# Characteristic evaluation of diabetes data using clustering techniques

**P.Padmaja**
**Srikanth Vikkurty**
**Nilofer Inaz Siddiqui**
**Praveen Dasari**
**Bikkina Ambica**
**V.B.V.E.Venkata Rao**
**Mastan Vali Shaik**
**V.J.P. Raju Rudraraju**

Department of Information Technology
Gitam University, Visakhapatnam, Andhra Pradesh, INDIA.

**Abstract**
Background & objectives: Taking into account the prevalence of diabetes among women the study is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of women suffering from diabetes.

Methods: Data mining functionalities like clustering and attribute oriented induction techniques have been employed to track the characteristics of the women suffering from diabetes. Information related to the study was obtained from National Institute of Diabetes, Digestive and Kidney Diseases.

Results: As clustering techniques have been utilized, the results were presented in the form of clusters showing the concentrations of the various attributes and the percentage of women suffering from diabetes with such characteristics. The results were evaluated in five different clusters and they show that 23% of the women suffering from diabetes fall in cluster-0, 5% fall in cluster-1, 23% fall in cluster-2, 8% in cluster-3 and 25% in cluster-3. It was also found that the characteristics seem to be varying for each cluster.

Interpretation & Conclusion: From the results it can be interpreted that the characteristics of the women suffering from diabetes are unique with respect to a cluster and no similarity can be found with respect to other clusters. The study helps in predicting the state of diabetes i.e., whether it is in an initial stage or in an advanced stage based on the characteristic results and also helps in estimating the maximum number of women suffering from diabetes with specific characteristics. Thus patients can be given effective treatment by effectively diagnosing the characteristics.

***Key words:***
*Clustering – diabetes – characteristic evaluation.*

## 1. Introduction

Clustering in data mining is a discovery process that groups a set of data such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized which indicates that, a cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications. Especially by clustering, one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlations among data attributes[1].

The overall performance of clustering is dependent on the discovery of quality clusters of items. Thus a majority of algorithms are concerned with efficiently determining the set of clusters in a given transaction or relational database[2]. The problem is essentially to generate the set of clusters in the database. Thus, different algorithms were introduced which aim at generating quality clusters which differ from one another in the context of generating the clusters

Manuscript received November 5, 2008
Manuscript revised November 20, 2008

*Applications:*

Clustering can be used as a stand alone tool to gain insight into the distribution of data, to observe the characteristic of each cluster, to focus on particular set of clusters for further analysis.

- Cluster analysis has been widely used in numerous application domains, including biology, medicine, anthropology, economics and market research.
- Clustering applications include plant and animal classification, disease classification, pattern recognition, data analysis, image processing etc.
- Recent uses include examining Web log data to detect usage patterns.

It may serve as a preprocessing step for other algorithms, such as characterization and classification, which would then operate on the detected clusters.

## 2. Material & Methods

The evaluation of the characteristics is performed taking into account the data obtained from National Institute of Diabetes, Digestive and Kidney Diseases[3]. Initially the data is grouped into clusters by using the clustering techniques. Once the data is grouped into clusters, the quality of the cluster needs to be identified as clusters without good quality are not helpful in effectively determining the characteristics. The estimation of the quality of a cluster helps us in identifying the algorithm that forms good quality clusters. Thus the characteristics of the diabetes data (available in form of clusters) can effectively be evaluated by applying the technique of *Attribute Oriented Induction* [4] for the clusters generated by the identified algorithm (i.e., algorithm that generates good quality clusters).

Figure I shows the general procedure for determining the characteristics of the diabetes data. Initially, the diabetes data is given as input to the normalization algorithm[5] which helps in generating the normalized file. The normalized file is then given as input to the appropriate clustering algorithm to generate the set of clusters. The best algorithm that generates clusters of good quality is determined by estimating the quality of the clusters generated by the respective algorithms.

