MAT JOURNALS

Journal of Computer Science Engineering and Software Testing
Volume 3 Issue 3

The Influence Select Feature on The Clustering Algorithm

MaysamToghraee[1] , Hamid Parvin[2], Farhad Rad[3]

[1]Department software computer engineering, science and research, Islamic Azad University, Yasouj, Iran, Email**: may.toghraee@yahoo.com**

[2]Faculty of engineering, department of computer science, Islamic Azad University, noorabad mamasani, fars, Iran, Email**: parvinhamid@gmail.com**

[3]Faculty of engineering, department of computer science, Islamic Azad University, Yasouj, Iran, Email**: f_rad@hotmail.com**

Abstract

The methods available for structuring the collections are: Classification methods and clustering methods. The following is a summary of the basic principles of the text mining process. Then some of the important methods for classifying the texts are evaluated together. Clustering is the process of organizing anarchy into groups whose components are similar. A cluster is an irregular set of similarities that are heterogeneous with other components of the cluster. The goal of clustering is to achieve a steady and reliable correlation, and to identify the logical connection between them. Therefore, clustering algorithms can be used in a wide range of subject areas. Since clustering results can be varied with the number of terms used, several empirical methods are proposed to diagnose the approximate number of terms that can be expected to provide an appropriate distribution of data among clusters and to define the upper and lower limits of the clustering algorithm. Our goal is to study data mining on different data, the results of this method show that the meta-heuristic method is suitable for the meta cluster algorithm compared to other clusters.

Keywords: *feature selection, data mining, meta cluster algorithm*

INROUDUCTION

Decision trees as one of the data mining techniques are widely used in the validation of bank customers Identify them for credit facilities. The main issue is the complexity of decision trees, excessive size, lack of flexibility and low accuracy in categorization. The purpose of this paper is to present a hybrid model for tree decision making by genetic algorithm technique to solve the problems mentioned above in order to validate bank customers. It seems that by choosing appropriate features and making decision trees, the genetic algorithm can reduce the complexity and increase the flexibility of decision trees. In the proposed hybrid model, first, credit data is split into two clusters using SimpleKmeans clustering technique. Then, using the genetic algorithm, five feature selection algorithms based on three Wrapper filter approaches and an embedded design based on a genetic decision tree determine the important validation features in the data set. In the following, five decision trees based on C4.5 blegorithm are made in each cluster with a set of selected features. The best decision trees in each cluster will be chosen on the optimality criteria considered in this paper and will be combined to create a final decision tree for validating bank customers. The Gatree Machine Learning Tool and the GATree software are used to achieve results. The results of this study indicate that the use of the proposed hybrid model in decision making tree leads to increase classification accuracy compared to many of the algorithms compared in this paper, but the complexity of the proposed hybrid model algorithm is compared with some sorting algorithms This article is more.It makes the distance between both samples less

representing the real distances. So, the quality of classifying or clustering are unpleasantly unreal and drop. It can be stated in another way. It can be said that some clusters or branches in feature's atmosphere are more coherent with some special features; Three general ways have been submit to overcome the above dimension problem: (a)using subspaces determined for clusters or branches by user, (b) using feature selection methods or decreasing dimensions like analyzing main factors and finally (c) using subspace clustering or subspace classifying methods. We discuss about the feature selection methods (b) in this report. A lot of solutions and algorithm have been represented for feature selection issue. A lot of solutions and algorithms have been represented for feature selection issue, some of which are 30 or 40 years old. The problem about algorithms when represented was their calculating feature. However, fast computers and big saving sources have made this problem unimportant, beside, big data sets for new issues has made it important to find a fast algorithm for this issue. Feature selection has 2 types (toghraee .M & et al, 2016) :

supervised feature: Labels are used during feature selection algorithm(Zhao &Liu , 2012)

unsupervised feature :Labels are not used during feature selection algorithm(G.D,2012)

Research domain is just limited to the supervised feature selection while labels are used during feature selection algorithm.

We are seeking for the below targets using local searching methods and imitating the nature searching algorithm; a) discuss about the Hierarchical algorithms, are far more prom partition methods. For example, clustering algorithm including non-isolate of such clustering chains or center clusters, works well. But partition techniques such as k-mean clustering good sample feature selection efficiency methods(Toghraee M& et al, 2016).

b) explain the assessing algorithm of this research

c) looking for tests and results from these methods using real data sets.

