International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 03 - Issue 02 || February 2017 || PP. 51-64

Using local optimization algorithms in ensemble clustering with maximize diversity

SeyedAhad Zolfagharifar[1], FaramarzKaramizadeh[2] , Hamid Parvin
[1]*Department of Computer Engineering, Yasouj Branch, Islamic Azad University, Yasouj, Iran*
[2]*,Department of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran*
[3]*Department of Computer Engineering, Yasouj Branch, Islamic Azad University, Yasouj, Iran*
Correspondence: FaramarzKaramizadeh, Department of Electrical and Computer Engineering, Shiraz University,

Abstract: Clustering data means partition samples in clusters which are similar, so that sample of each cluster has maximum similarity with each other and maximum distance with other sample. Unsupervised clustering is due to the choice of a particular algorithm for clustering an anonymous collection is risky and usually failed.Because of the complexity of the issue and lack of basic clustering methods, today the majority of studies directed towards hybrid clustering methods. Dispersion primary result is one of the most important factors that can effect on the quality of the final results.Also, quality of the early results is another factor which is effective on| quality of the results of the combination. Both factors have been considered in recent studies hybrid clustering. Here, proposed a new framework to improve the efficiency of hybrid clustering that is based on the use of a subset of primary clusters.The selection of these subsidiaries plays a crucial role in the performance of the Assembly. The selection is done with help of intelligent methods. The main ideas proposed methods for selecting a subset of the clusters; the clusters are stable with intelligent search algorithms.To evaluate the clusters, use the stability criterion based on mutual information. Finally, we collect the selected cluster to the final mix with help of several ways. Experimental results on several standard datasets show that the proposed methods can effectively improve the perfect combination method.

Key words: Hybrid Clustering, Local Optimization, Diversity, Evolutionary Algorithms, Correlation Matrix, Diversity.

Introduction

Clustering is the branches of unsupervised learning and is automated process, during which the samples are divided into clusters that its members are similar to each other and with other existed sample in other cluster have a maximum distance (Azimi, 2007, Alizadeh et al., 2014).

In general,in hybrid clustering two important issues should be considered. first a diversity of different clustering algorithms so that each of this clustering algorithms emphasis oncharacteristics of data, and the second component algorithms is the results for final clusters. In relation to the first issue, can be used four different ways as follows:

1- Use different clustering algorithms (sterile and gas, 2002).

2. Change the initial values and other parameters selected clustering algorithm (Fred and Jane, 2002).

3. Select some of characteristics or create new characteristic (Azimi, 2007, Parvin. Minai, 2015)

4. The classification of original data to different and distinct sub-collections (Taichi et al., 2003; Ferd and Lorenzo, 2008; Aid and the Kamel, 2008; Minai et al., 2002).

For hybrid algorithm result for final clusters there have been extensive studies and different articles have been printed (Alizadeh et al., 2013; Parvin et al., 2013 Barthelme and Leclerc, 1995, Fern and Bradley, 2003).

Dispersion in rankings information means that if a classifier has erroneous in some cases, then we search for another classifier that has multiple instances of errors in classification errors in the first set of learning to achieve a better outcome.

Absence of sets of learning Strips off this potential from information clustering methods and we have tried to enter a discussion of the concept of clustering our information (Parvin et al., 2013; Parvin et al., 2015; Dodiot and Freddliand, 2003; Fischer and Bohman, 2003).The concept dispersion has been used widely in recent years research(Fred and Jane, 2005 Fred and Jane, 2006; Kenchva and Whitaker, 2003; Kenchva and Hajitodorof, (2004).in the field of hybrid clustering a lot of work has been done that in following points to a few of them. Methods of hybrid clustering try to combine different partitions produced from a basic clustering methods, produce a solid partition of data (Easterly and Ghosh, 2002; Alizadeh et al., 2014; Parvin and Minai, 2015).In the most recent studies, all applications presence with an equal weight in the final composition, and all

existed clusters in all partitions with equal weight participate in the final composition. Azimi (2007), from the concept for the intelligent the clusters are combined.Is methods of Alizadeh et al. (2014), whichthere for all data collection first arranged based on the stability and then selected and 33% more stable.Baumgartner (2000), a method that indicating resampling based on fuzzy clustering to investigate the validity (Baumgartner et al., 2000). (Breckenridge,1989); Shamshiri and Tishbi, 2007; Roth et al, 2002 (a).The initial ideas for validating clusters using vector resampling (Lapointeand Lgndr, 1991) provided later in (Baumgartner et al., 2000; Roth et al., 2002 b) is complete. Sha and Das (2015) have developed a method to determine the number of clusters that automatically weighting the features.

1.1 methods for searching revelation

In this section, two algorithms used based on evolutionary methods in this paper is an overview study.1-1-1 GA

This method is imitation of evolution using computer algorithms. The most fundamental principle of evolution is heredity. GA innovator John Holland in the seventies inspired characteristics of evolutionary theory, invented the algorithm search algorithm of the same principles that nature on the evolution of gene symbols do Para clinic (Melanie, 1999).

2. A review of literature of the subject

Typically most of the hybrid clusters use k-means algorithm for their primary clustering (Fred and Jane, 2002; Fred and Jane, 2003; Fredliand and Dodiot, 2001). But the proposed methods is shown that according to behavior of each data set sometimes a specific clustering method found that gives better accuracy than k-means for some data sets (Parvin and Minai, 2015).But k-means algorithm due to its simplicity and fit ability in clustering always has been studied as the first choice of ensemble hybrid clustering.Another way to increase dispersion is changing on primary clustering parameters. For example, changing the number of clusters in K-means algorithm or changing the Seed Points of algorithm prototypes has effect in increasing the dispersion in clustering and plays an important role in clustering information.In Figure 1 effect of Seed Points in the final clustering are clearly visible. In Figure 1 first the distribution of samples is shown on the left and then shown implement the results of three different algorithms start with 3 different examples.

