International Journal of Scientific Research in Engineering
www.ijsrejournal.com
Volume 3 Issue 1ǁ Jan.- June 2018 ǁ PP. 01-08

Comparative Analysis of Diabetes Characteristics through Various Clustering Algorithms

Shubhangi Pahwa
Department of Computer Science DAVIET, Jalandhar
Corresponding Author: Shubhangi Pahwa

Abstract: *This study aims at finding the features that can determine the presence of diabetes and also can find out the quantity of women suffering from diabetes. The functionalities of data mining like attribute oriented induction and clustering have been used to track the features of women suffering from diabetes. Data related to this research was taken from National Institute of Diabetes, Digestive and Kidney Diseases. As clustering technique is used, the results are shown in the form of clusters. The clusters show the percentage of women suffering from diabetes and concentration of various characteristics with such characteristics. The results were accessed in five clusters which show that 22 % of women suffering from diabetes reside in cluster-0, 5% reside in cluster 1, 24% reside in cluster 2, 8% reside in cluster 3 and 25% reside in cluster 4. It was found that for each cluster the characteristics seem to vary. It can be interpreted from the results that the characteristics of women suffering from diabetes are different with respect to a cluster and no resemblance can be found with respect to other clusters. This research helps in forecasting the state of diabetes that if it is in an initial stage or it is in an advanced stage. This study also helps to estimate the maximum number of women that are suffering from diabetes with specific characteristics. By diagnosing the characteristics the patients can be given effective treatment.*

Keywords: *Data Mining, Clustering, K-means, K-medoid, KNN, MST*

Date of Submission: 25-06-2018 Date of acceptance: 10-07-2018

I. INTRODUCTION

A. Clustering:

Clustering is a data mining technique that groups a set of objects in such a way that object in the same clusters are more similar to each other than those in other groups. A cluster of objects can be considered as a group in many applications. With the help of clustering one can recognize the dense and the sparse regions which lead to discover interesting correlations and overall distribution patterns among the data attributes [1].

The performance of clustering is reliant on the detection of quality clusters of data item that is why most of the algorithms aim at determining the set of clusters efficiently for a given relational database or transaction [2]. The main problem is to generate clusters in the database. Therefore various algorithms were introduced that intent at generating quality clusters. These algorithms were different in the context of creating the clusters.

B. Clustering Applications:

Clustering can be used as a tool to observe the characteristics of every cluster, to understand the distribution of data and also to concentrate on a specific set of cluster for further examination.

1. Analyzing clusters has been used in various application domains like medicine, biology, anthropology, market research and economics.
2. The applications of clustering includes pattern recognition, image processing, animal and plant classification, data analysis, disease classification and so on.
3. Recently clustering has been used for analyzing web log data to identify usage patterns.

Clustering can also act as a preprocessing stage for other algorithms like classification and characterization which is then applied on the clusters formed.

II. MATERIAL AND METHODS:

The analysis of the characteristics is implemented taking into consideration the data acquired from National Institute of Diabetes, Digestive and Kidney Diseases [3]. Firstly the data is grouped together into clusters using various clustering techniques. After creating the clusters their quality is required to be identified as clusters with poor quality are not useful in defining the characteristics efficiently. Also the estimation of the cluster quality helps in figuring out the best algorithm that forms clusters of good quality. Therefore the characteristics of the data available can be effectively assessed by implementing Attribute Oriented Induction [4] technique for the clusters created by identified algorithm.

Figure 1 represents the procedure for defining the characteristics of the diabetes data. Firstly the data is provided as input to the normalization algorithm [5] that helps in creating the normalized file. Then the normalized file is provided to the appropriate clustering algorithm as input to create the set of clusters. The best algorithm that creates good quality clusters is determined by calculating the quality of clusters created by respective clustering algorithms.