EFFICIENCY FUNCTION METHODS

Different types of efficiency function for various subsets includes (1) wrapper methods, (2) embedded methods, (3) filter methods

Wrapper methods: Categorical issues often include a large number of features, but they are not all useful for categorization. Relative and exaggerated features may even reduce classification accuracy. Since the selection of suitable features is an NP-hard problem, searches for fast and efficient algorithms continue. In this paper, a distributed learning method is proposed to select appropriate features for classification issues. The results of implementing the proposed method on several datasets describe the performance of the proposed method in comparison with other methods(toghraee M& et al,2016) .

Filter methods Simultaneous analysis of information has become widespread in many online analytical systems, such as document analysis or analysis of purchase history. Contrary to ordinary multivariate observations, each object is detected by its degree of simultaneous occurrence with different items, and the goal is often to develop cluster structures between objects and items so that mutually identical pairs of items-items form a common cluster . One of the common uses of common cluster analysis can be found in the CF filtering. CF is the basis for obtaining personalized recommendations in various Web services, in which priority of similarity is considered among users. In this paper, a kind of fuzzy joint clustering model is introduced, which is derived from the simultaneous statistical clustering

model and, while briefly reviewing the CF framework, demonstrates its applicability for CF tasks. (Toghraee M& et al , 2016).

Efficiency function

To calculate the efficiency function, we should first calculate the relationship of each feature with other features and label.
Capacity building is one of the issues that is widely used in data mining. This is followed by the partition of a set of n elements of a p-dependent cluster so that all members of a cluster are assigned to the point determined as the center of gravity of that cluster. The purpose of this problem is to minimize the unevenness of all points of a cluster from the center of gravity of the cluster by observing the capacity constraint in each cluster, so that each element is assigned to only one cluster. In this paper, two different methods of solving the problem of clustering are presented. The first method is an ultra-innovative solution based on refrigeration simulation, which uses the answer-finding mechanism of different neighborhoods. The second method is based on the Genetic Algorithm, which uses an innovative local search process. The methods presented have been tested using various sample problems. The computational results indicate the efficiency and ability of the proposed solving methods (toghraee M & et al,2017).

Fit $_{ch}$ = $\sum$ $_{leaser\ (\ max\ ^{j}_{i=1}}$ (*cor*(X_j,X_k) * and (*ch*$_j$,*ch*$_k$) , *th*$_1$) * α + greater (*cor*(X_j,X_k) , *th*$_2$)

Where *f* is the amount of features and α is the big positive number, *th*$_1$ and *th*$_2$ are two thresholds which should be adjusted by the user, *ch*$_j$shows *Ith* of chromosome $_{,\ |\ |}$shows the absolute value, *cor*(X_j,X_k) shows the relationship of *I th* and *k th* features , and (*ch*$_j$,*ch*$_k$) show the logical operator (output is 1 when both inputs are 1,otherwise function output is 0) , *(a,b)* is greater than1 if a≥b ,otherwise function output is *0* and function*(a,b)* is less than 1, if *a≤b* . otherwise function output is *0* (toghraee . M, 20016) .

Evaluation Algorithms

single linkage and complete linkage clustering algorithm

Algorithm 1-3: single linkage and complete linkage clustering algorithm

1. each sample is placed into a cluster and a list of the distance between individual samples and it all sort them in ascending order.
2.Using the sorted list, for any amount $d_k$that represents the dissimilarity pattern , a pattern that the distance between them is less $d_k$the graph is made. If all the samples of this graph, the algorithm will terminate , otherwise repeat this step.
3.The result of this algorithm , hierarchical graph is that the intra you can graph with horizontal cuts, at each level of dissimilarity, a partition of the samples obtained.

Figure1

single linkage and complete linkage clustering algorithm

not have classification. A hierarchical clustering algorithm in the form figure(3.2.1) has been introduced. Based on the similarity matrix update methods in step2, are designed in various hierarchical algorithms.

Algorithm1-3: hierarchical clustering

1.calculating the similarity matrix contains between each sample pair
2.find the nearest two cluster together according to similarity matrix, and combining the two in a cluster and update the similarity matrix,
3.if all the samples were placed in cluster algorithm, otherwise, proceed to go step2.

