International Journal of Database Theory and Application
Vol.8, No.1 (2015), pp.11-22
http://dx.doi.org/10.14257/ijdta.2015.8.1.02

A Clustering Based Study of Classification Algorithms

Muhammad Husnain Zafar[1] and Muhammad Ilyas[2]

[1]Dept. of Computer Science and Information Technology, University of Sargodha Sargodha, Punjab, Pakistan
[2]Assistant Professor
Dept. of Computer Science and Information Technology, University of Sargodha Sargodha, Punjab, Pakistan
[1]husnainzafarmscs@gmail.com, [2]m.ilyas@uos.edu.pk

Abstract

A grouping of data objects such that the objects within a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups. Many of clustering algorithm is available to analyze data. This paper intends to study and compare different clustering algorithms. These algorithms include K-Means, Farthest First, DBSCAN, CURE, Chameleon algorithm. All these algorithms are compared on the basis of their pros and cons, similarity measure, their working, functionality and time complexity.

Key Words: *Clustering algorithms, k-means, farthest first, DBSCAN, CURE, chameleon algorithm*

1. Introduction

In recent years, the dramatic rise in the use of the web and the improvement in communications in general have transformed our society into one that strongly depends on information. The huge amount of data that is generated by this communication process contains important information that accumulates daily in databases and is not easy to extract. The field of data mining developed as a means of extracting information and knowledge from databases to discover patterns or concepts that are not evident [1].

One of the most complexes, well-known and most studied problems in data mining theory is clustering. This term refers to the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [2].

In this paper we study and compare five different clustering algorithms. These algorithms include K-Means, Farthest First, DBSCAN, CURE, Chameleon algorithm. All these algorithms are compared on the basis of their pros and cons, similarity measure, their working, functionality and their time complexity.

Rest of the paper compromise the following sections. Section II describes related work. Section III describes all the algorithms individually. Section IV describes the comparison of these algorithms. Section V compromises the evaluation and results of these algorithms. Section VI describes conclusion and future work.

2. Related Work

N. Sharma *et al.*, [8] present the comparison between different clustering algorithms. All these algorithms are implemented in Weka Tool. The aim of their study is to show that which algorithm is more suitable. These algorithms include DBSCAN, EM, Farthest First, OPTICS and K-Means algorithm. In this paper they show the advantages and

ISSN: 2005-4270 IJDTA

disadvantages of each algorithm but on the basis of their research they found that k-means clustering algorithm is simplest algorithm as compared to other algorithms.

K. H. Raviya and K. Dhinoja [9] introduce clustering technique in the field of Data Mining. They defined Data Mining and Clustering Technique. Data mining is used in many fields like fraud detection, AI, information retrieval, machine learning and pattern recognition *etc*. They compare two clustering algorithms K-Means and DBSCAN. They did the comparison on the basis of time, cost and effort. Their aim is to provide best technique and provide the fruitful comparison so that best clustering algorithm will be chosen. Their result shows that K-Means is better than DBSCAN algorithm.

T. Kinnunen *et al.*, [10] works on speaker identification. Speaker recognition is a generic term used for two related problems: speaker identification and verification [11]. In this research they study the role of vector quantization. Different clustering algorithms are compared and influenced to check which one is good and which algorithm gives the improvement in accuracy of speaker recognition. The result shows that Iterative splitting technique (SPLT) gives good result when database is very large otherwise Randomized local search (RLS) is best.

O. J. Oyelade *et al.*, [12] use K-Means algorithm in the field of education. They develop a system for analyzing students' results based on cluster analysis and uses standard statistical algorithms to arrange their scores data according to the level of their performance is described. K-Means algorithm is used to analyze student result data. N. V. A. Kumar and G. V. Uma [13] also apply different Data Mining Techniques for improving academic performance of students.

3. Clustering Algorithms

We select five clustering algorithms for comparison. Their one by one individual description is given below.

3.1 K-Means Algorithm

K-means clustering is a method of classifying/grouping items into k groups (where k is the number of pre-chosen groups). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. The steps in K-Means algorithm are given below [3, 9].

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Advantages of K-means are given below.

- With a large number of variables, K-Means may be computationally faster than hierarchical clustering.
- K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Disadvantages of K-means are given below.