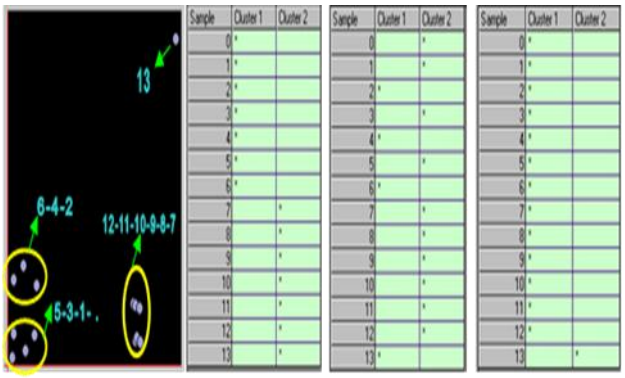


Figure 1: Examples of the results of the k-means algorithm figures respectively from left to right are: (1) display space 14 scattered in space. 2. The results of the prototypes 1 and 8. 3 results with Seed Points, 2 and 3. 4. The results obtained with the Seed Points1 and 13.

3- Suggested method

Is assumed X is the data collection includes N examples:

(2)

$$X = \{x_1, x_2, \dots, x_N\}$$

If π_j (x_i), the output corresponding to the j-th clustering algorithm is based on $x_i$example, we have:

(3)

$$x_i \to \{\pi_1(x_i), \pi_2(x_i), \dots, \pi_H(x_i)\}$$

$\pi_j(x_i)$ ،$j = 1, 2, \dots, H$ ، $i = 1, 2, \dots, N$Output from the implementation of k-means on $x_i$is smart. New feature space with H dimension. Each corresponding to one dimension of space-based clustering algorithm will be a new feature. If the set X consists of N with m attributes, created a new series, X ', a set of N samples will be with H characteristic.

(4)

$$X \equiv \{x_1, x_2, \dots, x_N\} \qquad X = \{x'_1, x'_2, \dots, x'_N\}$$
$$x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\} \implies x_i^{'} = \{x'_{i1}, x'_{i2}, \dots, x'_{iH}\}$$
$$i = 1, 2, \dots, N \qquad i = 1, 2, \dots, N$$

A property value in the new space is performed by using a mechanism that will be introduced in the next section. After the samples were create in the new feature space, the final clustering simple is done by using one of the clustering based methods.

3.1adaptive function

Collecting cluster primary cluster and achieving the final result is one of the most important steps a of hybrid clustering. In following we introduce several famous and new methods in this field.

1-Approach based could graph

In approach based could graph, we first change hybrid clustering toGraph partitioning and then we solve it with help of Graph partitioning algorithms,clusters displayed with could of a graph. (Parvin, and Minai, 2015; Easterly, and Ghosh, 2002).A CSPA the CSPA data points in the feature space to feature space mapped hyper graphs solidarity. Then, an algorithm METIS used to at least for hyper graphs like the newly spaced data points.As before, this method assumes that the more data points in a cluster on the primary partition, the more likely that the data points are inherently belong to a cluster.

A. CSPA is the easiest Revelation. The computational complexity of O (kN ^ 2 M), where k is the number of clusters, N is the number of points Data and M number of areas. Two-dimensional methods have less computational complexity.
B. HGPA
C. HGPA algorithm assumes that the top-of stems and clusters of data points that have come out of the primary partition of it.
D. MCLA

MCL algorithm first primary partition to partition from the cluster and then from a polling-based mechanism uses to generate partitions Assembly.

-2Methods of voting

This method is same with majority voting method. In this case, the cluster of each sample is determined by majority vote. (Levine and domaini, 2001); (Minai et al., 2004; Aid and Kamel, 2008)

3. The correlation matrix

The first algorithm that is basic and usable is k-means algorithm. In the first step, k-means algorithm applies in $X = \{X_1, X_2, \dots X_{B1}\}$so we can by using produced P_i obtain the following correlation matrix.

(5)

$$Co - association(xy) = \sum_{i=1}^{B1} \lambda(P_i(x), P_i(y))$$

3.2 Details related to the proposed method

In this section we have offer approach from the heart of previous problems that both optimize the dispersion and consider the accuracy. To do this, we first set (consensus) of the initial clusters that name RefSet. RefSet consensus with size | RefSet | that Shows the number of elements in this collection. It is noteworthy that RefSet_i represents the i- th member of this consensus. Then produce main consensus called for a consensus or Ensemble. It is noteworthy that Ensemble represents the i- th member of this consensus. For each of the Ensemble that i is changed to B, Calculate the stability of the calculation. Average similar partition of the partition Ensemble is stability in the reference collection. The levels of similarity of the two partitions calculate with standards famous equation Fisher. Now we have to say that how standard Fisher calculation is made. The criteria in this article are intended to evaluate a partition, or Fisher's exact test (F-Measure) respectively.

(7)

$$FM(P, L) = \max_{\tau} \sum_{i=1}^{K_P} \frac{2 \times N_i^P \times \left(\frac{N_{i\tau(i)}^{PL}}{N_i^P} \times \frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}}\right)}{N \times \left(\frac{N_{i\tau(i)}^{PL}}{N_i^P} + \frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}}\right)}$$

K_P number of clusters that partition P, NI ^ P represents the number of data in i- cluster partition P, N_j ^ L represents the number of data in j- cluster partition L, N_ij ^ PL represents the number of partitioning the data together in clusters i- I've j- P and the cluster is partitioned L, N number of data shows and τ is a permutation of numbers from one to N (i) is. If the partitionP & L be quite similar label, then the maximum value FM is one and if the two are quite different partitioning zero.It is worktable to say that stability in the form of partition Ensemble I is calculated as following.
(8)

$$tability(\text{Ensemble}_i) = \frac{1}{|RefSet|} \sum_{j=1}^{|RefSet|} FM(\text{Ensemble}_i, RefSet_j)$$

Here the selected clusters action is carried out in two phases. First in Phase 1, an evolutionary algorithm tries to find a subset of clusters that have the greatest stability. The evolutionary algorithm has a chromosome bit that is long with the total number of produced clusters in different production part.Each of the genes in this chromosome can take number of one or zero. Number one indicates that a number of genes in the clusters-is that cluster is selected and zero in a gene cluster m- number m that have not been selected among the clusters. To calculate the fitness function of evolutionary algorithm, the mean difference in the stability of clusters selected from number one (maximum stability mean that-is selected clusters), the calculation is made.To do so, first we raise the example. Suppose that 13 we have the following data.