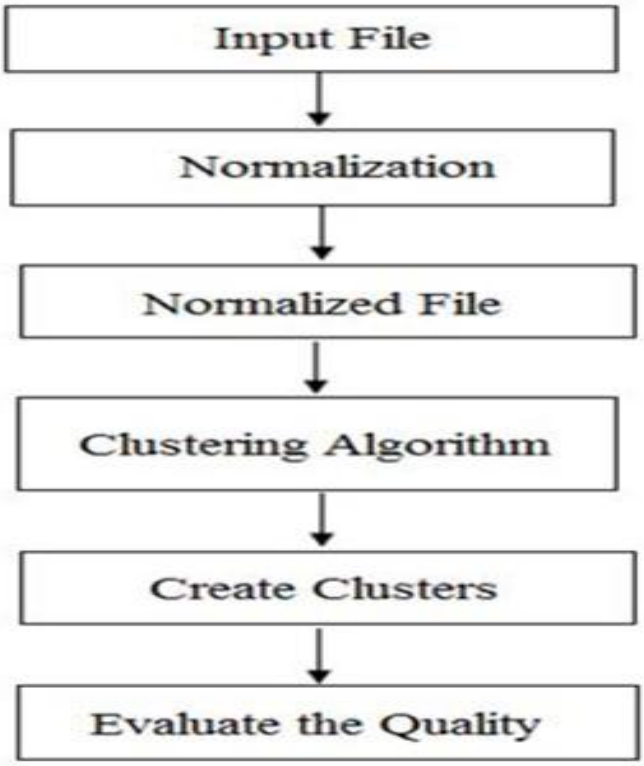


Figure 1: Cluster Formation and Quality Estimation Process

The quality of clusters can be denoted by their diameter and the maximum distance between any of the two objects in cluster [6]. The alternative measure for cluster quality is the centroid distance. Centroid distance is defined as the average distance between centroid and each object.

In this study four algorithms are considered for creating the clusters that are Minimum Spanning Tree (MST), K-Means [7], Nearest Neighbor and Partitioning Around Medoids (PAM). The quality of the clusters generated by these algorithms is used to identify the best algorithm [8] among these four algorithms.

After identifying the best algorithm we can evaluate the characteristics of the clusters using Attribute Oriented Induction technique. In this technique initially the number of attributes with distinct values is identified. After this, the attributes with maximum number of distinctive values are removed.

From rest of the attributes minimum and maximum values are identified. The objects are then grouped together by using the conception of set grouping by taking into consideration some threshold value.

III. RESULTS AND DISCUSSION

For identifying the best algorithm among the four the quality of the clusters can be compared graphically. Table I represents the outcomes of quality for the four given algorithms for different number of clusters alongside the graph. It can be determined from the graph that Partitioning Around Medoids (PAM) gives the clusters of better quality. Thus it can be considered for analyzing the characteristics of given data. The diabetes data is categorized in two classes that are the normal class which is consisted of normal data and the target class which is consisted of tested positive.

TABLE I: Results of Algorithms for different number of clusters

S.No.	ALGORITHM	Number of Clusters				
		5	10	13	19	25
1.	k-means	0.385	0.168	0.125	0.078	0.059
2.	k-medoid	0.379	0.162	0.120	0.074	0.056
3.	KNN	0.504	0.225	0.177	0.124	0.094
4.	MST	0.472	0.222	0.170	0.112	0.079