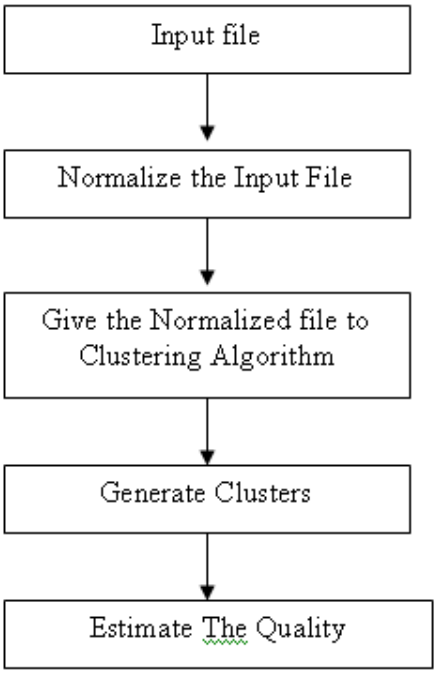



**Figure1** Cluster generation and quality estimation process

The "quality" of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster[6]. Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the centroid. In this study Centroid distance is considered to measure the quality of the cluster.

Four different algorithms are considered namely K-Means[7], Partitioning Around Medoids(PAM), Minimum Spanning Tree (MST) and Nearest Neighbor for generating the clusters and their quality is determined to identify the best algorithm that generates good quality clusters[8].

Once the algorithm is identified the characteristics of the clusters (diabetes data) can be evaluated using the approach of Attribute Oriented Induction. In this approach the first step is to identify the number of distinct values of various attributes. After identifying such attributes, those attributes with maximum number of distinct values are removed. From the remaining attributes maximum and minimum values are identified and the items are grouped using the concept of set grouping[9] by taking into account some threshold value.

## 3. Results & Discussion

The quality of the clusters generated by four algorithms can be compared graphically for identifying the best suitable algorithm. Table I shows the comparative analysis of the results of quality for the four algorithms for different number of clusters along with the graph. From the graph it can be identified that Partitioning Around Medoids algorithm provides clusters of good quality and hence this can be considered for evaluating the characteristics of diabetes data.

The diabetes data is classified into two classes; the target class, consisting of data which tested positive and the normal class which consists of normal data. Now the target class data (in clusters) is subjected to Attribute Oriented Induction for determining the characteristics.

The target class consists of 268 women who are actually affected with diabetes. However, it is unknown that in what stage the disease is. But after characterization it was observed that 23% of the women falling in cluster-0, 5% of them falling in cluster-1, 23% of them falling in cluster-2, 8% falling in cluster-3 and 25% falling in cluster-4 are the maximum number suffering from diabetes with the concentrations indicated in Table III. From this it can also be predicted that whether the disease is in an advanced state or in an initial state. The results of characterization are shown as graphs in Table II.

Hence this study will not only help in estimating the maximum number of people suffering from diabetes with specific characteristics but also helps in identifying the state of the disease.

## References

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

[2]. Kanungo, T., Mount, D., Piatko, C., Silverman, R., Wu, A.," An efficient k-means clustering algorithm: analysis and implementation," 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]"Jiawei Han","Yongjian Fu","*Advances in knowledge discovery and data mining book contents"*,Pages: 399 - 421 ,Year of Publication: 1996,ISBN:0-262-56097-6

[5] *"A control-flow normalization algorithm and its complexity (1992)*",by Zahira Ammarguellat, IEEE Transactions on Software Engineering

[6] Cluster Analysis By Mark S. Aldenderfer, Roger K. Blashfield

[7]http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

[8] "Javed A. Aslam","Katya Pelekhov","Daniela Rus","Generating, Visualizing, and Evaluating High-Quality Clusters for Information Organization Source Lecture Notes In Computer Science; Vol. 1481 archive Proceedings of the 4th International Workshop on Principles of Digital Document Processing table of contents,Pages: 53 ,- 69,Year of Publication: 1998 ISBN:3-540-65086-5"