Table 1: A collection of data with set of 13 hypothetical data
This data provided in Figure 2 on the features space.

Index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
characteristic x	**1**	**1**	**2**	**2**	**3**	**4**	**5**	**4**	**3**	**2.5**	**3**	**2.5**	**3.5**
characteristic Y	**1**	**2**	**1**	**2**	**3**	**5**	**4**	**4**	**2**	**2**	**2.5**	**3**	**3.5**

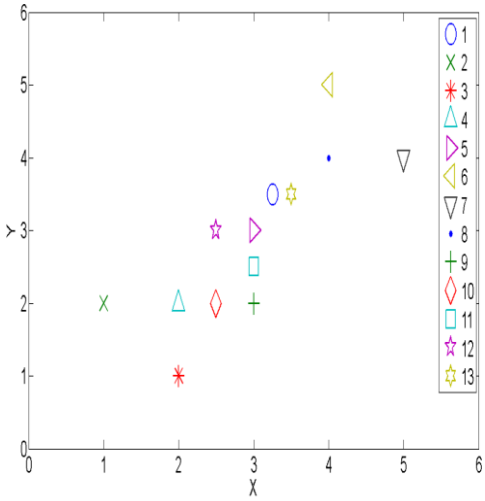


Figure 2: Represent the data set table in the feature space.

Table 2 shows a consensus ReSet of arbitrary size 7. In the table (2) amounts gray rectangle indicates that the clustering RefSet_2 data number11 (as it is in line 11) is reserved to cluster 1.

Table 2: a consensus ReSet with size 7

Data index	$RefSet_1$	$RefSet_2$	$RefSet_3$	$RefSet_4$	$RefSet_5$	$RefSet_6$	$RefSet_7$
1	2	1	3	3	3	3	3
2	1	2	1	1	1	1	2
3	1	2	1	1	1	1	2
4	1	2	1	1	1	1	2
5	2	1	2	2	2	3	1
6	3	3	3	3	3	2	3
7	3	3	3	3	3	2	3
8	3	3	3	3	3	2	3
9	2	1	2	2	2	3	1
10	1	2	2	2	2	1	1
11	2	1	2	2	2	3	1
12	2	1	2	2	2	3	1
13	2	1	3	3	3	3	3

After performing all the steps we have:
(25)

$$FitnessFunction([1,1,0,1]) = 14$$

So the same can be easily calculated and found that fitness of below chromosome is zero.

1	0	1	1

Figure 4: Showing a candidate solution (chromosome)

Then $FitnessFunction([1,0,1,1]) = 0$

On the other hand, after a nationwide we will see that the best chromosome is below chromosome. The level of performance function on this chromosome is 21.

0	0	1	1

Figure 5: Showing the best solution (chromosome)

Final correlation matrices that defined with chromosomes figure (5) that are given in the table (9).
In the final step of the correlation matrix obtained from third consensus optimized, is considered as a similarity matrix. In this case, a hierarchical clustering algorithm is considered as a function of final collector and gives

the resulting correlation matrix as input and returns the clustering of a final deal.The other possibility is that the correlation matrix to be considered as a new data collection and clustering done in this space. The new data set consider each column as one feature and each row as one data. We called this new data collection **interface space**.
Then, in the intermediate space we do a k-means clustering algorithm or fuzzy k-means.

Table (9): the final optimized correlation matrix for partitions of Table 7

1	0.5	0.5	0	0.5	0.5	0.5	0.5	0.5	0	0	0	1
0	0	0	0.5	0	0	0	0	0	1	1	1	0
0	0	0	0.5	0	0	0	0	0	1	1	1	0
0	0	0	0.5	0	0	0	0	0	1	1	1	0
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
0.5	0	0	0	0	1	1	1	0	0	0	0	0.5
0.5	0	0	0	0	1	1	1	0	0	0	0	0.5
0.5	0	0	0	0	1	1	1	0	0	0	0	0.5
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
0	0.5	0.5	1	0.5	0	0	0	0.5	0.5	0.5	0.5	0
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
0.5	1	1	0.5	1	0	0	0	1	0	0	0	0.5
1	0.5	0.5	0	0.5	0.5	0.5	0.5	0.5	0	0	0	1

4- The criteria for measuring quality
One of the drawbacks to measure is the similarity between two clusters difficult to label. To implement cluster label in 2 sets should try different permutations of the clusters similar to a name and a number to be addressed so that we can measure the similarity of two sets. For example, consider 2 sets Figure 6 with 9 samples that each cluster and each cluster has 3 with 3 samples. The two are quite similar clustering while labels such clusters is quite different.