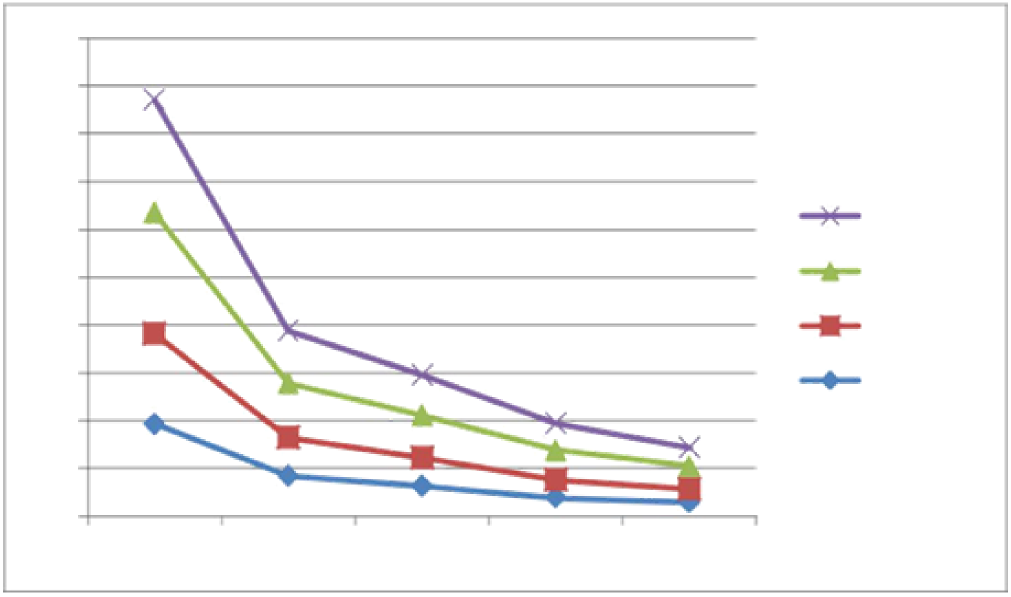


Fig 2: Result of Characterization

The target class data is then given to Attribute Oriented Induction as input for defining the characteristics. The target class comprises of 268 women that are actually under the impact of diabetes. Though it is unidentified that in what stage diabetes is.

However, it was observed after characterization that that 22 % of women suffering from diabetes reside in cluster-0, 5% reside in cluster 1, 24% reside in cluster 2, 8% reside in cluster 3 and 25% reside in cluster 4. From this it can be forecasted that if the disease is in an initial state or an advanced state. The fig 2 represents the results of characterization.

Therefore this research will not only help in evaluating the maximum number of people affected from diabetes with particular characteristics but also helps in recognizing the state of disease.

