IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue- 2 E-ISSN: 2347-2693

Comparing clustering Algorithms with Diabetic Datasets in WEKA Tool

G.G.Gokilam[1*] and K.Shanthi[2]

[1]*Department of Computer Science and Engineering, PRIST University, TamilNadu, India*
[2] *Department of Computer Science and Engineering, Principal of Ponnaiyah Ramajayam Polytechnic College, TamilNadu, India*

www.ijcaonline.org

Received: 6 January 2015 Revised: 10 January2015 Accepted: 26 January 2015 Published: 28 February 2015

***Abstract*—** Data mining is the process of discover useful information from large datasets. The data mining techniques are used to analyze and evaluate diabetic dataset in the field of bio-medical. One of the most important techniques of data mining is clustering which is used to analyzing data from different perspectives and summarizing into useful information. Clustering is the task of assigning a set of objects into group called clusters. This paper discusses different clustering algorithms like cobweb, DBSCAN, EM, Farthest first, filtered cluster hierarchical cluster, OPTICS, simple Kmeans. The algorithms are used to compare its performance by Time taken to build the clusters, the cluster differentiated by its true positive and true negative values. Our main aim to show the comparison of the different cluster algorithms are evaluated in weka tool (Data mining Tool) and find out which algorithm will be most suitable for the diabetes dataset.

***Keywords*—** Cluster, Diabetes , Weka ,Data Mining

I. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amount of data or dataset for their development and finding useful patterns or important in raw data has been called KDD large number of data mining algorithms has been developed for mining of knowledge in databases. There are quite a lot of arguments that could be sophisticated to support the use of data mining in the health sector. In data mining Clustering [1] is the task of discovering groups and structures in data that are in some way or another similar without using known structures of data .Mostly these data are temporal in nature.

Diabetes is a disease in which levels of blood glucose, also called blood sugar, are above normal. People with diabetes have problems converting food to energy. Diabetes is not a newly born disease, it has been with human race from long back but, came to knew about it in 1552 B.C. Diabetes [6] mellitus is a set of related diseases in which the body cannot regulate the amount of sugar in the blood. Glucose in the blood gives you energy to perform daily activities, walk briskly, run for a bus, ride your bike, take an aerobic exercise class and perform your day-to-day chores. It is assumed that the execution of the Data Mining technology would be dealing out, memory and data demanding task as in opposition to one that require continuous interaction with the database.

Cluster analysis or Clustering is the assignment of a set of observations into subset called clusters so that observations in the same cluster are similar in some sense. Clustering [2] is a method of unsupervised learning and a common technique for statistical data analysis used in many fields like machine learning, data mining, pattern recognition, and image analysis and bio informatics.

WEKA is most powerful Data mining Tool Created by researchers at the University of Waikato in New Zealand[8] .It is Java based also open source. It functions like Preprocessing Filters, Attribute selection, Classification/Regression, Clustering, Association discovery, Visualization. Different types of clustering algorithms are compared by using diabetes dataset in WEKA Tool.

II. CLUSTERING ALGORITHMS

A) Cobweb

The cobweb algorithm produce a balanced tree with sub cluster at their levels and then K-means is applied to find resulting sub clusters. The algorithm reads one instance per iteration from a dataset and incorporates it into the tree by descending the tree along an appropriate path to a node where the category utility is maximal after absorbing the instance and updating statistical information then find proper place to hold the instance cobweb tries one or several or all of the following four possible operations at each node on the path 1) place the instance in an existing cluster 2) create a new cluster by itself 3) mere the best two cluster with respect to the values of category utility 4) split a cluster into several clusters by lifting its children one level in the tree to replace itself. The operation resulting in the largest value of category utility is the final choice on that node. This process is recursively performed until a leaf node is reached or a new leaf is created. This cobweb algorithm is tested with diabetes

dataset[10] in WEKA tool, it produce 931 clusters in 3. 20 seconds.