Figure2
hierarchical clustering

squared error clustering algorithm

One of the objective for partition techniques the squared error, with is separated for compact clusters. Square error algorithm l on a data set x includes k clusters are defined as follows:

$$e^2(x,l) = \sum_{j=l}^{k} \sum_{i=l}^{n_j} \left| x_i^{(j)} - c_j \right|^2 \qquad (3-2)$$

The *i-th* pattern of *the j-th* cluster and the center of gravity of cluster *j-th*is made. *K-mean* algorithm is one of the simplest and most common error is that the square function(ward,1963). It was started with an initialized for the number of clusters and the assignment of new continues so that no example of a pattern to another pattern not more or function of square error is not after a few steps significant changes. Because simplicity of implementation and complexity (n is the number of samples), k-means algorithm is very common. One of the main problem in this algorithm it is sensitivity to initialize clusters, and may be trapped in a local minimum of the objective function.

Algorithm2-3: squared error clustering algorithm

1.random selection of k clusters
2.assign each sample to the nears cluster center and recalculating the center of the cluster.
3.Repeat step2 to replace samples that tend to cluster ends, when membership in the clusters were stable.
4.Division in the clusters based on some heuristic information

Figure3
squared error clustering algorithm

To do this, methods depended on evolutional algorithms have been represented for selecting subsets of features , in this chapter we discuss about the efficiency function of this algorithms.

Mixture resolving clustering algorithm

Data clustering method for mixture resolving a combination of several types have been proposed. The method is that the data generated by one or more of the probability distribution function and purpose , specify the parameters of each of these functions in further work in this area , generating function is assumed as a Gaussian function. Traditional approaches to solving the problem of estimating the most likely (maximum likelihood method) vector parameters is possible. Expectation maximum algorithm as the most common method used to estimates a general purpose method missing data suits. Expectation maximum parameters start and at each stage of sample points by these parameters then the samples are used to update parameters. Quadratic mutual information between individual partition to maximize the agreement. Space used specifically for labeling the clusters to solve less computational complexity and the labeling provides clusters in specific patterns.

Cluster Based Similarity Partitioning Algorithm

In this way, the relationship between the clustering of samples taken within a cluster, and therefore it can be used to generate a measure of similarity between samples pairwise. Then, the same evaluation criteria showed that this sample is used for clustering. If two instances within a cluster, thay are completely unlike each other. This is the simplest method and cluster based similarity partition algorithm is used, it can resemble a fraction of partition so that two samples are within a cluster define. The similarity matrix s with dimensions n× n can be calculated by multiply a sparse matrix $s = \frac{1}{r} HH^{+}$. Now, can any similarity based clustering algorithm on the matrix s used and reused clustering samples. In this clustering, could vertex of the graph, for example, from the edge weight is used as the similarities.

Hyper Graph-Partitioning Algorithm

In the algorithm using of minimum cut the most common information has been estimated. Therefore, the cluster group as a hyper graph cluster each edge of a cluster is defined. Thus, a could graph using cutting the minimum number of hyper edge, the partition is formatted. This method of clustering was call algorithm hyper graph. It is assumed that all hyper edge and vertices, has the same weight. Now divide that can separate hyper edgeand hyper graph to k part with almost equal size the clustering. By keeping a heed maximum 5% imbalance, clustering size is kept the same which is achieved with relationship $k.max_{i\epsilon\{1....k\}} \frac{n_l}{n} \leq 1.05.$

Meta cluster algorithm

Meta cluster algorithm is based on the clustering of cluster. In the way, an evaluation of the sample to the membership gives us cluster, each cluster is also shown as a hyper edge. The purpose meta cluster algorithm grouping and compression associated with each other and assign each sample to a hyper edge is pressed. Hyper edge connected to compression by clustering hyper edge by a clustering algorithm based to graph are determined. To Any cluster containing say hyper edge $C^{(m)}$. compression of the k reduces the number hyper edge $\sum_{q=1}^{r} k_q$.

Evaluation Methods

In this section, the results of applying the proposed method on different data sets and used parameters has been reported.