Figure 6: An example of the problem of matching labels

Object	1	2	3	4	5	6	7	8	9
Partition 1	1	1	1	2	2	2	3	3	3
Partition 2	3	3	3	1	1	1	2	2	2

To solve the problem of cluster tag label can still consider the first set label stable and change second set of clusters label so obtain the maximum similarity with the first set and introduce those circumstances match (or accuracy) of the partition as the similarity of two set.
One way to avoid compliance labels is using MI (Mutual Information), (Alizadeh et al., 2014; Parvin and Minai, 2015). Confusion matrix for the series A and B in this case, consider that the rows and columns represent clusters is set A set B. Consider the following definitions:
N_{ij} :The entry (i, j) is a matrix that represents the number of samples in cluster i in cluster j sets A and B are set.
$N_{i.}$:Total amount of the row i or the total samples in cluster i in the set A.
C_A:The number of clusters of A
C_B:The number of clusters of A

It is worth noting to say 2 be equal there is no requirement that the number of clusters collection 2. Due to high definition MI relationship is defined as below
(26)

$$MI(A,B) = \sum_{i=1}^{c_A}\sum_{j=1}^{c_B}\frac{N_{ij}}{N}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)$$

Since there is no upper limit on the amount of MI, relationship mutual information normalized (which is normalized MI (we define above normal relationship that NMI is as follows:

(27)

$$MI(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}\frac{N_{ij}}{N}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{j=1}^{c_A}N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{c_B}N_{.j}\log\left(\frac{N_{.j}}{N}\right)}$$

If two sets A and B be quite similar then the NMI maximum value 1 and if 2 sets are quite different from each other it returns a value of 0.For example, consider the series A and B that are presented in Table 10. In following set NMI is calculated as follows.

Table (10): A set of data

Object	1	2	3	4	5	6	7	8
Partition A	1	1	2	2	3	3	4	4
Partition B	2	1	3	3	2	1	2	2

Interaction matrix 2 sets for the above is like figure 7, and amount of NMI is equal to**:**
(28)

$$NMI(A,B) = \frac{-2\times 5.5452}{-11.0904-8.3178} \cong 0.5714$$

Observed that amount of BMI value obtained with the same label is right on top of two sets.

	B →			
A ↓	1	2	3	Total
1	1	1	0	2
2	0	0	2	2
3	1	1	0	2
4	0	2	0	2
Total	2	4	2	

Figure 7: Example provided matrix interference

Rand is a simple method for measuring the similarity between the two series as follows screw. See following definitions:
n_{11}:The number of pair's samples in sets A and in Category B, with a cluster

n_{00}: The number of pair's samples in sets A and in Category B in two different clusters, and in none of these two is not together in one cluster.

n_{10}:The number of pair's samples in series A are in one cluster, but in the set B is in 2 different clusters.

n_{01}:The number of pair's samples in a series B in a cluster, but are different in set A into 2 clusters. It is obvious that $n_{11}$and n_{00}case shows that the two sets have the same result on a pair and $n_{01}$and n_{10}cases in states that have two different conclusions about a particular pair. In general, $\frac{N(N-1)}{2}$pairs of different samples in a set of N member there, so:

(29)

$$n_{00} + n_{01} + n_{10} + n_{11} = \frac{N(N-1)}{2}$$

Randrelationshipis definedas follows:

$$r(A,B) = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} = \frac{2 \times (n_{00} + n_{11})}{N(N-1)}$$

Rand relationship for two sets quite similar returns a value of 1, but if 2 sets be different value is not logical. Relationship AR dissimilar fault Rand relationship when the two sets were dissimilar and return it to the reasonable amount. Suppose that two sets with the number of clusters with the equal number of sample. If the two are quite dissimilar as regards NMI AR algorithm returns a value close to 0 (Alizadeh et al., 2014; Parvin and Minai, 2015). AR is calculated as the definitions of the NMI relationship are as follows:

(31)

$$AR(A,B) = \frac{\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}\binom{N_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}$$

(32)

$$t_1 = \sum_{i=1}^{c_A}\binom{N_{i.}}{2};\ t_2 = \sum_{i=1}^{c_B}\binom{N_{.j}}{2};$$

For example, in the previous example amount of AR is equal with:

(33)

$$AR(A,B) = \frac{2 - \frac{8}{7}}{\frac{1}{2}(4+8) - \frac{8}{7}} = \frac{3}{17}$$

Other criteria in the assessment of a clustering are Fisher criteria which introduce in the proposed method. Relation to this criterion is given below.

(34)

$$FM(P,L) = \max_{\tau}\sum_{i=1}^{K_P}\frac{2 \times N_i^P \times (\frac{N_{i\tau(i)}^{PL}}{N_i^P} \times \frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}})}{N \times (\frac{N_{i\tau(i)}^{PL}}{N_i^P} + \frac{N_{i\tau(i)}^{PL}}{N_{\tau(i)}^{L}})}$$

$K_P$is the number of clusters partition P; N_i^Prepresents the number of existed data in the cluster from the partition P is i-; N_j^Lrepresents the number of data in the cluster partition L is my j-; N_{ij}^{PL}^ PL represents the number of partitioning the data together in clusters i- I've j- P and the cluster partition is L; N shows the total number of data; τ is a permutation of numbers from one to N respectively.

If the partition P and label L be quite similar, then FM the maximum value is one and if the two partitioning are quite different it returns zero. For example, the amount of returned FM is calculated.

(35)

$$FM(P, L) = 0.5147$$

• Comparison of three methods of FM; AR, and NMI

Almost three FM, AR and NMI have better results than other methods and do not require implementing with different sets of clusters label. In the below picture we are trying to provide a comparison between the three methods to consider more appropriate methods for future studies. To provide the results of the first attempts to express some of our content.First consider an artifact of the data set with 1500 samples and take the 3 clustering. The synthetic data sets distributed in three clusters is the same as that in each cluster 500 is shown. Suppose we show labels of the data set with T.Also suppose $T_i$label i- Assume data is safe. After T_i is equal to 1 $1 \leq i \leq 5002$ if$501 \leq i \leq 1000$; and 3 is if$1001 \leq i \leq 1500$.