- Difficulty in comparing quality of the clusters produced (*e.g.*, for different initial partitions or values of K affect outcome).
- Fixed number of clusters can make it difficult to predict what K should be.
- Does not work well with non-globular clusters.

- K-Means will not identify outliers
- Complexity is O(n * K * I * d)
    - n = number of points,
    - K = number of clusters,
    - I = number of iterations,
    - d = number of attributes

3.2 Farthest First Algorithm

Farthest first is a Variant of K means that places each cluster center in turn at the point furthest from the existing cluster centers.

This point must lie within the data area. This greatly sped up the clustering in most cases since less reassignment and adjustment is needed [8].

Advantages of Farthest First are given below.

- Farthest-point heuristic based method has the time complexity O (nk), where n is number of objects in the dataset and k is number of desired clusters. Farthest-point heuristic based method is fast and suitable for large-scale data mining applications.

Disadvantages of Farthest First are same as K-means algorithm.

3.3 DBSCAN Algorithm

DBSCAN is density based clustering algorithm. The main concept of density based algorithm is given below [4, 9].

- In this method clustering is based on density such as density connected point.
- Each cluster has a considerable higher density of points than outside of the cluster
- Two global parameters:
    - Eps: Maximum radius of the neighborhood
    - MinPts: Minimum number of points in an Eps-neighborhood of that point
- Core Object: object with at least MinPts objects within a radius 'Eps-neighborhood'
- Border Object: object that on the border of a cluster.

Steps of DBSCAN are given below [5].

- Arbitrary select a point p
- Retrieve all points density-reachable from p wrt Eps and MinPts.
- If p is a core point, a cluster is formed.
- If p is a border point, no points are density reachable from p and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

Advantages of DBSCAN are given below.

- DBSCAN does not require you to know the number of clusters in the data a priori, as opposed to k-means.
- DBSCAN can find arbitrarily shaped clusters. It can even find clusters completely surrounded by (but not connected to) a different cluster.
- DBSCAN has a notion of noise.

Disadvantages of DBSCAN are given below.

- DBSCAN can only result in a good clustering as good as its distance measure is. The most common distance metric used is the euclidean distance measure.

Especially for high-dimensional data, rendering it hard to find an appropriate value for epsilon.

- DBSCAN cannot cluster data sets well with large differences in densities, since the minPts-epsilon combination cannot be chosen appropriately for all clusters then.
- The Algorithm is not partition able for multi-processor systems.

3.4 CURE Algorithm

CURE is an agglomerative algorithm in the hierarchical method which builds clusters gradually. The steps of CURE algorithm are given blow [6].

- Starts with each input as separate cluster and each successive step merge the closest pair of clusters.
- C representative points are stored to compute the distance between a pair of cluster.
- These are determined by first choosing C well scattered points with in the cluster and then shrinking them towards center of the cluster by a fraction α.
- Representative points of cluster are used to compute its distance from other clusters.

Representative points attempts to capture the physical shape and geometry of cluster. Shrinking the points towards center mitigates the effects of outlier. Larger movement in the outlier reduces their ability to cause the wrong cluster to be merged.

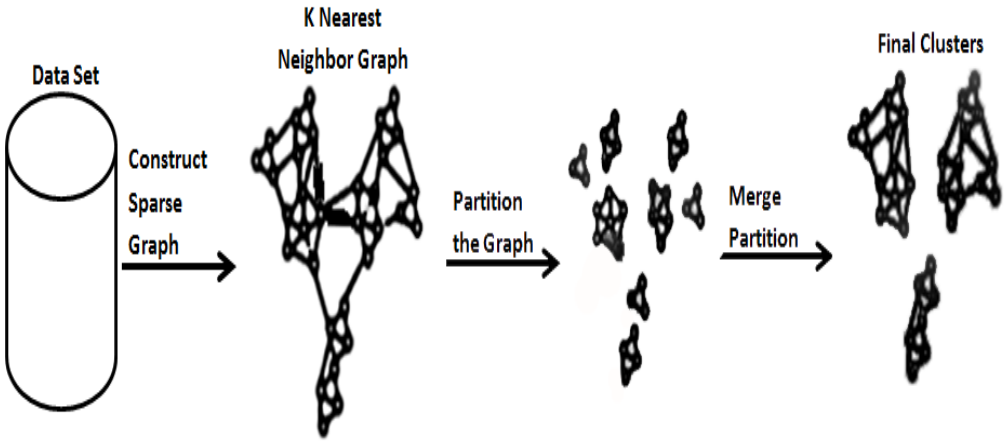

Figure 1. Chameleon Framework

Advantages of CURE algorithm is given below.