The most important criterion for determining the efficiency of a classification classification algorithm is the Classification Accuracy-Rate, which measures the accuracy of the entire category. In fact, this is the most popular and general criterion for calculating the efficiency of categorization algorithms, which shows that the designed class bits correctly categorized several percent of the entire set of test records.

The accuracy of the classification is obtained by using the relation I, which states that the two values of TP and TN are the most important values that should be maximized in a single problem. (In multi-series questions, the values placed on the main diameter of this matrix - which should be in the case of CA calculation deductions - should be maximized).

The error rate criterion is exactly the opposite of the ranking accuracy criterion obtained by using the equation II. The smallest value is zero, when we have the best performance, and its value is the same as when we have the lowest performance.

It is necessary to note that in actual problems, the classification accuracy criterion is by no means a suitable criterion for evaluating the efficiency of algorithms, because in relation to the accuracy of

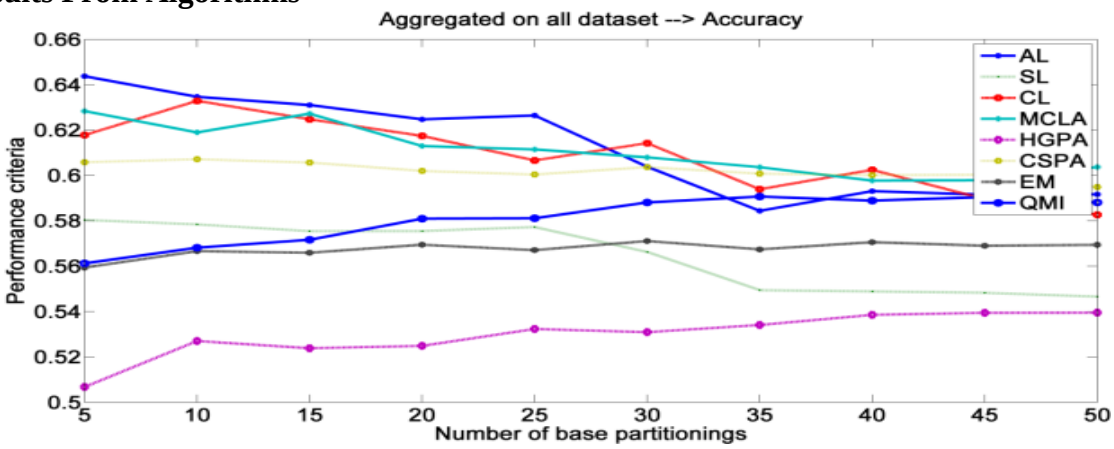
classification, the value of the records of different categories is considered the same. Therefore, in matters that are dealt with inequalities, in other words, in cases where the value of a batch is different from that of another, other criteria are used(toghraee & et al , 2016).

Table1: *Datasets Used In The First Experiment In This Thesis. Starred Data Set"*" Are Real Data Sets.*

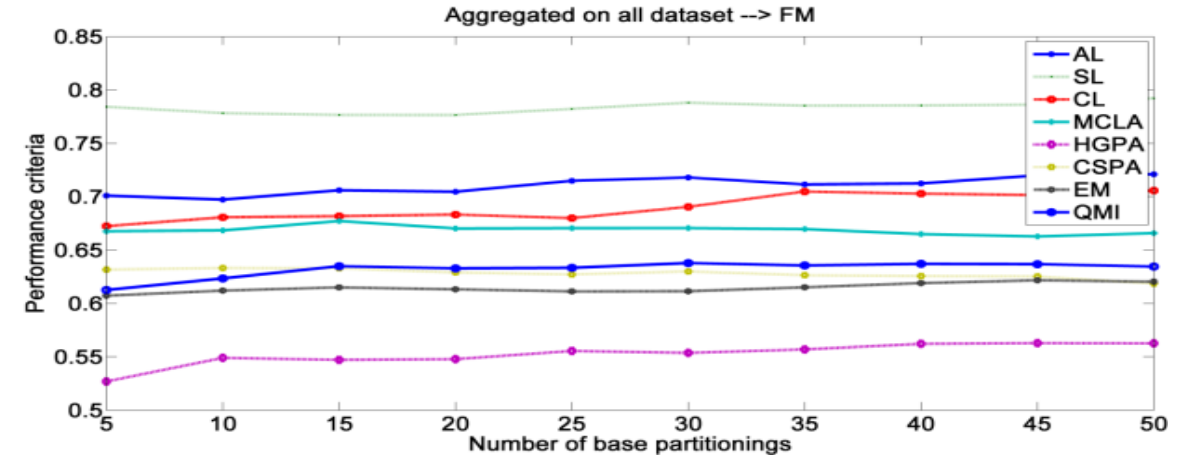
Dataset Name	# of data items	# of features	# of classes	Data distribution per clusters
Breast Cancer*	404	9	2	444-239
Bupa*	345	6	2	145-200
Glass*	214	9	6	70-76-17-13-9-29
Galaxy*	323	4	7	51-28-46-38-80-45-35
SAHeart*	462	9	2	160-302
Ionosphere*	351	34	2	126-225
Iris*	150	4	3	50-50-50
Wine*	178	13	3	59-71-48
Yeast*	1484	8	10	463-5-35-44-51-163-244-429-20-30

