Improved Optimization centroid in modified Kmeans cluster

G.G.Gokilam[a] and K.Saravanan[b]

[a] Research scholar, Department of Computer Science and Engineering, PRIST University, Thanjavur,India.
gggokilam@yahoo.co.in

[b] DEAN, Faculty of Computer Science, PRIST University, Thanjavur, India.
ks_tnj@yahoo.co.in

Abstract— **Data mining is the process which determines the interesting patterns from huge amounts of data which can be conducted by any kind of data as long as it forms meaningful for their application like database data. Data mining provides the methodology and knowledge to convert these mass of data into useful information for decision making in numerous fields like finance, Retail Industry, Telecommunication, Weather forecasting, Healthcare. Data mining process can split like supervised and unsupervised learning. Classification and regression are supervised; clustering and association are unsupervised learning. Data analysis dealings for clustering can be dichotomized as groping of assent data based on the accessibility of proper models. Clustering algorithms like cobweb, Density Based SCAN, Hierarchical, EM, Farthest first, filtered cluster, OPTICS and simple Kmeans are used to form the distinct class of objects. In this paper, simple kmeans algorithm are modified and implemented in diabetes dataset to find the patient affected by diabetes diseases with test condition like Tested Negative, Tested Normal, and Tested High which produce optimization result for large dataset.**

Keywords: Data mining, Clustering, Kmeans, Diabetes.

I. INTRODUCTION

Data mining is the use of computerized data analysis techniques to discover past undetected dealings between data items. Data mining often involve the analysis of data stored in a data warehouse. Some of the major data mining techniques are regression, classification and clustering. Data mining techniques are the effect of a long process of research and invention in the developments. This development begins when business data was first stored in computers, continue their development in data access and more recently it make knowledge that allow users to find data in real time. Data mining takes this development process beyond display data access and navigation to potential and practical information delivery [6]. Data mining is ready for application in the big business community for the reason that it is supported by some technologies like huge data collection, great multiprocessor computer and Data mining algorithms.

Clustering is the course of creation a collection of abstract objects into classes of related objects. A cluster of data objects can be treated as individual group. While forming the group, first the set of data is separated into groups based on data relationship and then assign the labels to the groups. Importance of clustering algorithm needs highly scalable to deal with large databases. Facility to deal with different kinds of attributes like interval-based numerical and categorical data. The clustering algorithm should be capable of perceive clusters of arbitrary shape. They should not be enclosed to only distance measures that tend to find spherical cluster of small sizes and able to handle low-dimensional data in high dimensional space [3]. Sometimes Databases have noisy, missing or invalid data. Clustering algorithms also handle this situation. Then cluster results should be intelligible, and usable. Cluster analysis is one of the most important data analysis methods which generally used for many practical applications like Bioinformatics. Clustering is the method of partitioning a given set of objects into disjoint clusters. The k-means algorithm [1] is successful in producing clusters for several practical applications. But the k-means algorithm have high complex for their calculation, particularly for large data sets. In addition, this algorithm results in dissimilar types of clusters depending on the arbitrary choice of initial centroids [9]. Some research work depends on increasing the performance of the k-means clustering algorithm. This paper deals with optimization method in simple Kmeans in data mining.

Nowadays Diabetes Mellitus is one of the worldwide health problems. Its birth in 17th century and lack of insulin is primary cause of diabetes was introduced in 18th century. This problem was differentiated like Type1, Type2, and Gestational diabetes. In type1 diabetes is fully depend on insulin, when the bête cell of pancreas

DOI: 10.21817/ijet/2017/v9i2/170902224

destroy insulin production and uncontrolled blood glucose level may damage the body organs but it is less common diseases [8]. Type 2 diabetes can indentify by that body does not produce enough insulin to meet its own needs. No exact symptoms at early stage, it may occur by obesity, hereditary and it controlled by diet food, medicines, and some exercise. Suppose the person affected by diabetes leads to suffer various difficulties likes visual destruction, cardio vascular ailment, leg deduction and renal failure if diagnosis is not completed in the right time [7]. Some women also have high sugar during pregnancy. Insulin production is less in their body this diabetes called as gestational diabetes that pregnant woman does not have pre-existing diabetes it may occurs second trimester of pregnancy and disappears after baby born but it affects 4 to 5% of pregnant women.

II. Related work

To improve the efficiency of the k-means algorithm [2], The k-prototypes algorithm [1] incorporate the k-means and k-modes processes for clustering the data. This method process the contrast measure is defined by taking both numeric and categorical attributes. Normally K means algorithm process like first decide how many number of clusters to form and then assign the data points which was closest to its centroid. When the centroid became stable means stop the process otherwise recalculate the centre point for the cluster and then assign data points by Huang Z[1]. Systematic method to find the initial centroids in distributed data which produced clusters with better accuracy, compared to the k-means algorithm by Fang Yuan [3]. Improving the accuracy and efficiency of the Kmeans clustering algorithm by M.P. Sebastian [4] research consist of two parts first determine initial centroids of the cluster by using an algorithm and second part assign each data point to the appropriate cluster . Enhancing Kmeans clustering algorithm with improved initial centre by Madhu yedla [5] Data points contain both positive and negative value then arrange the data points with equal range. Finding better initial cluster by koheri aria[6] both kmeans and hierarchical are used to find cluster.