- The clustering algorithm can recognize arbitrarily shaped clusters.
- The algorithm is robust to the presence of outliers.
- It appropriate for handling large data sets.

Disadvantages of CURE algorithm is given below.

- CURE ignores the information about the aggregate inter-connectivity of objects in two clusters. So it is introduced Chameleon algorithm.

3.5 Chameleon Algorithm

CHAMELEON operates on a sparse graph in which nodes represent data items, and weighted edges represent similarities among the data items. This sparse graph

representation of the data set allows CHAMELEON to scale to large data sets and to operate successfully on data sets that are available only in similarity space and not in metric spaces. CHAMELEON finds the clusters in the data set by using a two phase algorithm [7].

Phase -I

In first phase graph partitioning algorithm is used to divide the data set into a set of individual clusters.

Phase -II

In Second Phase agglomerative hierarchical mining algorithm is used to merge the clusters.

Figure 1 show the overall framework of CHAMELEON algorithm. The similarity measured in this algorithm is based on two parameters. These are relative inter-connectivity (RI) and relative Closeness (RC). Chameleon selects pairs to merge for which both RI and RC are high. Advantages are given below.

- There are several advantages of representing data using a k-nearest neighbor graph Gk.
- Firstly, data points that are far apart are completely disconnected in the Gk. Secondly; Gk captures the concept of neighborhood dynamically.

4. Comparison of Algorithms

Table I shows the comparison between K-means, Farthest First and DBSCAN algorithm.

Table 1. Comparison of K-means, Farthest First and DBSCAN Algorithm

Comparison Parameters	K Means Algorithm	Farthest First Algorithm	DBSCAN
Working & functionality	K means algorithm can be built by classify data to groups of objects (objects based clustering grouping) based on their attributes (attribute based clustering grouping)/ features in to k number of groups.	Farthest First Algorithm also build by classify data to groups of objects based on their attributes/ features in to k number of groups.	DB (Density Based Grouping) scan is based on the density reachability and density connectivity Where dense area objects separated by less dense areas.

	This algorithm partitions objects in a data set into a fixed number of K disjoint subsets. For each cluster, the partitioning algorithm maximizes the homogeneity	This is the variation of K Means algorithm. The Only Difference is that, in Simple K Means the Centre of cluster is initially selected may or may not from the data point but in Farthest First algorithm center of cluster is selected within the data points. Remaining working is same.	In this algorithm objects are initially assumes to be unassigned. Db scan then chooses arbitrary object p from data set .if it finds p as a core object then finds all density connected objects bases on eps and mints and if p is not a core object then considered as a noise and move to next object.
Advantages	It is simple to implement. It is very fast to execute and scalable. It works well for Euclidian data. Its convergence can be done to local minima not global minimum.	It is simple to implement. It is very fast to execute and scalable. It works well for Euclidian data. Its convergence can be done to local minima not global minimum. It speed up clustering process and there are less assignments and adjustment are needed.	It can find arbitrarily shaped clusters and also find clusters completely surrounded by different clusters. It is robust to noise. There is no need of any k deterministic. It requires just two points which are very insensitive to the ordering of the points in the database.
Disadvantages	It works only for well-shaped clusters and it is sensitive to outlier (noise). There is a need to know the pre-defined value of k.	It works only for well-shaped clusters and it is sensitive to outlier (noise). There is a need to know the pre-defined value of k.	In this algorithm datasets with varying densities are problematic and it requires connected region of sufficiently high density
Similarity Measure	Use Euclidean distance formula to find distance.	Use Euclidean distance formula to find distance.	Use Euclidean distance formula to find distance.
	Use centroid approach to find similarity	Use centroid approach to find similarity	Use density-connected points to find similarity.
Time Complexity	O(n * K * I * d) n = number of points, K = number of clusters, I = number of iterations, d = number of attributes	O(nk)	O(m) m = number of points

Table 2 shows the comparison between CURE and CHAMELEON algorithm.