IV. ANALYSIS OF RESULTS OF CHARACTERIZATION

A. *Cluster- 0 Characteristics*

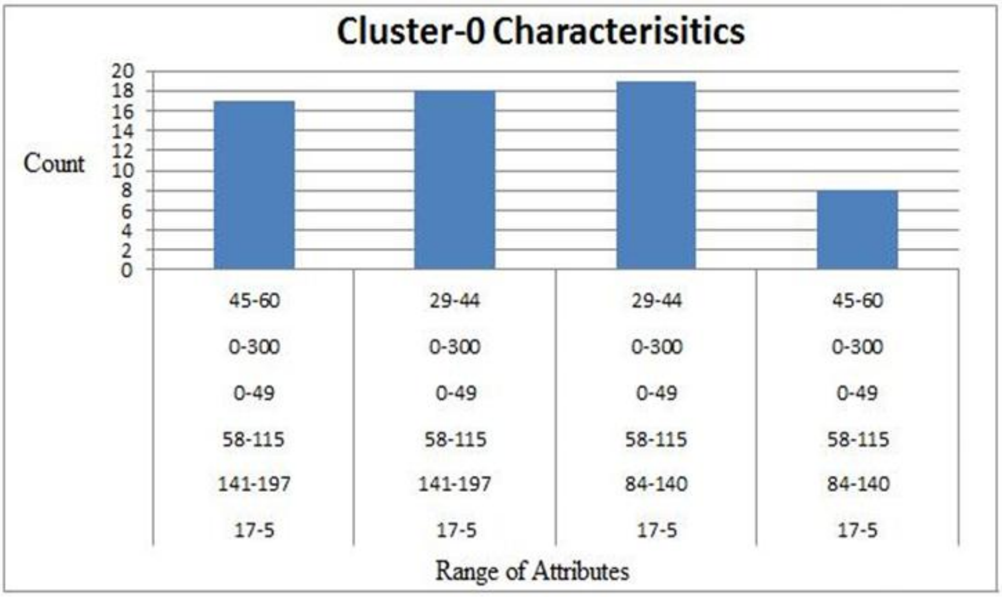


Fig 3: Cluster-0 characteristics ***B. Cluster- 1 Characteristics***

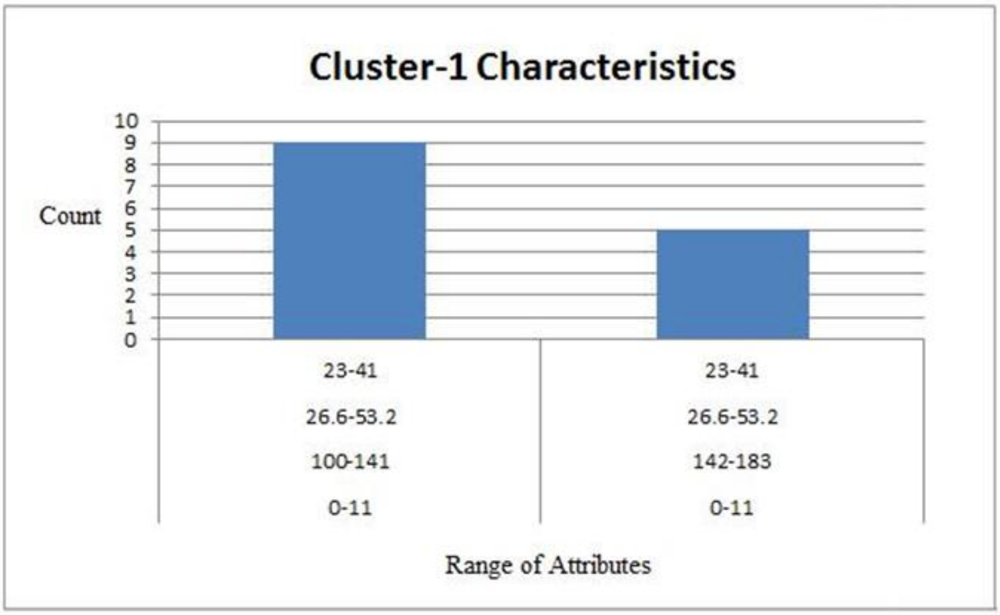


Fig 4: Cluster-1 Characteristics

C. Cluster- 2 Characteristics

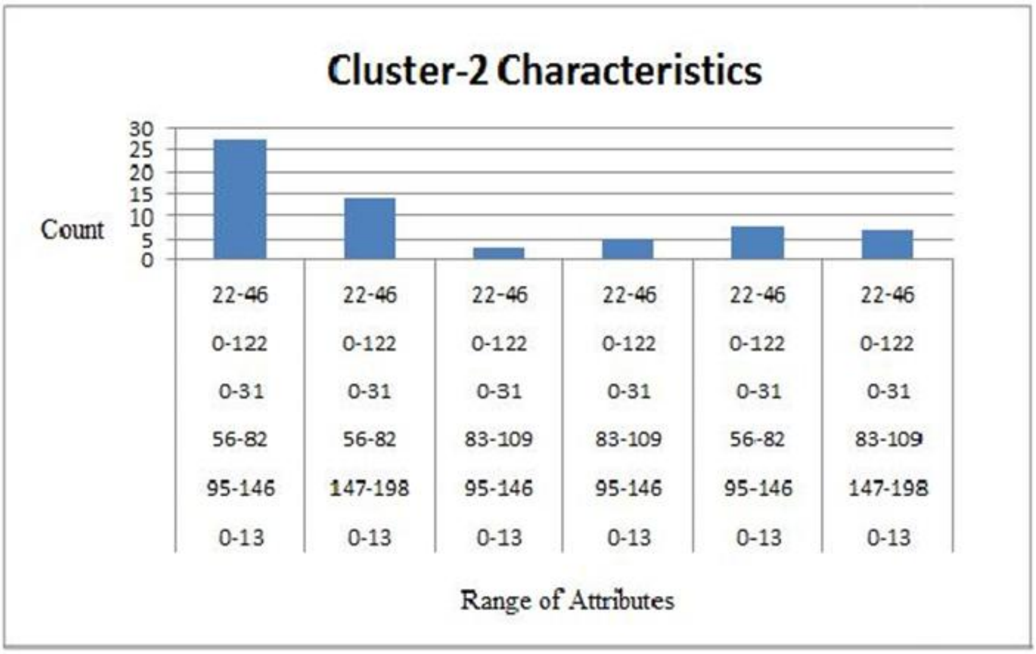