[9] "*Using Destination Set Grouping to Improve the Performance of Window-Controlled Multipoint Connections (1996)*" by Shun Yan Cheung, Mostafa H. Ammar

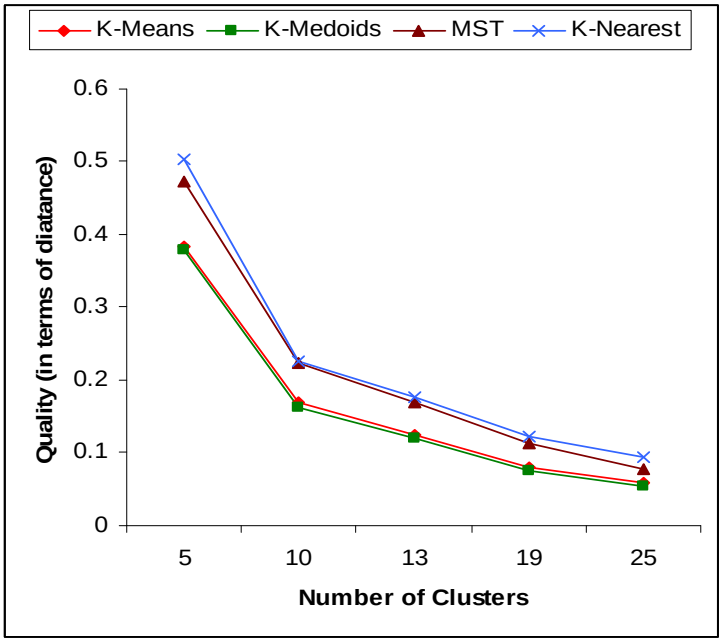
## Tables and figures

**Table I**. Quality results of algorithms for different number of clusters

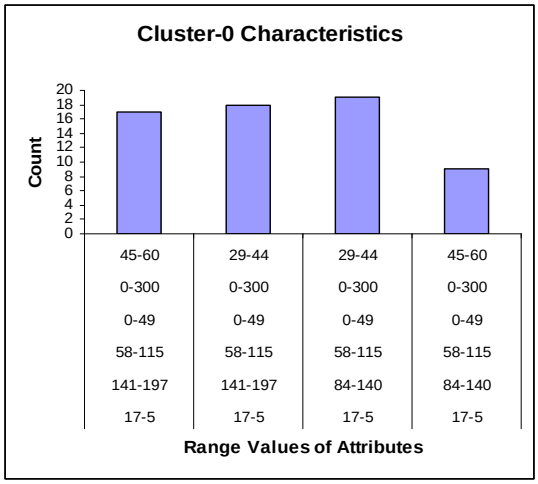
| Algorithm | number of clusters | | | | |
|---|---|---|---|---|---|
| | 5 | 10 | 13 | 19 | 25 |
| K-Means | 0.384 | 0.169 | 0.124 | 0.079 | 0.058 |
| PAM (K-Medoids) | 0.378 | 0.163 | 0.119 | 0.075 | 0.055 |
| MST | 0.473 | 0.223 | 0.169 | 0.112 | 0.078 |
| K-Nearest Neighbour | 0.503 | 0.226 | 0.176 | 0.123 | 0.093 |

**Graph** Quality results of algorithms for different number of clusters

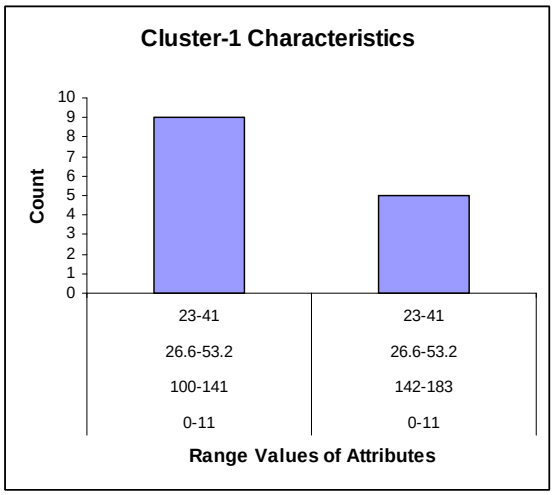
**Table II** Results of characterization