These data sets is presented in the table above. The results of this section, by changing various parameters clustering algorithm, we will try to provide an understanding of them and to reach a general conclusion. Intensive clustering algorithm parameters, including rate update (learning rate), sampling rate number of partitions thatconsensus the number of partitions consensus of series$\{5,10,15,\ldots,50\}$sampling rate of sets$\{.2,.4,.6,..,1\}$ update rate of sets$\{0,.1,.2,\ldots,1\}$ are select. In any measure the quality of an average of over 10 performance in a row as them quality is considered(Toghraee, M& et al. 2017).

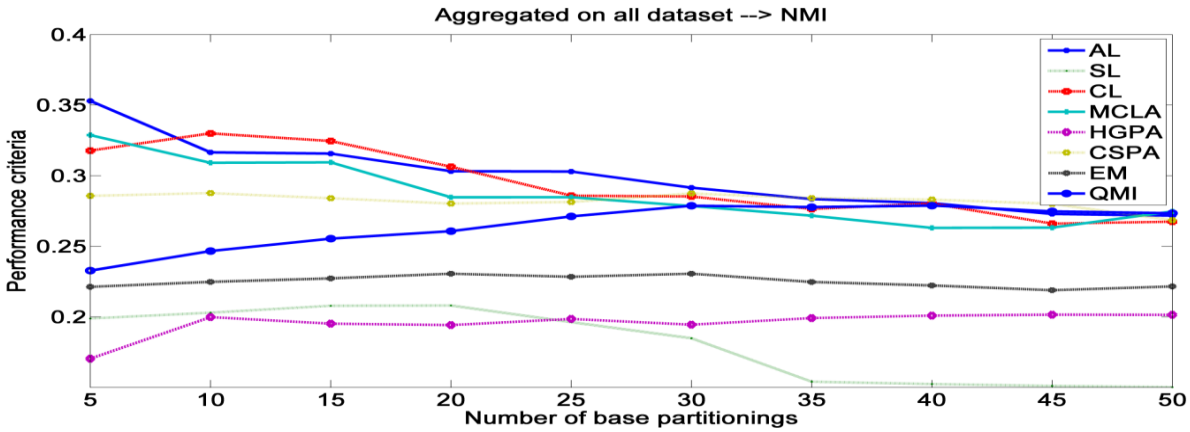
Results From Algorithms



Accuracy



fisher method

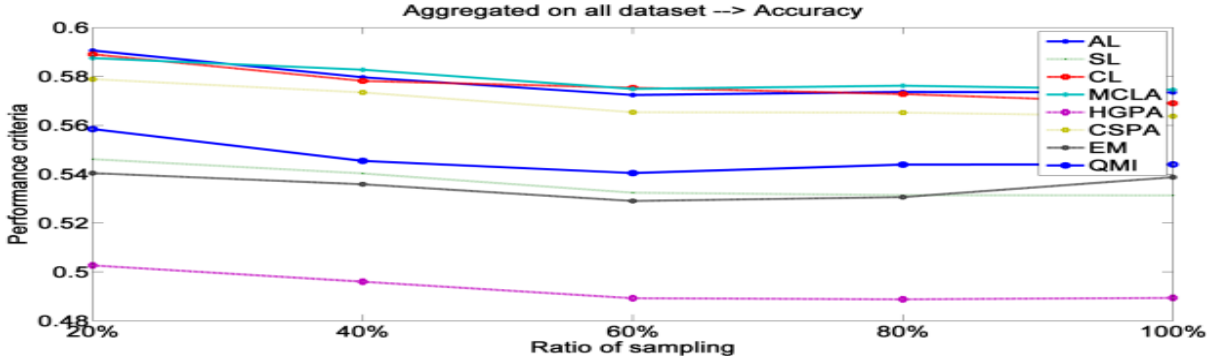