Table 2. Comparison between CURE and CHAMELEON Algorithm

Comparison Parameters	CURE	Chameleon
Working & functionality	It first partitions the random sample and partially clusters the data points in each partition. After eliminating outliers, the pre clustered data in each partition is then clustered in a final pass to generate the final clusters	It is based on two phases: at first partitions the data points into sub-clusters, then repeatedly merging sub-clusters, com from previous stage to obtain final clusters
	Firstly it will choose C representative points that are well scattered and then shrink them by a fraction towards center of cluster. Shrinking the points towards center reduces the outliers. These representative points are used to compute its distance from other clusters. These Representative points attempt to capture the physical shape and geometry of cluster.	In first phase, it uses the KNN to construct the sparse graph. After this Partition the KNN graph such that the edge cut is minimized because Since edge cut represents similarity between the points, less edge cut => less similarity. Chameleon selects pairs to merge for which both Relative inter-connectivity (RI) and Relative closeness (RC) is high.
Similarity Measure	Representative points are used to measure similarity	Relative inter-connectivity (RI) and Relative closeness (RC) of the clusters are used to measure similarity.
	It uses static model	It uses dynamic model because it considers the internal characteristics of the clusters themselves
Advantages	This clustering algorithm can recognize arbitrarily shaped clusters and it is robust to the presence of outliers. It is appropriate for handling large data sets	This algorithm is proven to find clusters of diverse shapes, densities, and sizes in two-dimensional space
Disadvantages	CURE ignores the information about the aggregate inter-connectivity of objects in two clusters	CHAMELEON is known for low dimensional spaces, and was not applied to high dimensions [14].
Time Complexity	This algorithm uses space that is linear in the input size n and has a worst-case time complexity of O(n2 logn). For lower dimensions (e.g., two), the complexity can be shown to further reduce to O(n2)	For large n, the worst-case time complexity of the algorithm is O(n(log2 n + m)), where m is the number of clusters formed after completion of the first phase of the algorithm Time complexity of CHAMELEON algorithm in high dimensions is O(n2)

On the basis of some characteristics we draw a graph that shows that either a characteristic is present in these algorithms are not. This is shown in the Figure 2. In this

graph X-Axis shows the characteristics that are present in these algorithms. Y-Axis shows that which characteristic is present in which algorithm.

