022), 1–20.

[23] D. Huang, J. Lai, and C.-D. Wang, *Ensemble clustering using factor graph*, Pattern Recog. **50** (2016), 131–142.

[24] N. Iam-On and T. Boongoen, *Diversity-driven generation of link-based cluster ensemble and application to data classification*, Expert Syst. Appl. **42** (2015), no. 21, 8259–8273.

[25] A.K. Jain, *Data clustering: 50 years beyond k-means*, Pattern Recog. Lett. **31** (2010), no. 8, 651–666.

[26] A.K. Jain and M.H.C. Law, *Data clustering: A user's dilemma*, Int. Conf. Pattern Recog. Machine Intell.,

Springer, 2005, pp. 1–10.

[27] J. Jia, X. Xiao, B. Liu, and L. Jiao, *Bagging-based spectral clustering ensemble selection*, Pattern Recog. Lett. **32** (2011), no. 10, 1456–1467.

[28] I. Kononenko, *Estimating attributes: Analysis and extensions of relief*, Eur. Conf. Machine Learn., Springer, 1994, pp. 171–182.

[29] G. Li, M.R. Mahmoudi, S.N. Qasem, B.A. Tuan, and K.-H. Pho, *Cluster ensemble of valid small clusters*, J. Intell. Fuzzy Syst. **39** (2020), no. 1, 525–542.

[30] X. Li, Y. Zhang, H. Cheng, F. Zhou, and B. Yin, *An unsupervised ensemble clustering approach for the analysis of student behavioral patterns*, IEEE Access **9** (2021), 7076–7091.

[31] B.P. Marques and C.F. Alves, *Using clustering ensemble to identify banking business models*, Intell. Syst. Account. Finance Manag. **27** (2020), no. 2, 66–94.

[32] B. Minaei-Bidgoli, H. Parvin, H. Alinejad-Rokny, H. Alizadeh, and W.F. Punch, *Effects of resampling method and adaptation on clustering ensemble efficacy*, Artif. Intell. Rev. **41** (2014), no. 1, 27–48.

[33] E. Mueller, J.S.O. Sandoval, S. Mudigonda, and M. Elliott, *A cluster-based machine learning ensemble approach for geospatial data: Estimation of health insurance status in missouri*, ISPRS Int. J. Geo-Inf. **8** (2019), no. 1, 13.

[34] F. Najafi, H. Parvin, K. Mirzaie, S. Nejatian, and V. Rezaie, *Dependability-based cluster weighting in clustering ensemble*, Statist. Anal. Data Min. ASA Data Sci. J. **13** (2020), no. 2, 151–164.

[35] A. Nazari, A. Dehghan, S. Nejatian, V. Rezaie, and H. Parvin, *A comprehensive study of clustering ensemble weighting based on cluster quality and diversity*, Pattern Anal. Appl. **22** (2019), no. 1, 133–145.

[36] H. Niu, N. Khozouie, H. Parvin, H. Alinejad-Rokny, A. Beheshti, and M.R. Mahmoudi, *An ensemble of locally reliable cluster solutions*, Appl. Sci. **10** (2020), no. 5, 1891.

[37] P. Panwong, T. Boongoen, and N. Iam-On, *Improving consensus clustering with noise-induced ensemble generation*, Expert Syst. Appl. **146** (2020), 113138.

[38] M. Pividori, G. Stegmayer, and D.H. Milone, *Diversity control for improving the analysis of consensus clustering*, Information Sciences **361** (2016), 120–134.

[39] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, and C. Lau, *Prediction of central nervous system embryonal tumour outcome based on gene expression*, Nature **415** (2002), no. 6870, 436–442.

[40] E. Rashedi and A. Mirzaei, *A hierarchical clusterer ensemble method based on boosting theory*, Knowledge-Based Syst. **45** (2013), 83–93.

[41] F. Rashidi, S. Nejatian, H. Parvin, and V. Rezaie, *Diversity based cluster weighting in cluster ensemble: an information theory approach*, Artif. Intell. Rev. **52** (2019), no. 2, 1341–1368.

[42] M. Robnik-Šikonja and I. Kononenko, *Theoretical and empirical analysis of relieff and rrelieff*, Machine Learn. **53** (2003), no. 1, 23–69.

[43] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M.J. Er, W. Ding, and C.-T. Lin, *A review of clustering techniques and developments*, Neurocomput. **267** (2017), 664–681.

[44] R.I. Seetan, J. Bible, M. Karavias, W. Seitan, and S. Thangiah, *Consensus clustering: A resampling-based method for building radiation hybrid maps*, 15th IEEE Int. Conf. Machine Learn. Appl. (ICMLA), IEEE, 2016, pp. 240–245.

[45] Y. Shi, Z. Yu, W. Cao, C.L.P. Chen, H.-S. Wong, and G. Han, *Fast and effective active clustering ensemble based on density peak*, IEEE Trans. Neural Networks Learn. Syst. **32** (2020), no. 8, 3593–3607.

[46] A. Spira, J.E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.-M. Dumas, P. Calner, and P. Sebastiani, *Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer*, Nature Med. **13** (2007), no. 3, 361–366.

[47] R. Srivastava, P. Singh, K.P.S. Rana, and V. Kumar, *A topic modeled unsupervised approach to single document*

extractive text summarization, Knowledge-Based Syst. (2022), 108636.

[48] A. Strehl and J. Ghosh, *Cluster ensembles—a knowledge reuse framework for combining multiple partitions*, J. Machine Learn. Res. **3** (2003), no. Dec, 583–617.

[49] D. Sun, K. Yang, and Z. Ding, *Confidence-based simple graph convolutional networks for face clustering*, IEEE Access **10** (2022), 6459–6469.

[50] A. Topchy, A.K. Jain, and W. Punch, *A mixture model for clustering ensembles*, Proc. 2004 SIAM Int. Conf. Data Mining, SIAM, 2004, pp. 379–390.

[51] Z. Wang, H. Parvin, S.N. Qasem, B.A. Tuan, and K.-H. Pho, *Cluster ensemble selection using balanced normalized mutual information*, J. Intell. Fuzzy Syst. **39** (2020), no. 3, 3033–3055.

[52] F. Yang, X. Li, Q. Li, and T. Li, *Exploring the diversity in cluster ensemble generation: Random sampling and random projection*, Expert Syst. Appl. **41** (2014), no. 10, 4844–4866.

[53] X.-S. Yang, *Nature-inspired algorithms and applied optimization*, vol. 744, Springer, 2017.

[54] M. Ye, W. Liu, J. Wei, and X. Hu, *Fuzzy c-means and cluster ensemble with random projection for big data clustering*, Math. Prob. Engin. **2016** (2016).

[55] C.T. Zahn, *Graph-theoretical methods for detecting and describing gestalt clusters*, IEEE Trans. Comput. **100** (1971), no. 1, 68–86.

[56] H. Zarzour, F. Maazouzi, M. Al-Zinati, Y. Jararweh, and T. Baker, *An efficient recommender system based on collaborative filtering recommendation and cluster ensemble*, Eighth Int. Conf. Soc. Network Anal. Manag. Secur. (SNAMS), IEEE, 2021, pp. 01–06.

[57] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, *Sequence clustering in bioinformatics: an empirical study*, Brief. Bioinf. **21** (2020), no. 1, 1–10.