normal Mutual Information

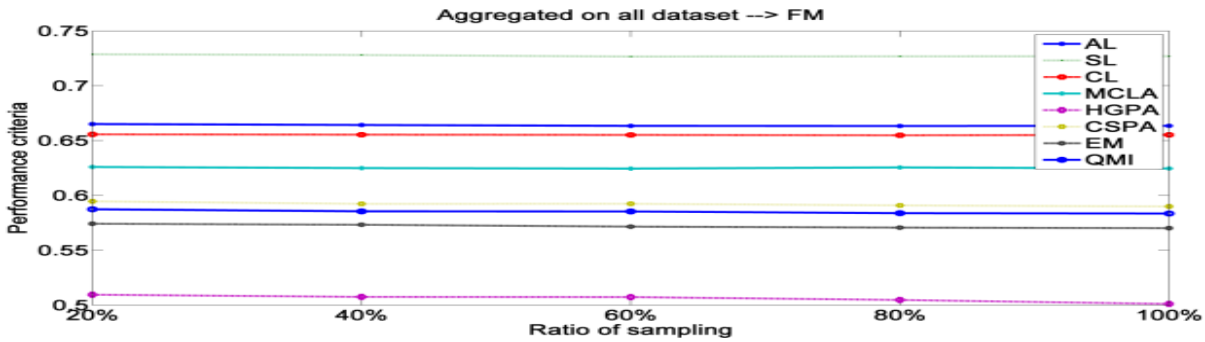
figure.4.

average performance clustering resent on all data sets deal with a variety of different function depending on the number of partition consensus – curve horizontally in the assembly shows the number of partitions and vertical curve shows the quality of the final partition.

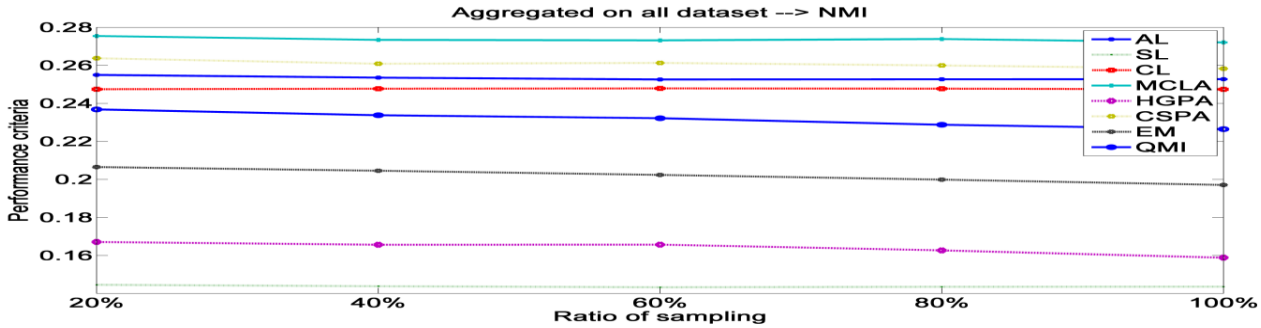
Figure.4. as you can see the number of partition, performance does not improve. It is very important that the number of partitions does not always improve performance because one of our goals is to find the point where the combination of clustering to increase the number of partitions and will actually improve performance is not meeting the number of partition 10 to 20 will be.