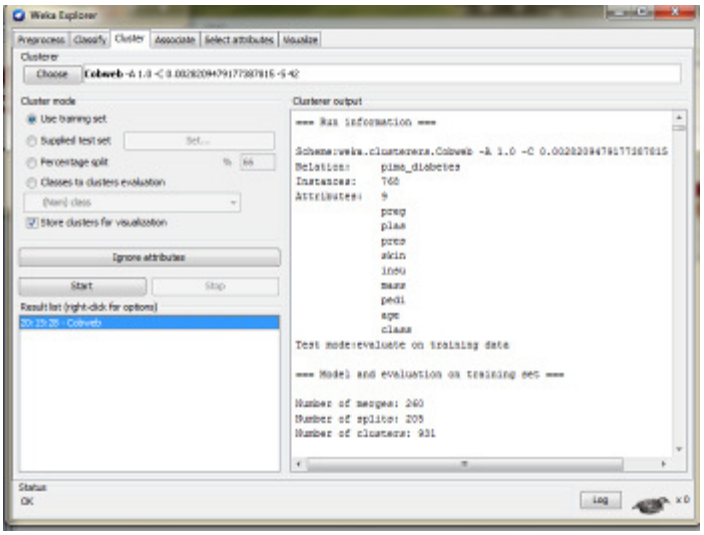


Figure 1: cobweb clustering algorithm in WEKA Tool

B) EM algorithm

The EM (**Expectation–Maximization)** algorithm is one such elaborate technique. The EM algorithm[7] is a general method for finding the maximum-likelihood estimate of the parameters from a given data set when the data is incomplete or has missing values. There are two main applications of the EM algorithm, The first one occurs when the data obtain from the observation process. The second one occurs when optimizing the likelihood function can be implied by assuming the existence values for additional but missing (or hidden) parameters. The latter application is more common in the computational pattern recognition community. It is an iterative method for finding maximum likelihood or Maximum a Posteriori (MAP) estimate of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. This EM algorithm is tested with diabetes dataset in WEKA tool. It produces 9 different clustered instances, Log likehood: -28.54483 and Time taken to build model in 156. 26 seconds approximately.

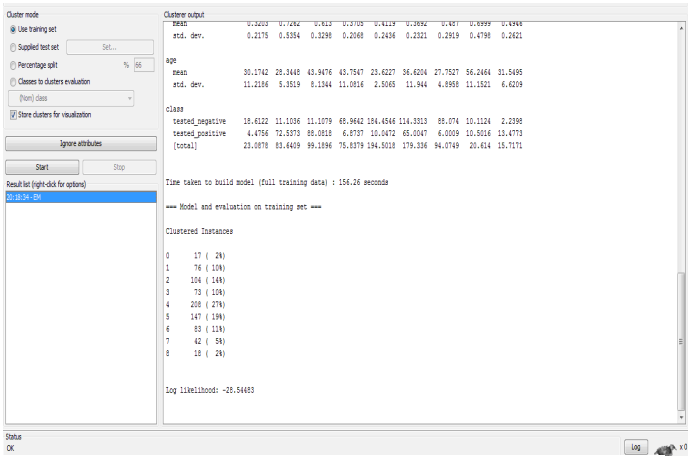


Figure 2: EM clustering algorithm in WEKA Tool.

C) DBSCAN

The DBSCAN is Density-Based clustering algorithms [8]to find clusters based on density of data points in a region and it use only one input parameter for their process so minimal knowledge is required. Density-Based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). DBSCAN separates data points into three classes: 1)Core points: points the interior of a cluster, 2) Border points: points neighborhood of a core point,3)Noise points: points which is not a core point or a border point. To find a cluster, DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and Min Pts. This DBSCAN algorithm is tested with diabetes dataset in WEKA tool, it produces two different clustered instances, clusters 0: tested_negative, cluster 1: tested_positive and Time taken to build model (full training data): 1. 23 seconds approximately.