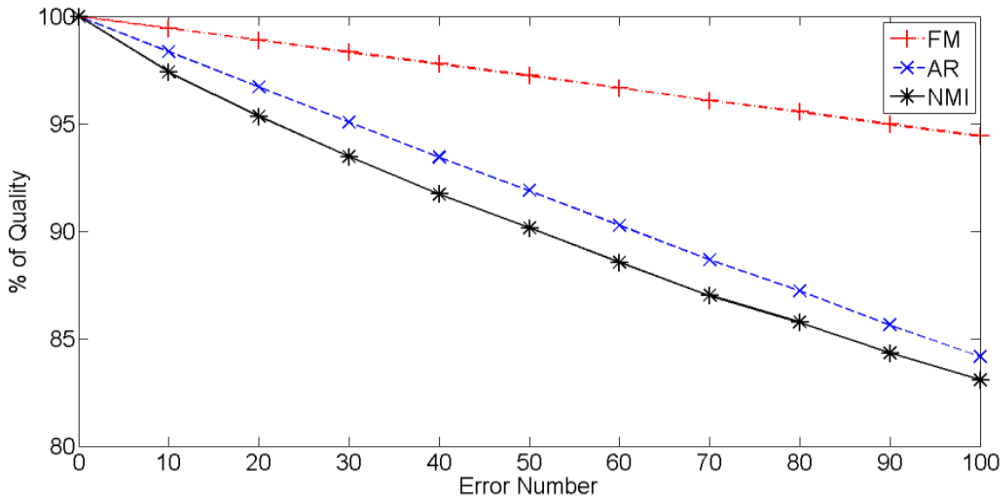


Figure 8: The result of Matching between the results of clustering algorithm hypothetical and actual labels of samples in a data set with 1500 samples and 3 clusters

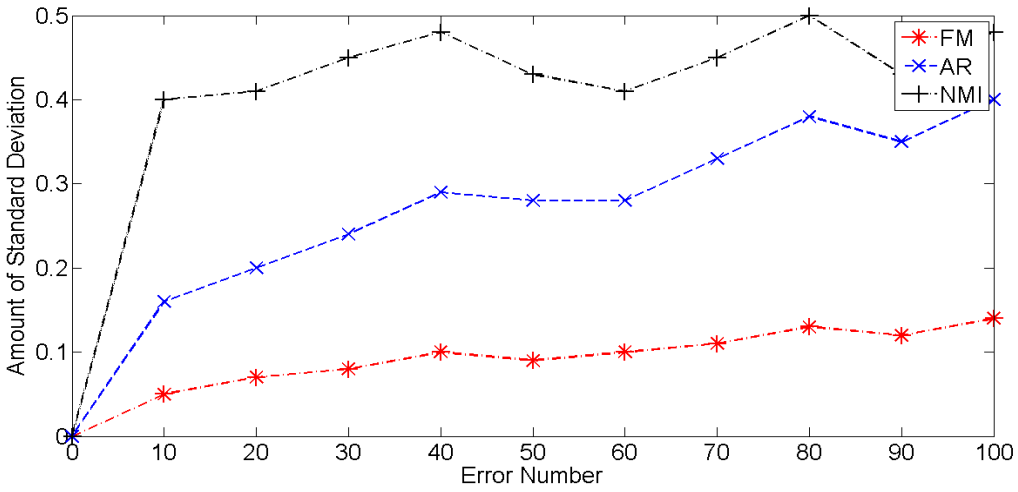


Figure 9: The level of instability as a result of matching between the results one hypothetical clustering algorithm

Now suppose thatone clustering algorithm X have been run on synthetic data sets. A hypothetical output of this algorithm is supposed to be shown with A. Also suppose $A_i$is label of i.Also suppose $A_i$is equal with 2 if be$1 \leq i \leq 500$also let AB = 2; is 3 if be$501 \leq i \leq 1000$; and is 1 if$1001 \leq i \leq 1500$.Now we calculate Fisher (FM), the index Rand (AR) or normalized mutual information (NMI) between A and T, we will discover that the accuracy of this algorithm is 100%. Now suppose the size of Error (which is an arbitrary numerical value), this algorithm has error. If it is Error = 3, we select three random value r, p and q between 1 and 1500.Then $A_q A_p A_r$we replaced with other values, it means we add one to their value. For example, we add one toA_q; if it is 1, we change it to two, if it is two, then we change it to and if it is 3 we change it to 1. In short, we can say:
(36)

$$A_q = (A_q + 1) mod\ 3$$

(37)

$$A_p = (A_p + 1) mod$$

(38)

$$A_r = (A_r + 1) mod\ 3$$

Now that the ASEAN amount (random r, p and q) is varies with T, we calculate FM, AR or NMI respectively. If r = 782, p = 1467 and is q = 1186, FM, AR and NMI will be 99.83, 99.50 and 99.14 percent. If r = 772, p = 905 and q = 1262 is, FM, AR and NMI will be 99.87, 99.60 and 99.43 percent.So a hundred times with the percentage of random values for positions r, p and q Repeat and percentage amounts set for FM, AR and NMI are calculated.

The average percentage amount showrespectivelyμ_{FM}^{3}, $\mu\_\mu_{AR}^{3}$ andμ_{NMI}^{3}, as the values of FM, AR and NMI with three errors (Error = 3) consider. As well as the standard deviation percentage value show respectivelyσ_{FM}^{3}, σ_{AR}^{3}andσ_{NMI}^{3}, we consider the error values of FM, AR and NMI when three errors (Error = 3). Now change the error of 0 to 100, values μ_{FM}^{Error} ‘μ_{AR}^{Error} μ_{NMI}^{Error} σ_{FM}^{Error} ‘σ_{AR}^{Error} σ_{NMI}^{Error} and we calculate in figure represent in (8) and (9).