Accuracy



Fisher Method

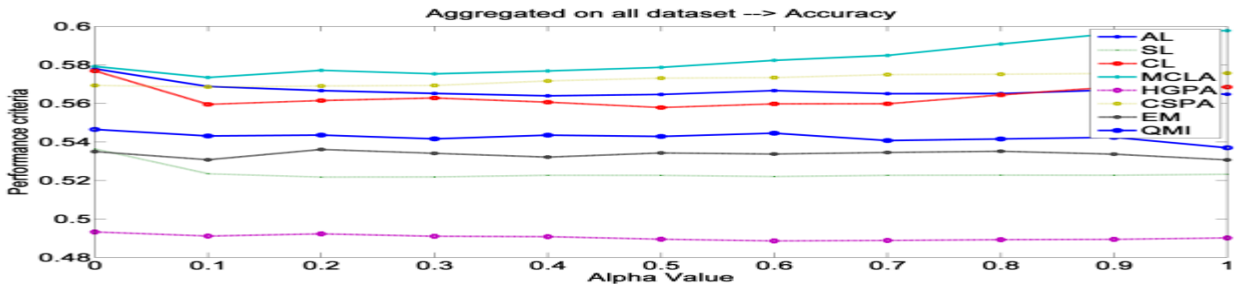


normal Mutual Information

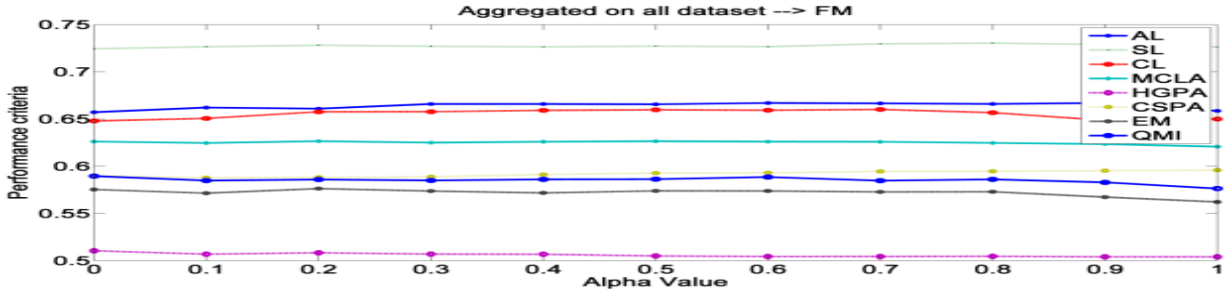
figure5

average performance clustering resent on all data sets deal with a variety of different function depending on the number of partition sampling rate – curve horizontally in the assembly shows the number of partitions and vertical curve shows the quality of the final partition.

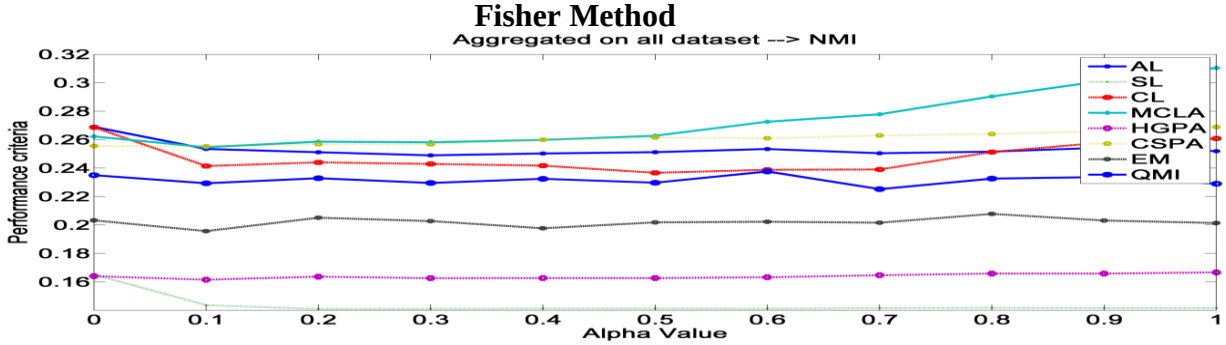
Figure5, average performance clustering resonant on all data sets deal with verity of different functions in different sampling rate shows. Fig.5. as well as you can see by increasing the sampling rate, steadily performance does not improve that is actually the best sampling rate will be 20%. The sampling rate increase will the running time of the time however the dividend clustering will be intensive.