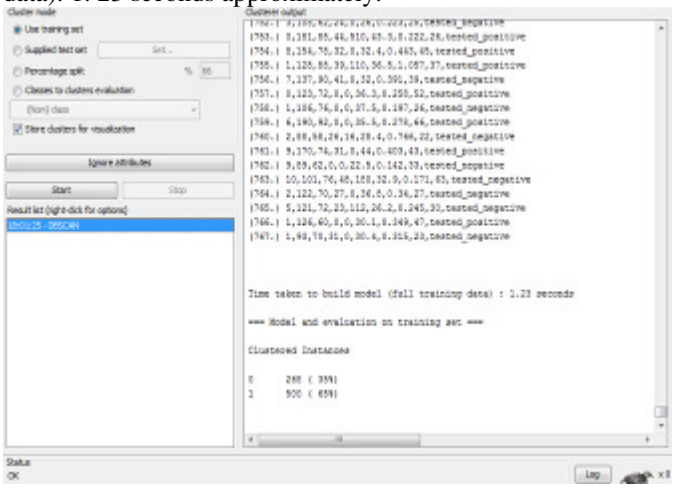


Figure 3: DBSCAN clustering algorithm in WEKA Tool.

D) FARTHEST-FIRST

Farthest first find its variant of K_Means[4], each cluster centre point furthermost from the existing cluster centre is placed by the K_Mean and this point must be positioned within the data area. So that it greatly speeds up the clustering in most cases but it need less move and adjustment for their fast performance. It use heuristic approach for finding its points. It's arbitrary point is p1, pick an another point p2 far from p1, pick pi to maximize the distance to the nearest of all centroid, the maximize the min{ dist(pi, p1),dist(pi,p2),…}.After all K representatives are chosen then we define the partition of data area D: cluster is Cj consists of all points closer to pj than to any other representative. This Farthest First algorithm is tested with diabetes dataset in WEKA tool; it produces two different clustered instance clusters 0: tested_negative, cluster 1: tested_positive and Time taken to build model (full training data): 0.02 seconds.

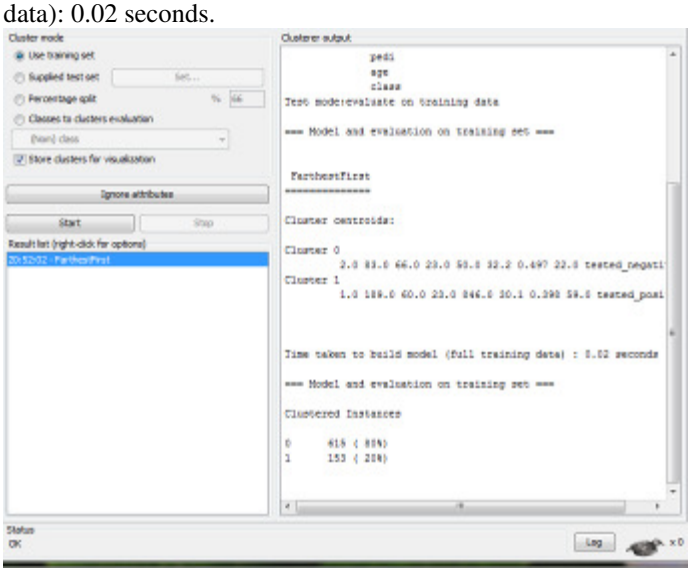


Figure 4: Farthest First Clustering algorithm in WEKA Tool.