Table 11: Summary of used characteristics of standard data set

number of samples	Number of features	Number of classes	Data collection
178	13	3	Wine
683	9	2	breast-cancer
345	6	2	Bupa
323	4	7	Galaxy
214	9	6	Glass
400	2	2	Halfring
150	4	3	Iris
351	34	2	Ionosphere
462	9	2	Saheart
1484	8	10	Yeast

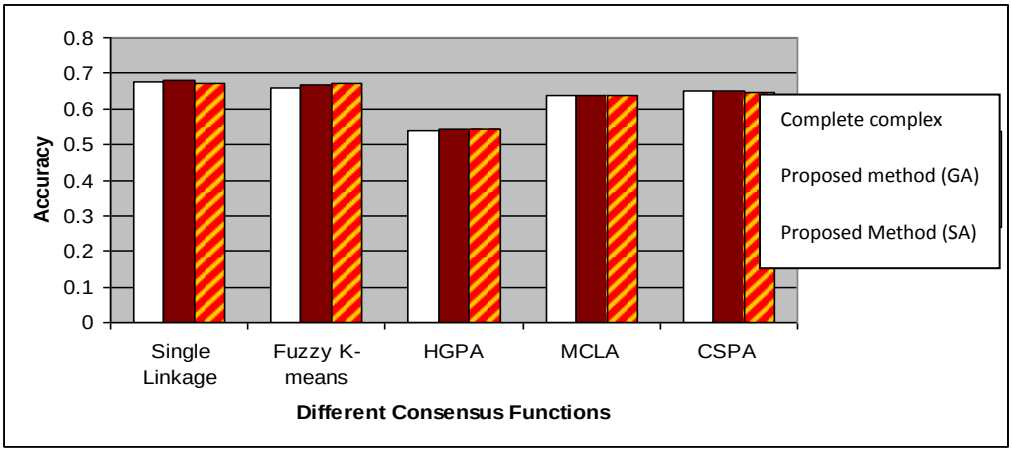


Figure 10: Diagram of average accuracy on all data sets against various methods

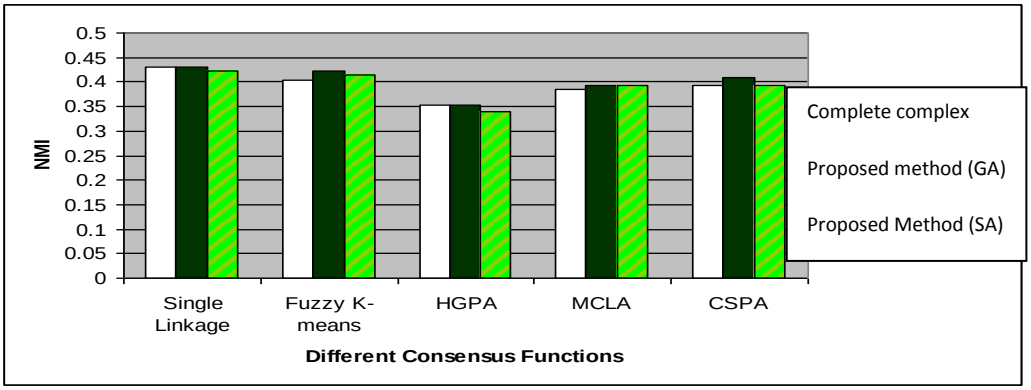
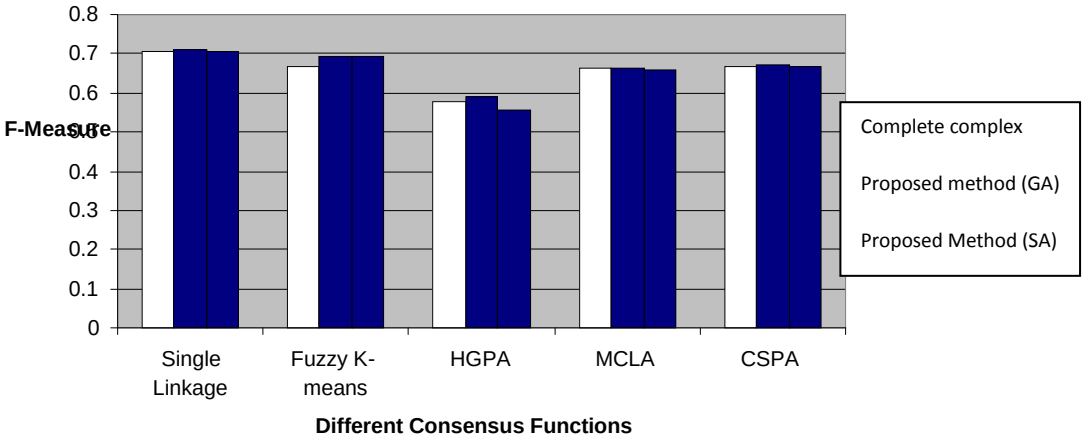


Figure 11: Diagram of average normalized mutual



information on all data sets against various methods

Figure 12: DiagramFisheraveragestandardonalldata setsagainstvariousmethods

Figure 8 shows the results of the implementation of the results of a clustering algorithm X and labels hypothetical instances in a cluster is a data set with 1500 samples and 3 clusters. By analyzing the figure above, we have tried to identify a method that have more accurate and closer results to the actual accuracy of clustering (obtained based on the error). In this section has report the results of the application of the method and parameters used different data collection.The proposed method has been tested and implemented in MATLAB 7.1 environment. The results are reported on average 10 times the performance of stand-alone application.The performance of different methods of clustering Calculated with three criteria, accuracy, NMI and F-Measure Figure (10), (11) and (12) results indicate that as is seen not only does not decrease in most cases