Accuracy



Fisher Method



normal Mutual Information

Figure6

Average Performance Clustering Resent On All Data Sets Deal With A Variety Of Different Function Depending On The Number Of Partition Update Rate – Curve Horizontally In The Assembly Shows The Number Of Partitions And Vertical Curve Shows The Quality Of The Final Partition. Figure6, Average Performance Clustering Resonant On All Data Sets Deal With A Variety Of Different Functions In Different Learning Rate Shown. Figure6, Unlike Previous Forms As You Can See By Increasing The Rate Of Learning, Performance Steadily Improved Almost As Long To Reach The Learning Rate .9. If The Learning Rate .9, In Addition, It Is Noteworthy That The Increase Rate Of The Learning Rate Will Increase Running Time. From Fig(4,5) And 6, Will Know That Meta Cluster Method(MCLA) With Sampling Rates 0.2, Learning Rate 0.9 And The Number Of Partitions Between 5 To 10 With The Best Performance Base Clustering Algorithm K-Means Produce.

CONCLUSION

Experimentally shown meta clustering algorithm favorable impact on the efficiency of the final partition and an outstanding result in clustering ensemble, the number of partitions and sampling rate is inversely related to the performance of the final partition . In addition, the experimental results show that the proposed framework , comparable to (and even sometimes better than) the best combination is based clustering methods .

Future Works

As future work can be changed in such a way that the weighted fuzzy clustering and the contributions involved in the manufacture of various weights vary. Because data loss may be forced into a cluster weights to ruffle it , and if a mechanism for automatically preventing the problem was , was very helpful.

REFRENCE

1. Dempster A.P., Laird N.M., and Rubin D.B.,(1997). *Maximum likelihood from incomplete data via the EM algorithm*, J, Royal Statistical Society B(39)1:1-38.

2. Dy J.(2015). Unsupervised feature selection. *Computational Methods of FeatureSelection*, pages 19-39.
3. Anderberg M. R.(2010). *Cluster Analysis for Applications*, Academic Press, Inc., New York,
4. AZad L. A.,(1965), *Fuzzy Sets*, Information and Control, 8:338-353.
5. Dash M., Liu H.(1997). *Feature Selection for Classification*. Intelligent Data Analysis 1:131-136
6. *classification*, Machine Learning, Vol. 61, No. 3, pp. 129 – 150**.**
7. Ghanbarzadeh A.. Pham D. T., , Koc E., Otri S., Rahim S., Zaidi M..(2006). *Intelligent Production Machines and Systems*.
8. Gower J. C. and Ross G. J. S.,(2010). *Minimum Spanning Trees and Single-Linkage Cluster.*
9. Gu Q., Li Z., and Han J.(2012). *Generalized _sher score for feature selection*.arXiv preprint arXiv:1202.3725.
10. Guyon I. and Elissee A. (2013*). An introduction to variable and feature selection*. Journal of Machine Learning Research, 3:1157-1182.
11. Guyon I., Weston J., Barnhill S., and Vapnik V.(2012). *Gene selection for cancer classification using support vector machines*. Machine learning, 46(1):389-422.
12. Toghraee M, rad F, parvin H,.(2016). *THE impact of feature selection on meta heuristic algorithm to data mining methods.* International journal of modern education and computer science. Volum 8; issue 10., page(33).
13. Toghraee M, rad F, parvin H,.(2016). *Evaluation of meta heuristic algorithm for stable feature selection.* I.J.iformation technology and computer science (ijitcs),.volum8. issue:2074-9015; pp: 22-29.
14. Toghraee M, rad F, parvin H,.(2016). *Effect neural networks on selected feature by meta heuristic*. i.j. mathematical science and computing (ijmcs).volum2.issue:2310-9033.pp:41-48.
15. Toghraee M, Esmaeili M , parvin H,.(2016).*evaluation neural networks on selected feature by meta heuristic algorithms*. Artificial intelligent system and machine learning.volum8.pp:108-115.
16. Toghraee M, rad F, parvin H,.(2017). Evaluation average total data set learning machine on the meta heuristic algorithm. International journal of emerging trend& technology in computer science .volum6.issue:2278-6856.page(7).
17. Neumann J, C and Schnar G. S. (2011*).Combined SVM-based feature selection and*