E) FILTERED CLUSTER

The filtered cluster algorithm is based on storing the multidimensional data points in a kd-tree. The process of the tree is like a binary tree approach, which represents a hierarchical subdivision of its data point set's bounding box using their axis and then splitting is aligned by hyperplanes. Each node of the kd-tree is associated with a closed box, called cell. The root's cell is the bounding box of the point in the dataset. If the cell contains at most one point, then it is declared to be a leaf. Then the finding points in the cell are then partitioned to one side or the other of this hyper plane. The resulting sub cells are the children of the original cell, this leads to a binary tree structure. This Filtered cluster algorithm is tested with diabetes dataset in WEKA tool; it produces two different clustered instance clusters 0: tested_negative, cluster 1: tested_positive and Time taken to build model (full training data): 0. 04 seconds.

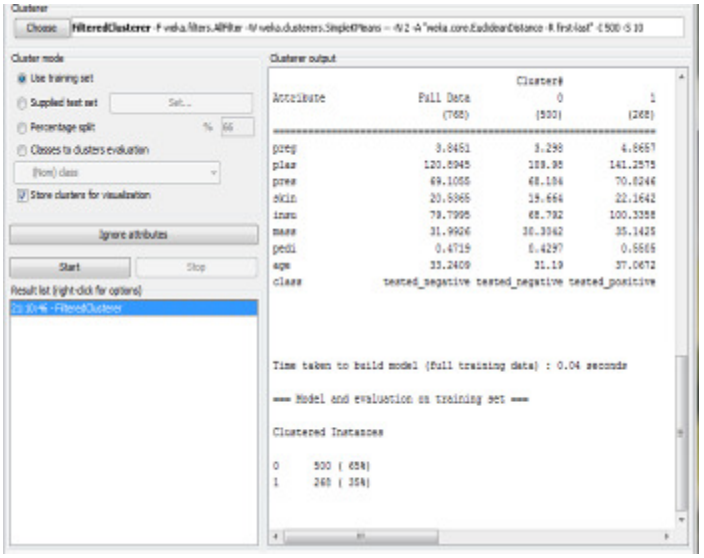


Figure 5: Filtered Cluster algorithm in WEKA Tool.

F) HIERARCHICAL CLUSTERING

Hierarchical cluster [3] divides the clusters in a sequential manner with nested portioned. It consists of Agglomerative approach and divisive approach.

i) Agglomerative: This is a "bottom up" approach, each observation starts in its own cluster, and similar clusters are merged as one moves up the hierarchy until all its data form into one cluster. The algorithm will look for the two most similar data points and merge them to create a new "pseudo-data point", which represents the average of the two merged data points. Each iterative step takes the next two closest datapoints (or pseudo-datapoints) and merges them. This process is continued until one cluster containing all of its original datapoints.

ii) Divisive: This is a "top down" approach, and this hierarchical clustering having all its objects into one cluster then split the cluster into smaller cluster. In its splitting process needs minimum relation for the different cluster and maximum relation in the same cluster.

This Hierarchical clustered algorithm is tested with diabetes dataset in WEKA tool, it produces two different clustered instances, clusters 0: tested_negative, cluster 1: tested_positive and Time taken to build model (full training data): 5.37 seconds.