improved.Select primary clusters with genetic algorithm and smelting to improve the performance of the proposed method in choosing the most optimal primary cluster for helps of the final clustering combination.In practice, applying the evolutionary algorithm to select clusters obtained two clusters stable and unstable clusters. Full Assembly and the governing body of clusters using evolutionary algorithm selects in NMI and F-Measure and calculate carefully the various data sets and to facilitate the analysis, average results on 10 sets of data are shown in Figures 10, 11 and 12.UCI standard used data set is data collection that results of almost all of them in recent studies reported by using this data set. The proposed method has been tested on 10 normal standard set. To perform these experiments tried to data collection in terms of the number of classes, the number of features and the number of samples to be varied from maximum possible strength and repeatable results of operations. For normalization method each the data set characteristics zero mean and variance, N(0,1)were normal.For all this data set, the number of clusters and label specimens are known from before. Therefore, the percentage of cases that were diagnosed correctlyare used as a measure of performance clustering method. In fact, after solving the correspondence between the labels and the actual clusters can determine the error rate. In all ways using the K-means algorithm is used as the basic algorithm. The numbers of initial results arestable and equal with 120.In fact, the number of K-means algorithm is obtained by manipulating the parameter k. In this way, the four groups of 30 each of the first results, taking into account the number of clusters used by the algorithm size k, k + 1, k + 2 and k + 3 is obtained. Also, to create greater dispersion in the preliminary results of sampling without replacement at the rate of 50% has been used. Also, for the ultimate partitioning of single connection methods on the correlation matrix is used Fuzzy K-means and methods of graph-based method HGPA, MCLA and CSPA. Table (11) shows summary of standard data sets used in the experiments.

As it is seen in the average, in most cases the result is improved efficiency, therefore we can conclude that not only reduce the selected clusters cause in decreasing efficiency, but also it cause to increasing performance in many cases.

Also, since this is the continuation of what has already been done by Alizadeh and Minai, A comparison is done between their work and work which is done before by Mr. Azimi.

Although the proposed method in terms of accuracy in Table 12, is better than other methods, but still cannot claim that the proposed method is best. Remains to be seen whether these results are not coincidence that change again with different initialization parameters and algorithms, otherwise the results will not eat. For a closer look and find out whether this advantage is significant, it should be one of the statistical survey took refuge indeed.

Here we use the method of truly poll Friedman. This method is suitable for the following reasons account for comparing several methods simultaneously. This method classify all of method based on their performance on a regular data collection and it consider one for rank the most effective way the way with the highest performanceand consider M (M Methods is a Total) with less performance.In cases which exists some methods with same ranking, the average ranking is considered for them. For example, if A and B are the second and third method performance between methods, ie their rank be 2 and 3, but the performance of them be equal, their rank respectively will be equal with 2.5 and 2.5. This method is in detailed below.

Friedman hypothesis suggests that tis of no signified methods have not difference method. To refute this hypothesis, for showing significant difference methods should act as follows:

First, supposesr_i^j represents rank i- way I've j- is in the data set. I j- calculation method is an average rating of about 12. (12)

$$R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j$$

We calculate average rating of all methods. Degree of freedom is k-1 where k represents the number of methods. 5 ways because we have 4 degrees of freedom-is the issue. Now we calculate the following relationship χ_F^2value of about 13. (13)

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right)$$

By calculating about 13 times the amount χ_F^2will be 16.10 that the amount of the expected value in table 4 kegs with degrees of freedom equal to the greater 489.9.

Then the null hypothesis is rejected and we conclude that the difference between the methods is meaningful. Since the average of each methods (a) the full Assembly, (B) GA, (c) SA, (d) Alizade (a) Azimi, respectively 2.3, 2.0, 3.4, 3.5 and 3.8. So GA method meaningfully is better than other methods.

5. Conclusion

In this paper, we propose a method in hybrid clustering that changes its choice quality based on the type of data collection. The proposed method first performs an initial hybrid clustering and then based on dispersion between the results of initiating clustering algorithms and initiating hybrid clustering algorithms to explore the possible choices in each data set.Then select the best subset of preliminary results based on algorithms revelation has been, well. Then it is done to final clustering on the selected subset to obtain the final clusters. The proposed method for the selection of cluster primary classifications based on each data set into the final composition of the cluster is in a dynamic manner. The results obtained indicate the efficiency and ability of the proposed method of clustering information.

Table 12: Comparison of Total full accuracy of the proposed method and previous methods

Azimi	**Alizade**	**SA**	**GA**	**Data collection**	**Full Assembly**
96.63	96.63	96.63	96.63	**96.74**	Wine
95.91	95.73	95.17	95.29	**97.03**	Breast-Cancer
54.75	54.33	55.07	**55.10**	55.01	Bupa
29.97	31.27	30.65	**32.82**	30.03	Galaxy
55.05	57.76	45.79	**57.86**	56.81	Glass
67.70	74.48	74.50	74.50	**76.38**	Halfring
89.33	89.33	89.33	89.33	**89.60**	Iris
70.74	70.60	70.65	**70.74**	70.71	Ionosphere
56.06	63.36	63.27	63.29	**63.44**	Saheart
43.40	42.75	43.05	42.93	39.42	Yeast
65.954	67.642	66.411	**67.831**	67.517	ALL

References

[1]. Azimi, G,"The distribution of the hybrid clustering", MSc Thesis, University of Science and Technology, 2008.
[2]. Aarts E. H. L. and Korst J. ***Simulated Annealing and Boltzmann Machines***, John Wiley & Sons, Essex, U.K, 1989.
[3]. Akbari E., Dahlan H.M., Ibrahim R., Alizadeh H.: ***Hierarchical cluster ensemble selection***. Eng. Appl. of AI 39: 146-156 2015.
[4]. Alizadeh H., Minaei-Bidgoli B., Parvin H. ***Optimizing Fuzzy Cluster Ensemble in String Representation***.IJPRAI 27(2), 2013.
[5]. Alizadeh A., Minaei-Bidgoli B., Parvin H. ***Cluster ensemble selection based on a new cluster stability measure***. Intell.Data Anal. 18(3): 389-408, 2014.
[6]. Ayad H.G. and Kamel M.S., ***Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters***, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, VOL. 30, NO. 1, 160-173, 2008.
[7]. Barthelemy J.P. and Leclerc B., ***The median procedure for partition***, *In Partitioning Data Sets, AMS DIMACS Series in Discrete Mathematics*, Cox, I. J. et al eds., 19, pp. 3-34, 1995.
[8]. Baumgartner R., Somorjai R., Summers R., Richter W., Ryner L., and Jarmasz M., ***Resampling as a Cluster Validation Technique in fMRI,****JOURNAL OF MAGNETIC RESONANCE IMAGING* 11: pp. 228–231, 2000.
[9]. Breckenridge J., ***Replicating cluster analysis: Method, consistency and validity,****Multivariate Behavioral research*, 1989.
[10]. Dudoit S. and Fridlyand, J., ***Bagging to improve the accuracy of a clustering procedure, Bioinformatics***, 19 (9), pp. 1090-1099, 2003.