Fig-5: Cluster 2 Characteristics ***D. Cluster-3 Characteristics***

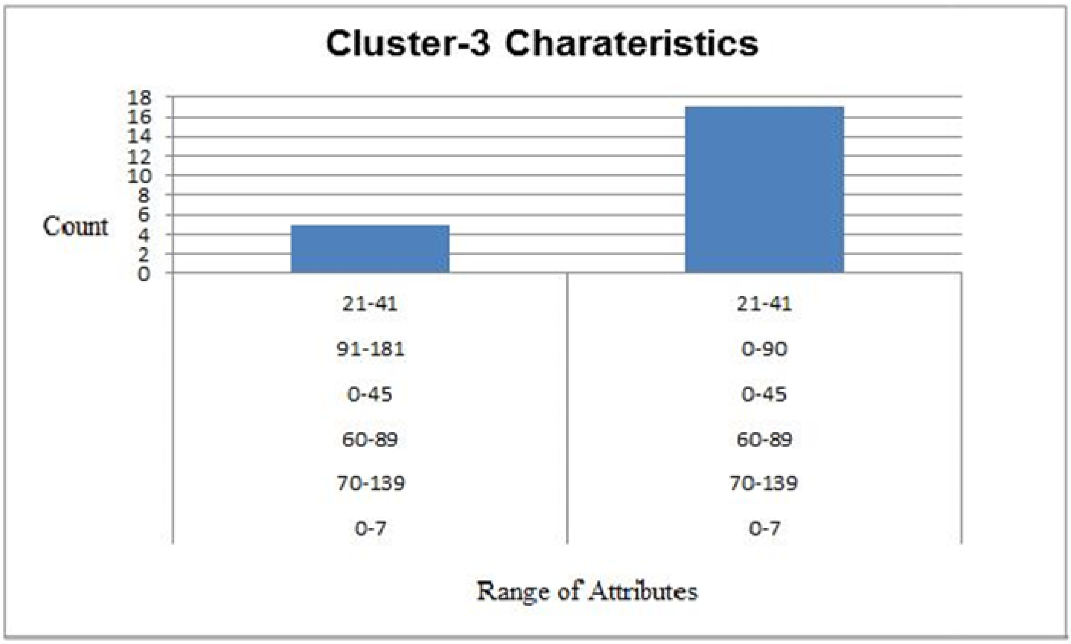


Fig 6: Cluster 3 Characteristics