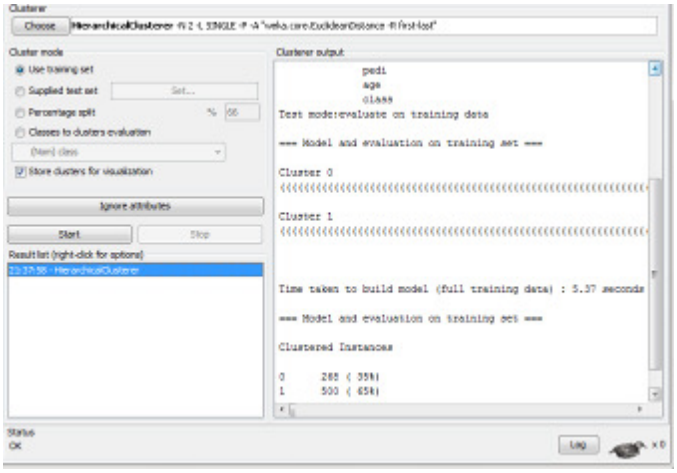
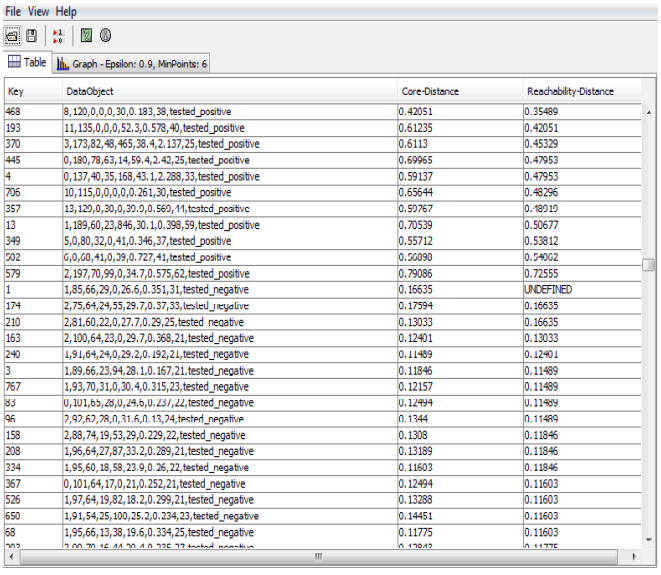


Figure 6: Hierarchical clustering algorithm in WEKA Tool.

G) OPTICS: Ordering Points To Identify the Clustering Structure

OPTICS to create an ordering of a data set with respect to its density-based clustering structure is presented Optimization based partitioning algorithms typically represent clusters by a prototype. Objects are assigned to the cluster represented by the most similar (i.e. closest) prototype. An iterative control strategy is used to optimize the whole clustering such that, e.g., the average or squared distances of objects to its prototypes are minimized. The OPTICS algorithm generates the augmented cluster-ordering consisting of ordering the points, reach ability-values and core-values. This OPTICS is tested with diabetes dataset in WEKA tool; it produces two different clustered instance and Time taken to build model (full training data): 1. 42 seconds.



Key	DataObject	Core-Distance	Reachability-Distance
468	8,120,0,0,0,30,0.183,38,tested_positive	0.42051	0.35489
193	11,135,0,0,0,52.3,0.578,40,tested_positive	0.61235	0.42051
370	3,173,82,48,465,38.4,2.137,25,tested_positive	0.6113	0.45329
445	0,180,78,63,14,59.4,2.42,25,tested_positive	0.69965	0.47953
4	0,137,40,35,168,43.1,2.288,33,tested_positive	0.59137	0.47953
706	10,115,0,0,0,0,0.261,30,tested_positive	0.65644	0.48296
357	13,129,0,30,0,39.9,0.569,44,tested_positive	0.59767	0.48919
13	1,189,60,23,846,30.1,0.398,59,tested_positive	0.70539	0.50677
349	5,0,80,32,0,41,0.346,37,tested_positive	0.55712	0.53812
502	6,0,68,41,0,39,0.727,41,tested_positive	0.56090	0.54062
579	2,197,70,99,0,34.7,0.575,62,tested_positive	0.79086	0.72555
1	1,85,66,29,0,26.6,0.351,31,tested_negative	0.16635	UNDEFINED
174	2,75,64,24,55,29.7,0.37,33,tested_negative	0.17594	0.16635
210	2,81,60,22,0,27.7,0.29,25,tested_negative	0.13033	0.16635
163	2,100,64,23,0,29.7,0.368,21,tested_negative	0.12401	0.13033
240	1,91,64,24,0,29.2,0.192,21,tested_negative	0.11489	0.12401
3	1,89,66,23,94,28.1,0.167,21,tested_negative	0.11846	0.11489
767	1,93,70,31,0,30.4,0.315,23,tested_negative	0.12157	0.11489
83	0,101,65,28,0,24.6,0.237,22,tested_negative	0.12494	0.11489
96	2,92,62,28,0,31.6,0.13,24,tested_negative	0.1344	0.11489
158	2,88,74,19,53,29,0.229,22,tested_negative	0.1308	0.11846
208	1,96,64,27,87,33.2,0.289,21,tested_negative	0.13189	0.11846
334	1,95,60,18,58,23.9,0.26,22,tested_negative	0.11603	0.11846
367	0,101,64,17,0,21,0.252,21,tested_negative	0.12494	0.11603
526	1,97,64,19,82,18.2,0.299,21,tested_negative	0.13288	0.11603
650	1,91,54,25,100,25.2,0.234,23,tested_negative	0.14451	0.11603
68	1,95,66,13,38,19.6,0.334,25,tested_negative	0.11775	0.11603