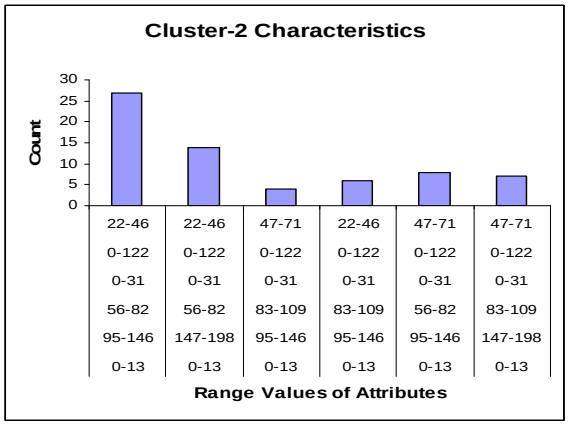
**Cluster-0**



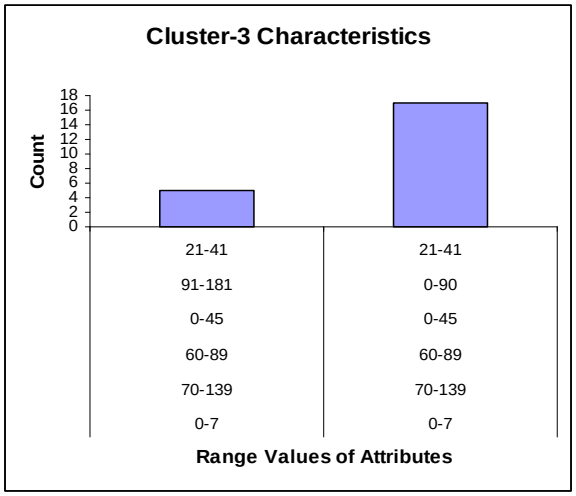

**Cluster-1**




**Cluster-2**




**Cluster-3**




**Cluster-4**

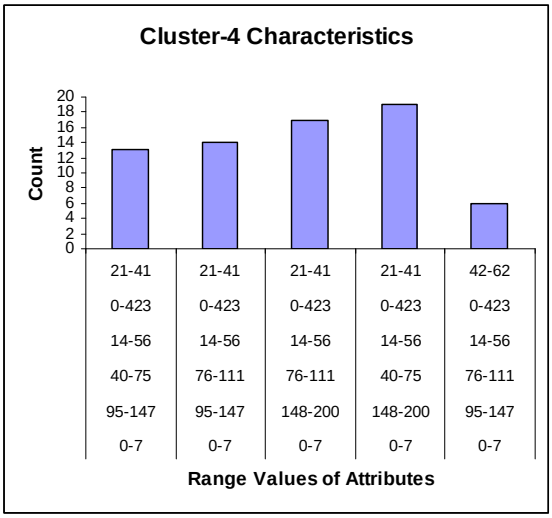


**Table –III**

## Analysis Of the Results Of Characterization

| Cluster Number | Maximum number of Patients ( in%) | | Characteristics | |
|---|---|---|---|---|
| | with respect to individual cluster | With respect to all cluster | | |
| **Cluster 0** | 50% | 30% | A1 | 5-17 |
| | | | A2 | 141-197 |
| | | | A3 | 58-115 |
| | | | A4 | 0-49 |
| | | | A5 | 0-300 |
| | | | A8 | 29-60 |
| | 40% | 10% | A1 | 5-17 |
| | | | A2 | 141-197 |
| | | | A3 | 58-115 |
| | | | A4 | 0-49 |
| | | | A5 | 0-300 |
| | | | A8 | 29-60 |
| **Cluster-1** | 93% | 5% | A1 | 0-11 |
| | | | A2 | 100-183 |
| | | | A3 | 26.6-53.2 |
| | | | A4 | 23-41 |