E. Cluster- 1 Characteristics

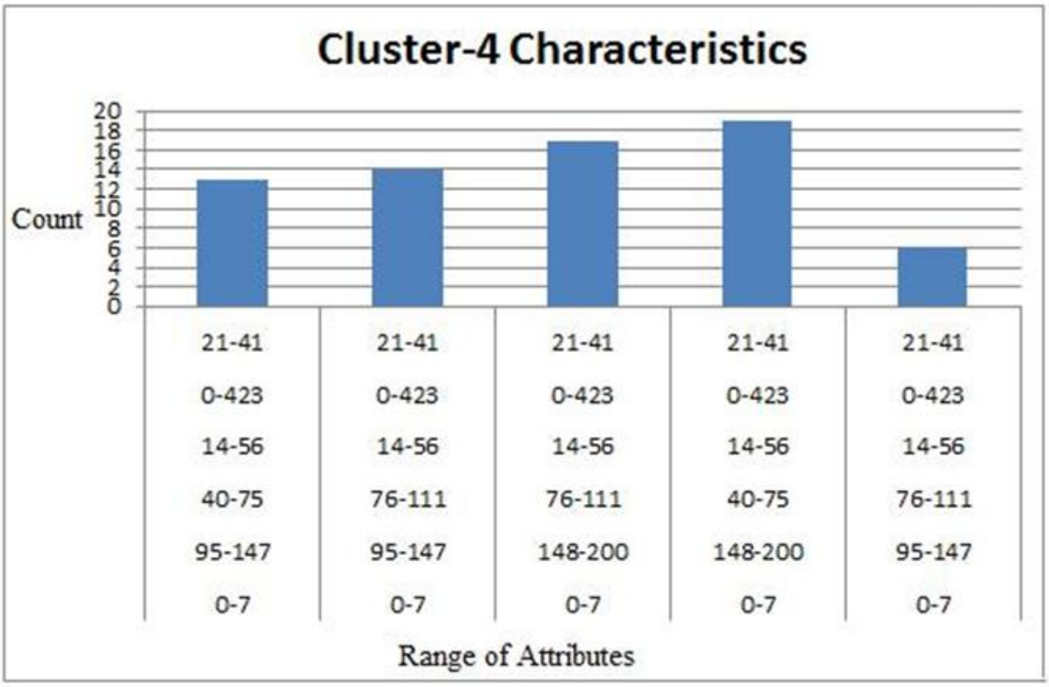
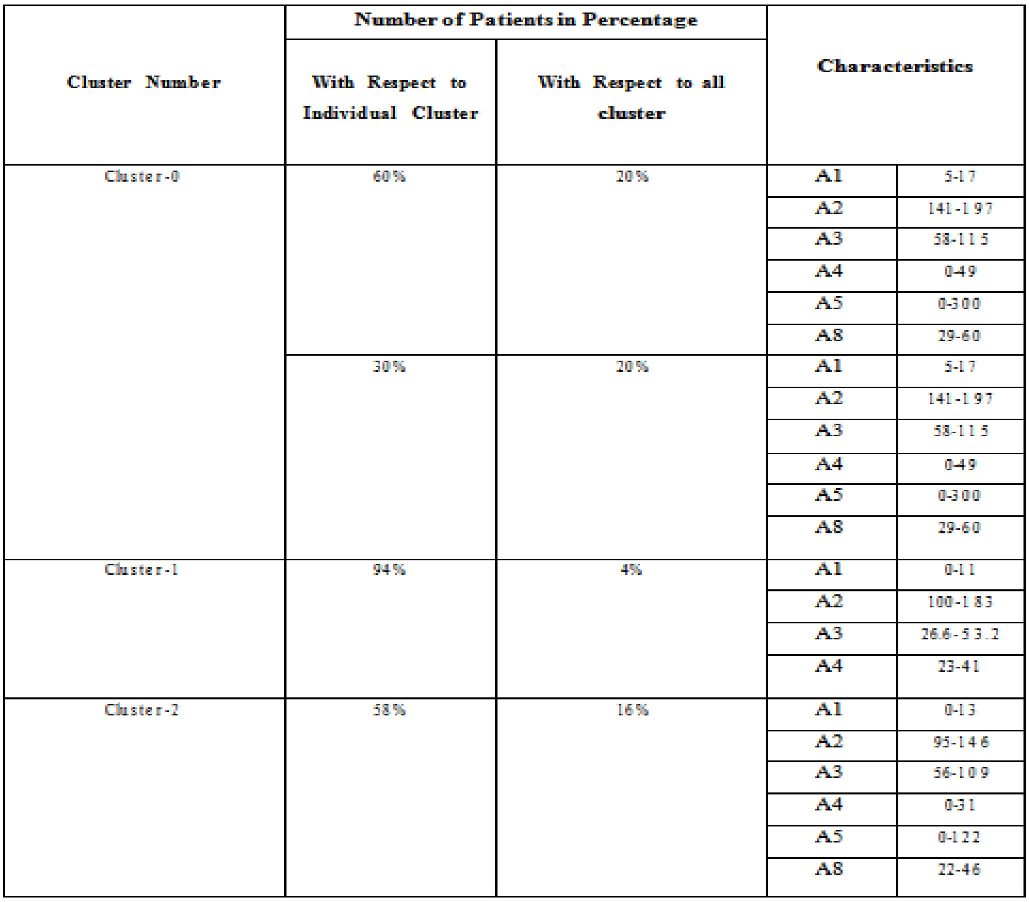
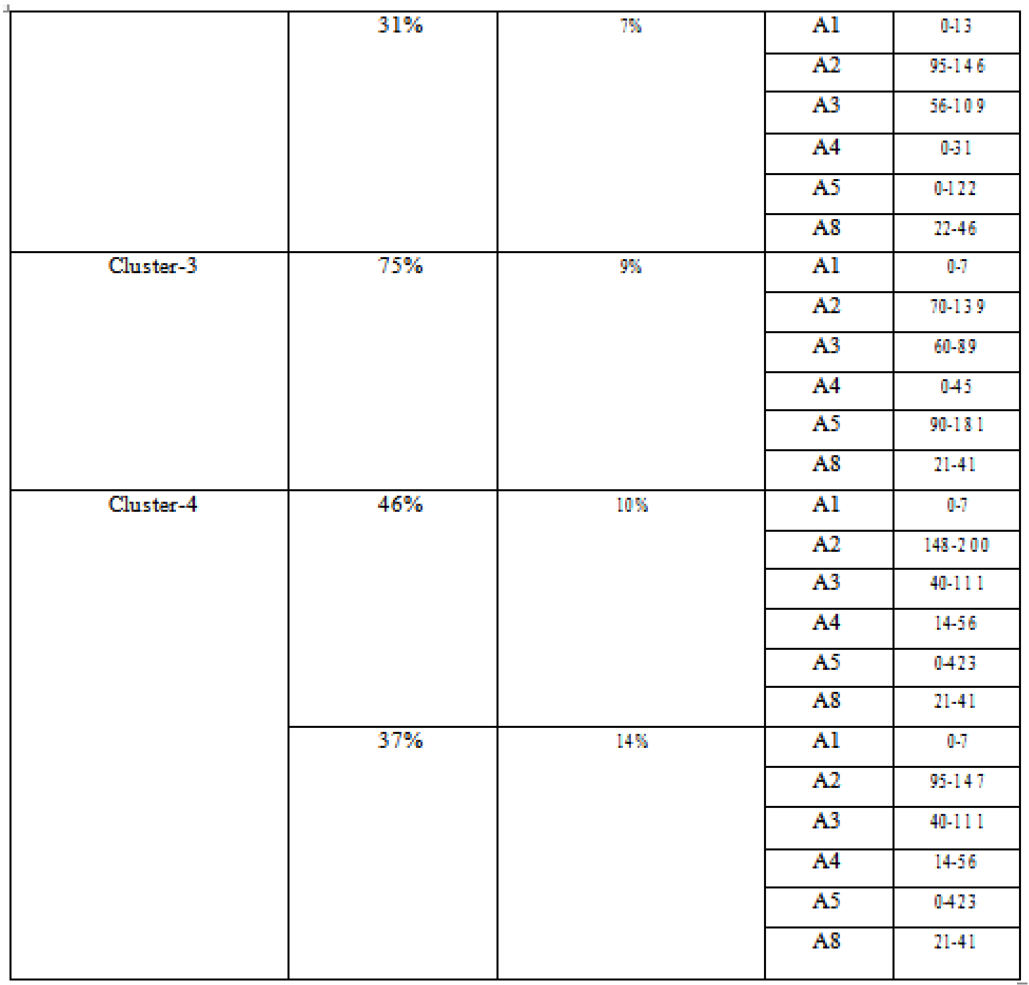