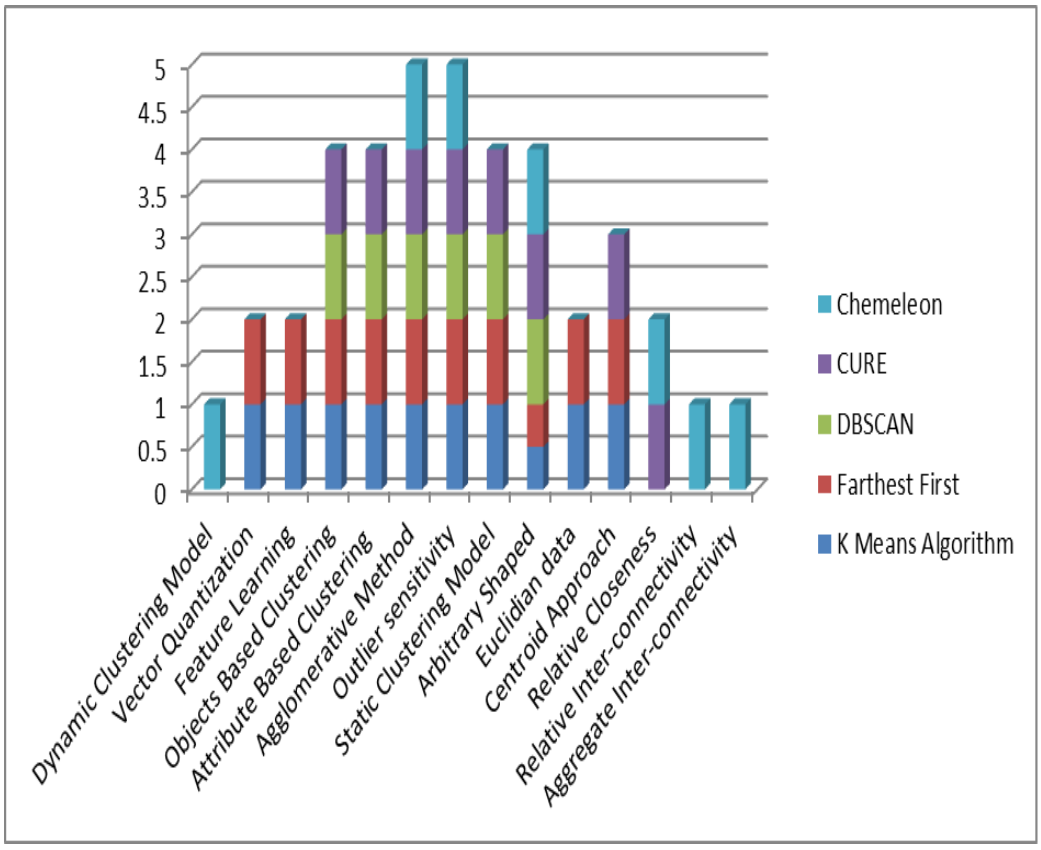

Figure 2

5. Evaluation and Results

For implementation of these algorithms, we choose Weka tool. K-Means, Farthest First and DBSCAN are implemented in Weka tool. For getting the results we choose data set. Dataset belongs to wine. It consists of 178 instances and 14 attributes. These attributes are Class, Alcohol, Malic Acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline.

On this data set we applied all three algorithms. The result of Farthest First, K-Means and DBSCAN is shown in the Table 3, 4 and 5 respectively.

The result of these algorithm shows that K-Means and Farthest First takes less time to execute as compared to DBSCAN. Result of DBSCAN also shows that 1 instance remains un-cluster but in other two algorithms no instance remains un-cluster. Both have disadvantage. In K-Means and Farthest First we have to set the value of K. K is the number of clusters to be build and in DBSCAN we have to set the values of Eps and MinPts. From this result it is considered that Farthest First is much better than two other algorithms. We can also run these algorithms on another data set and that data set also shows the same result. Name of dataset is Vowel. It consists of 990 instances and 14 attributes. These attributes are train or test, Speaker number, Sex, Feature 0 to Feature 9 and Class.

Table 3. Farthest First Algorithm Results

Cluster centroids:													
Cluster 0													
	3	13.7	4.36	2.2	22.5	88	1.28	.47	.52	1.1	6.6	.78	1.7	520
Cluster 1													
	2	11.56	2.05	3.23	28.5	119	3.18	5.08	0.47	1.87	6	0.93	3.69	465
Cluster 2													
	1	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547
Time taken to build model (full training data) : 0.01 seconds													
Clustered Instances													
0	64	(	36%	)									
1	55	(	31%	)									
2	59	(	33%	)									

Table 4. K-Means Algorithm Result

Number of iterations: 5				
Within cluster sum of squared errors: 49.998510705570595				
Missing values globally replaced with mean/mode				
Attribute	Full Data	0	1	2
	178	59	48	71
==========				
class	2	1	3	2
Alcohol	13.00	13.7	13.1	12.27
Malic_acid	2.336	2.01	3.333	1.9327
Ash	2.366	2.45	2.437	2.2448
Alcalinity_of_ash	19.49	17.0	21.41	20.238
Magnesium	99.74	106	99.31	94.549
Total. Phenols	2.295	2.84	1.678	2.2589
Flavanoids	2.029	2.98	0.781	2.080
Nonflavanoid_phenols	0.361	0.29	0.447	0.363
Proanthocyanins	1.590	1.89	1.15	1.630
Color_intensity	5.058	5.52	7.39	3.086
Hue	0.957	1.06	0.68	1.056
OD315_of_diluted_wines	2.6117	3.1578	1.6835	2.7854
Proline	746.8	1115	629	519.5

Time taken to build model (full training data) : 0.06 seconds	
Clustered Instances	
0 59 (33%)	
1 48 (27%)	
2 71 (40%)	

Table 5. DBSCAN Algorithm Result

Clustered DataObjects: 178		
Number of attributes: 14		
Epsilon: 0.9; minPoints: 6		
Index: weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase		
Number of generated clusters: 3		
Elapsed time: .1		
Time taken to build model (full training data) : 0.1 seconds		
Clustered Instances		
0	59	(33%)
1	70	(40%)
2	48	(27%)
Unclustered Instances: 1		

We evaluate these algorithms on other data sets. On some data sets DBSCAN does not cluster any instance but other two algorithms cluster that instances. This dataset belongs to Car. It consists of 1728 instances and 7 attributes. These attributes are buying, maint, doors, persons, lug boot, safety and class. The result of DBSCAN shows the following results.