[11]. Faceli K., Marcilio C.P. Souto d., ***Multi-objective Clustering Ensemble***, *Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06)*, 2006.
[12]. Fern, X.Z. and Brodley, C. E. ***Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach,****In Proc. 20th Int. conf. on Machine Learning, ICML 2003*, 2003.
[13]. Fern X.Z., and Lin W.,***"Cluster Ensemble Selection"***. Statistical Analysis and Data Mining 1(3): 128-141, 2008.
[14]. Fischer B. and Buhmann J.M., ***"Bagging for path-based clustering"***, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.1411–1415, 2003.
[15]. Fred, A. and Jain, A.K. "***Data Clustering Using Evidence Accumulation***", *Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City*, pp. 276 – 280, 2002.
[16]. Fred A. and Jain A.K., "***Robust data clustering***", *in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR ,USA*, vol. II, pp. 128–136, 2003.
[17]. Fred A.L. and Jain A.K. ***"Combining Multiple Clusterings Using Evidence Accumulation***".*IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.
[18]. Fred A. and Jain A.K., "***Learning Pairwise Similarity for Data Clustering"***, *In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06)*, 2006.
[19]. Fred A. and Lourenco A. (2008), ***"Cluster Ensemble Methods: from Single Clusterings to Combined Solutions",****Studies in Computational Intelligence (SCI)*, 126, 3–30.
[20]. Fridlyand J. and Dudoit S. ***"Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method".****Stat. Berkeley Tech Report*.No.600, 2001.
[21]. Jain A., Murty M. N., and Flynn P. (1999), ***Data clustering: A review***. *ACM Computing Surveys*, 31(3):264–323.
[22]. Kuncheva L.I. and Hadjitodorov S. ***"Using diversity in cluster ensembles***".*In Proc. of IEEE Intl. Conference on Systems, Man and Cybernetics*, pages 1214–1219, 2004.
[23]. Kuncheva L.I. and Whitaker C. J., "***Measures of diversity in classifier ensembles***", *Machine Learning*, 2003.
[24]. Lapointe F.J. and Legendre P.***Thegeneration of random ultrametric matrices representing dendrograms***.*Journal of Classification, Springer New York*, Vol. 8, No. 2, pp 177-200, 1991.
[25]. Law M.H.C., Topchy A.P., and Jain A.K. ***"Multiobjective data clustering***".*In Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 424–430, Washington D.C, 2004.
[26]. Levine E., Domany E., "***Resampling Method for Unsupervised Estimation of Cluster Validity".****Neural Computation* 13: 2573-2593, 2001.
[27]. Melanie M., "***An Introduction to Genetic Algorithms***", A Bradford Book The MIT Press, Cambridge, Massachusetts. London, England, Fifth printing, 1999.
[28]. Minaei-Bidgoli B., Topchy A. and Punch W.F., ***"Ensembles of Partitions via Data Resampling"***, *in Proc. Intl. Conf. on Information Technology, ITCC 04, Las Vegas*, 2004.
[29]. Parvin H., Minaei-Bidgoli B., Alinejad-Rokny H., Punch W.F. ***"Data weighing mechanisms for clustering ensembles***". Computers & Electrical Engineering 39(5): 1433-1450, 2013.
[30]. Parvin H., Minaei-Bidgoli B. "***A clustering ensemble framework based on selection of fuzzy weighted clusters in a locally adaptive clustering algorithm***". Pattern Anal. Appl. 18(1): 87-112, 2015.
[31]. Parvin H., Mirnabibaboli M., Alinejad-Rokny H. ***"Proposing a classifier ensemble framework based on classifier selection and decision tree"***. Eng. Appl. of AI 37: 34-42, 2015.
[32]. Roth V., Braun M.L., Lange T., and Buhmann J.M., ***"Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data***", *ICANN 2002*, LNCS 2415, pp. 607–612, 2002a.
[33]. Roth V., Lange T., Braun M., and Buhmann J., ***A "Resampling Approach to Cluster Validation"***, *Intl. Conf. on Computational Statistics, COMPSTAT*, 2002b.
[34]. Saha A., Das S. ***"Categorical fuzzy k-modes clustering with automated feature weight learning"***. Neurocomputing 166: 422-435, 2015.
[35]. Shamiry O., Tishby N., ***"Cluster Stability for Finite Samples***", *21st Annual Conference on Neural Information Processing Systems (NIPS07)*, 2007.
[36]. Strehl A. and Ghosh J., ***"Cluster ensembles - a knowledge reuse framework for combining multiple partitions***".*Journal of Machine Learning Research*, 3(Dec):583–617, 2002.
[37]. Topchy, A., Jain, A.K. and Punch, W.F., "***Combining Multiple Weak Clusterings***", *Proc. 3d IEEE Intl. Conf. on Data Mining*, pp. 331-338, 2003.
[38]. Xiong S., Azimi J., Fern X.Z.,***"Active Learning of Constraints for Semi-Supervised Clustering"***. IEEE Trans. Knowl. Data Eng. 26(1): 43-54, 2014.