Fig 7: Cluster-4 Characteristics

TABLE II: Analysis of the results of Characteristics



Cluster Number	Number of Patients in Percentage		Characteristics	
	With Respect to Individual Cluster	With Respect to all cluster		
Cluster-0	60%	20%	A1	5-17
			A2	141-197
			A3	58-115
			A4	0-49
			A5	0-300
			A8	29-60
	30%	20%	A1	5-17
			A2	141-197
			A3	58-115
			A4	0-49
			A5	0-300
			A8	29-60
Cluster-1	94%	4%	A1	0-11
			A2	100-183
			A3	26.6-53.2
			A4	23-41
Cluster-2	58%	16%	A1	0-13
			A2	95-146
			A3	56-109
			A4	0-31
			A5	0-122
			A8	22-46



	31%	7%	A1	0-13
			A2	95-146
			A3	56-109
			A4	0-31
			A5	0-122
			A8	22-46
Cluster-3	75%	9%	A1	0-7
			A2	70-139
			A3	60-89
			A4	0-45
			A5	90-181
			A8	21-41
Cluster-4	46%	10%	A1	0-7
			A2	148-200
			A3	40-111
			A4	14-56
			A5	0-423
			A8	21-41
	37%	14%	A1	0-7
			A2	95-147
			A3	40-111
			A4	14-56
			A5	0-423
			A8	21-41

Here A1 represents number of pregnancies, A2 is plasma glucose concentration, A3 is Diastolic blood pressure, A4 represents Triceps skin fold thickness, A5 is Serum Insulin(2-Hour) and A8 represents age.

V. CONCLUSION

In this study comparative evaluation of diabetes characteristics is performed through various clustering algorithms. The four different algorithms used here are KNN (K- Nearest Neighbor), MST (Minimum Spanning Tree), K-medoid and K-means algorithm. The performance of these four algorithms was evaluated and it is concluded that k-medoid provides clusters of good quality. So it can be used for evaluating the characteristics of diabetes data. With the help of this study we can also identify the state of diabetes.

Therefore, this research will not only help to estimate the maximum number of diabetes patients but also the identifying the state of diabetes that whether the patient is in initial state or advanced state.

ACKNOWLEDGEMENT

The author would like to thank the department of computer science and engineering for their encouragement and support towards this research work.

REFERENCES

[1]. 'Data clustering: a review,' A.K. Jain, M.N. Murty, P.J. Flynn, ACM Computing Surveys, 31, 1999.

[2]. "An efficient k-means clustering algorithm: analysis and implementation," Kanungo, T., Mount, D., Piatko, C., Silverman, R.,Wu, A., IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.24, No.7, pp. 887-892, (2002).

[3]. www.diabetes.niddk.in.gov website of National institute of Diabetes, Digestive and Kidney Diseases.

[4]. Advances in knowledge discovery and data mining book contents", "Jiawei Han","Yongjian Fu"," Pages: 399 - 421 ,Year of Publication: 1996,ISBN:0-262-56097-6

[5]. Fartash.Haghanikhameneh, Payam. Hassany Shariat Panahy,, Nasim. Khanahmadliravi,, and Sayed Ahmad. Mousavi, "A comparison study between data mining algorithms over classification techniques in squid dataset", International Journal of Artificial Intelligence (IJAI), Volume 9, 2012.

[6]. Arpita M.Hirudkar and Mrs. S.S Sherekar , "Comparative analysis of data mining tools and techniques for evaluating performance of database system", ", International Journal of Computer Science and Applications, Vol. 6, 2013.

[7]. Nikhil N. Salvithal and Dr. R.B. Kulkarni, “Evaluating performance of data mining classification algorithm in weka”, International Journal of Application or Innovation in Engineering and Management (IJAIEM), Volume 2, Issue 10, October 2013.

[8]. Narendra Sharma, Aman Bajpai, and Mr.Ratnesh Litoriya, “Comparison the various clustering algorithms of weka tools” International Journal of Technology and Advanced Engineering (IJETAE), Volume 2, Issue 5, May 2012. [9] Osama Abu Abbas, “Comparison between data clustering algorithms”, The International Arab journal of Information Technology, Vol. 5, No. 3, July 2008.