| | | | | |
|---|---|---|---|---|
| **Cluster-2** | 59% | 15% | A1 | 0-13 |
| | | | A2 | 95-146 |
| | | | A3 | 56-109 |
| | | | A4 | 0-31 |
| | | | A5 | 0-122 |
| | | | A8 | 22-71 |
| | 30% | 8% | A1 | 0-13 |
| | | | A2 | 95-146 |
| | | | A3 | 56-109 |
| | | | A4 | 0-31 |
| | | | A5 | 0-122 |
| | | | A8 | 22-71 |
| **Cluster-3** | 76% | 8% | A1 | 0-7 |
| | | | A2 | 70-139 |
| | | | A3 | 60-89 |
| | | | A4 | 0-45 |
| | | | A5 | 90-181 |
| | | | A8 | 21-41 |
| **Cluster-4** | 43% | 13% | A1 | 0-7 |
| | | | A2 | 148-200 |
| | | | A3 | 40-111 |
| | | | A4 | 14-56 |
| | | | A5 | 0-423 |
| | | | A8 | 21-41 |
| | 39% | 12% | A1 | 0-7 |
| | | | A2 | 95-147 |
| | | | A3 | 40-111 |
| | | | A4 | 14-56 |
| | | | A5 | 0-423 |
| | | | A8 | 21-62 |

A1 - Number of pregnancies; A2 - Plasma glucose concentration; A3 - Diastolic blood pressure; A4 - Triceps skin fold thickness; A5 - 2-Hour serum insulin; A6- Body mass index; A7 - Diabetes pedigree function; A8 – Age;

**Smt. Poosapati Padmaja** working as Associate Professor, Dept. Of IT, GITAM University, Visakhapatnam. Her field of interest in research is Data mining. She has authored many papers for National and International Conferences in Classification and Clustering.

**ikanth Vikkurty** has completed his Post Graduation in formation Technology , GITAM University. He has published per in a Journal in the area of Data Mining. Currently he is orking as a Developer in Software Company, Wipro, Bangalore, dia.

**Nilofer Inaz Siddiqui** pursuing Bachelor Degree in Information Technology, Gitam University. She has presented a paper in MNGSA Conference, Coimbatore, Tamil Nadu and BITS PILANI, Rajasthan.

**Praveen Dasari** is currently completing his Under-Graduation(B.E) programme in Department of Information Technology, GITAM University. As of now, his authored research papers included in 2 International Conferences Proceedings (PES IT, Bangalore and MNGSA, Coimbatore) and National Conference (NCSCA).His areas of interest in research include Data mining, Graph mining and Web mining and clustering techniques.

**Bikkina Ambica** is currentl doing her B.E in Informatio Technology in GITAM University,Visakhapatnam.He r field of interests includ clusterting and graphmining.

**V.B.V.E.VenkataRao** currently pursuing final year engineering in the emerging field of Information Technology in GITAM University, Visakhapatnam. He has authored a research paper for NCSCA National Conference and Quark-08, Goa Campus. His research activitiesconcern Clustering in Data mining and Aspect Oriented Programming.

**Masthan vali Shaik** presently studying Bachelor degree in Information Technology in GITAM University, visakhaptnam. He presented paper in international conference MNGSA,Bangalore and in (NCSCA-07) Dec 2007.

**V.J.P. Raju Rudraraju** currently pursuing final year engineering in the Information Technology in GITAM University, Visakhapatnam. He has authored a research paper for National Conference on Optimized design and implementation of TCP/IP software architecture (NCSCA-07) Dec 2007.