Figure 7: OPTICS Clustering in WEKA Tool.

H) SIMPLE K-MEANS

K-means [5] is one of the simplest unsupervised learning algorithms that solve the clustering problem. It classifies a given data set through a certain number of clusters fixed a priori. The main idea is to define k centroids, one for each cluster. This process is iterated until there is no change in gravity centers. The algorithm works like, First place the point k into space represented by the object are clustered have initial group centroids. Each object assign to a group has closest centroids. While all objects assigned then recalculate the position of the K centroids. This type of cluster is tighter than other clusters. This Simple K-Means is tested with diabetes dataset in WEKA tool; it produces two different clustered instance and Time taken to build model (full training data): 0.09 seconds.

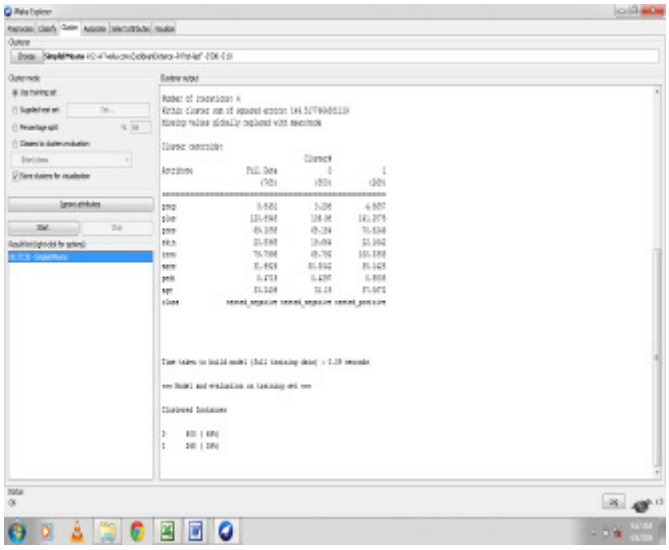
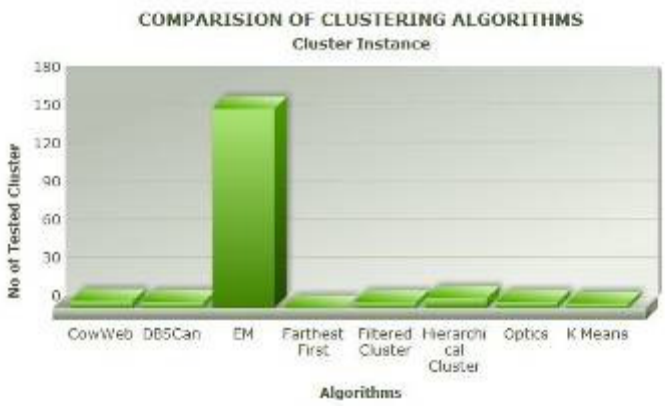


Figure 8: Simple K-Means clustering algorithm in WEKA Tool.