- Time taken to build model (full training data): 2.94. Seconds
- Clustered Instances: 0
- Unclustered Instances: 1728

From the above results, we can see that all 1728 instances of dataset remains un-cluster but K-Means and Farthest First cluster all the instances on this data set. So K-means and Farthest First are much better than DBSCAN but Farthest First takes less time to execute so it is better than K-means algorithm.

We did comparison of all five algorithms theoretically. Result of theoretical analysis shows that Chameleon algorithm is better than all other algorithms because it removes all the problems that occurs in other algorithms. Disadvantages of K-Means and Farthest First are solved by DBSCAN and problems of DBSCAN are resolved by CURE algorithms. Similarity drawback of CURE is removed by Chameleon algorithm.

6. Conclusion and Future Work

In his paper we present brief and easy comparison between different clustering algorithms. We also evaluate these algorithms on different datasets and present the results through tables. The aim of this comparison is the selection of good clustering algorithm as clustering technique is most popular technique for the classification of objects and used in many fields like Biology, Libraries, Insurance, and Marketing *etc*.

In Future we will implement and evaluate CURE and Chameleon algorithm and will be able to present the results of these algorithms with practical examples. After this we will compare all these five algorithms with practical examples and find that which algorithm is best among these five algorithms.

References

[1] G. Fung, "A Comprehensive Overview of Basic Clustering Algorithms", **(2001)**.
[2] N. T. Linh and C. C. Chau, "Application of CURE Data Clustering algorithm to Batangas State University Student Database", International Journal on Advances in Computing and Communication Technology, **(2013)**, pp. 108-115.
[3] J. Gao and D. B. Hitchcock, "James-Stein Shrinkage to Improve K-means Cluster Analysis", **(2009)** November 30.
[4] S. Kisilevich, F. Mansmann and D. keim, "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos", Computing of Geospatial Research and Applications., NY, USA, **(2010)**, June 21.
[5] R. Luo and C. Zaniolo, "DBSCAN & Its Implementation on Atlas", [Online]. Available: http://www.wis.cs.ucla.edu/wis/atlas/doc/dbscan.ppt, **(2002)** June.
[6] S. Guha, R. Rastogi and K. Shim, "CURE: an efficient clustering algorithm for large databases", International Conference on MOD, New York, USA, **(1998)** June.
[7] G. Karypis, E. Han and V. Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", IEEE Computer, **(1999)**, pp. 68-75.
[8] N. Sharma, A. Bajpai and R. Litoruya, "Comparison the various clustering algorithms of weka tools", International Jornal of Emerging technology and Advanced Engineering, vol. 2, no. 5, **(2012)** May.
[9] K. H. Raviya and K. Dhinoja, "An Empirical Comparison of K-Means and DBSCAN Clustering Algorithm", PARIPEX Indian Journal of Research, vol. 2, no. 4, **(2013)** April, pp. 153-155.
[10] T. Kinnunen, T. Kilpelainen and P. Franti, "Comparison OF Clustering Algorithms in Speaker Identification, **(2011)**.
[11] S. Furui, "Recent advances in speaker Recognition", Pattern Recognition Letters, **(1997)**, pp. 859-872.
[12] O. J. Oyelade, O. O. Oladipupo and I. C. Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", IJCSIS, USA, vol. 7, no. 1, **(2010)** January.
[13] N. V. A. Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique", European Journal of Scientific Research, vol. 34, no. 4, **(2009)**.
[14] M. K. Rafsanjani, Z. A. Varzaneh and N. E. Chukanlo, "A survey of hierarchical clustering algorithms", TJMCS, vol. 5, no. 3, **(2012)** December, pp. 229-240.

Authors

Muhammad Husnain Zafar, is a student of the Master of Science in Computer Sciences at the University of Sargodha. He has already completed his BS 4 year degree in computer sciences. His research interests include Software Engineering, Software Reusability and other topics.

Muhammad Ilyas, received a Master degree in Software Project Management in 2004 from National University of Computer and Emerging Sciences, Lahore and a Doctor of Informatics from Johannes Kepler University, Linz Austria in 2010. His research interests include Software Engineering, Design Pattern and knowledge base systems. He is currently an assistant professor in the Department of Computer Science and Information Technology at the University of Sargodha, Pakistan.