III COMPARISON OF CLUSTERING ALGORITHMS

Algorithm Used	Time Taken (in seconds)	Number Of clusters	Number Of Cluster Objects	Cluster Instances	
				0 tested_ negative	1 tested_ positive
Cobweb	3. 20	931	-	-	-
DBscan	1. 23	02	768	268	500
EM	156. 26	09	768	-	-
Farthest First	0. 02	02	768	615	153
Filtered Cluster	0.04	02	768	500	268
Hierarchical cluster	5.37	02	768	268	500
Optics	1.42	-	768	-	-
Simple KMeans	0.09	02	768	500	268

Table 1: Cluster algorithms tested in WEKA Tool.



IV. CONCLUSION

In the recent few years data mining techniques covers every area in our life. We are using data mining techniques in mainly in the medical, banking, insurances, education etc. There are different data mining clustering techniques can be used for the identification of diabetes disease among patients They names are: cobweb, DBSCAN, EM, Farthest first, filtered cluster, hierarchical cluster, OPTICS, These techniques are compared by using data mining tool WEKA with diabetes dataset which produce the result as tested_positive, tested_negative for the affected and not affected by the diabetes disease .It is the simplest tool for classify the data various types of cluster. It is the first model for provide the graphical user interface of the user while perform the clustering we used the promise data repository. It is providing the past project data for analysis. With the help of figures we are showing the working of various algorithms used in weka also time taken to form the cluster. Every algorithm has their own importance and we use them on the behavior of the data, but on the basis of this study we found that farthest first clustering algorithm requires minimum time taken to form the cluster and also it is simplest algorithm as compared to other algorithms. This paper shows only the clustering operations in weka using diabetes dataset.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", second edition, Morgan Kaufmann Publishers an imprint of Elsevier.

[2]A.K. JAIN Michigan State University, M.N.MURTY Indian Institute of Science AND P.J. FLYNN The Ohio State University: "Data Clustering".

[3] P. Vijaya, M N Murthy and D K Subramanian. Leaders-sub leaders, "An efficient hierarchical clustering algorithm for large data sets",Pattern Recognition Letters (2004) 505-513.

[4] Rama. B,"A Survey on clustering Current status and challenging issues" (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 09, 2010, 2976-2980.

[5] M. Pramod Kumar "Simultaneous Pattern and Data Clustering Using Modified K-Means Algorithm" International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 2003-2008.

[6] Miroslav Marinov, M.S.,1 Abu Saleh Mohammad Mosa, M.S.,1 Illhoi Yoo, Ph.D.,1,2 and Suzanne Austin Boren, Ph.D., MHA1,2 " Data-Mining Technologies for Diabetes: A Systematic Review" Journal of Diabetes Science and Technology Volume 5, Issue 6, November 2011 © Diabetes Technology Society.

[7] Celeux, G. and Govaert, G. (1992). "A classification EM algorithm for clustering and two stochastic versions. Computational statistics and data analysis", 14:315–332

[8] Narendra Sharma , Aman Bajpai , Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 5, May 2012).

[9] Dr. Wenjia Wang, "Tutorial for DM tool Weka 1 CMP: Data Mining and Statistics within the Health Services".

[10] K. Rajesh, V. Sangeetha ," Application of Data Mining Methods and Techniques for Diabetes Diagnosis" International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012

AUTHORS PROFILE

1. **G.G.Gokilam** received her MTech(Computer science and Engineering) in 2010 from PRIST University. Now Doing Ph.D in PRIST University.

2. **Dr K.Shanthi** received her PhD(Computer Science) in 2012 from Bharathidasan University , MTech(Computer Science and Engineering) from PRIST University in 2010, Mphil(Computer Science) in 2005 from Bharathidasan University, MCA from Bharathidasan University. in 2003 and MSc(Mathematics) in 1993 from Annamalai University. She has 15 years of teaching experience. Now working as Principal of Ponnaiyah Ramajayam Polytechnic College, Thanjavur, Tamilnadu.