In scientific data collection have some advanced methods produced resulted in the huge scale of talented data pertaining in different fields of science and technology. Outstanding developed techniques for generating and accumulate data, the increase of scientific databases become tremendous so practically impossible to extract useful information from them by using predictable database analysis techniques [10]. Effectual mining methods are extremely essential to expose implicit information from enormous databases. It works like that objects in the same cluster are related while objects belonging to another clusters vary considerably, with respect to their attributes.

III. Proposed Method

A new method has been proposed to increase performance of kmeans clustering algorithm. In this method the initial centroids are choose by a procedure[11] which has some systematic approach, so this method is very perceptive for the initial starting points and then produce the unique clusters [4, 5]. This algorithm processed to find the initial centroid of the cluster by arranging the data elements in sorting order then subtracting first element from the last element, resulting value are divided by number of cluster that values are taken intervals and form the cluster then use arithmetic mean to find initial centroid. Then apply this centroid value to form the clusters with minimum distance and these processes are continuing until no change in centroid value.

Prerequisite

d={d1,d2,d3….dn} //elements in dataset
k=expected number of cluster // were k≠0

Guarantee

Set of clusters.

Proposed Algorithm

1. Arrange data elements in ascending order.
2. Find initial cluster width $k_w = \frac{dn-di}{k}$ (i=1)
3. Use the value of k_w divide data elements and create k no. of cluster.
4. To find the centroid value by using mean for each cluster.

   $Ce = |\sum_{i=1}^{m} \frac{di}{m}|$ // m total member of elements in each cluster
5. For each element d, compute distance.

6. Add elements d to the corresponding cluster.
7. Repeat step5 and 6 for cluster formation until no change in centroid value.

IV RESULTS AND DISCUSSION

Proposed algorithm is applied in real time diabetes dataset contains Patients blood sugar level before and after food which applied to find diabetes test condition like Tested Negative, Tested Normal, Tested Positive as result. This method finds initial centroids steadily and requires the data values as an input for cluster. Patient age and before food sugar values are taken as input and apply kmeans to form clusters in figure (1) and figure (2). Same patient age and after food sugar values are taken as input and apply kmeans to form cluster in figure(3) and (4).

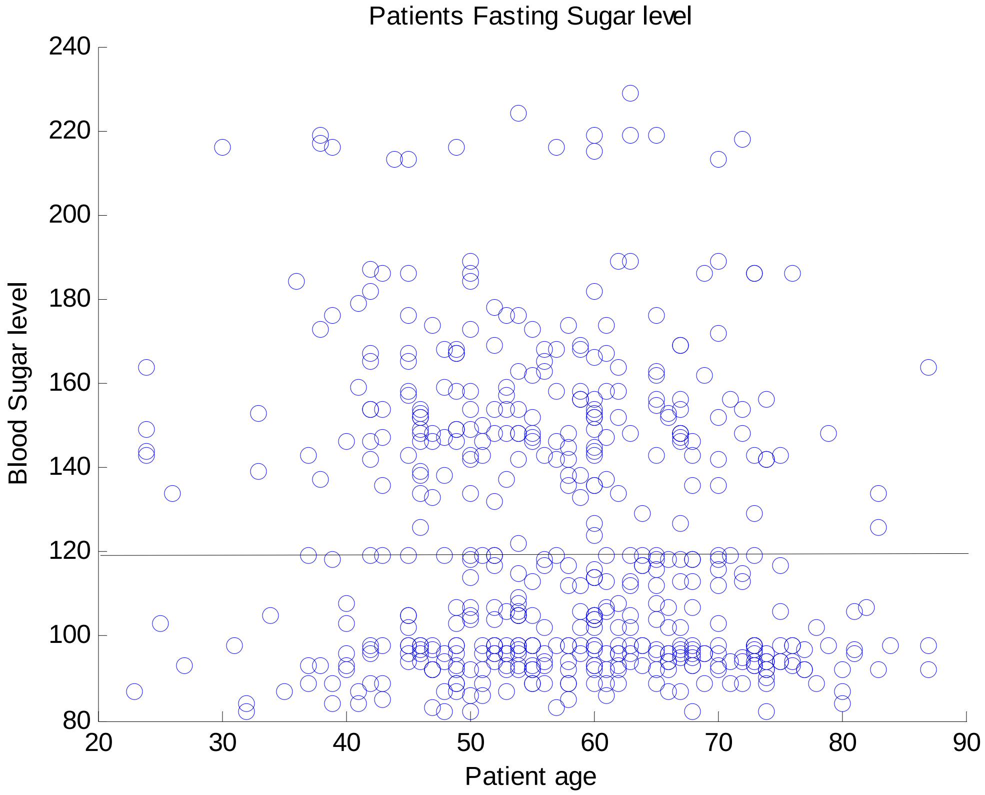

Figure 1: Patients age and Fasting glucose level.

Figure 1 illustrate patient age value in x-axis and blood sugar value in y- axis Normally the Fasting glucose level are taken in early morning before take any food even liquid food also avoid, some doctor advice this fasting sugar value can take 8 to 10 hours gap between dinner and blood taken time. Normally, this sugar value is ranging from 80 to 120. If patient sugar value above this range, it is treated as high in sugar level.

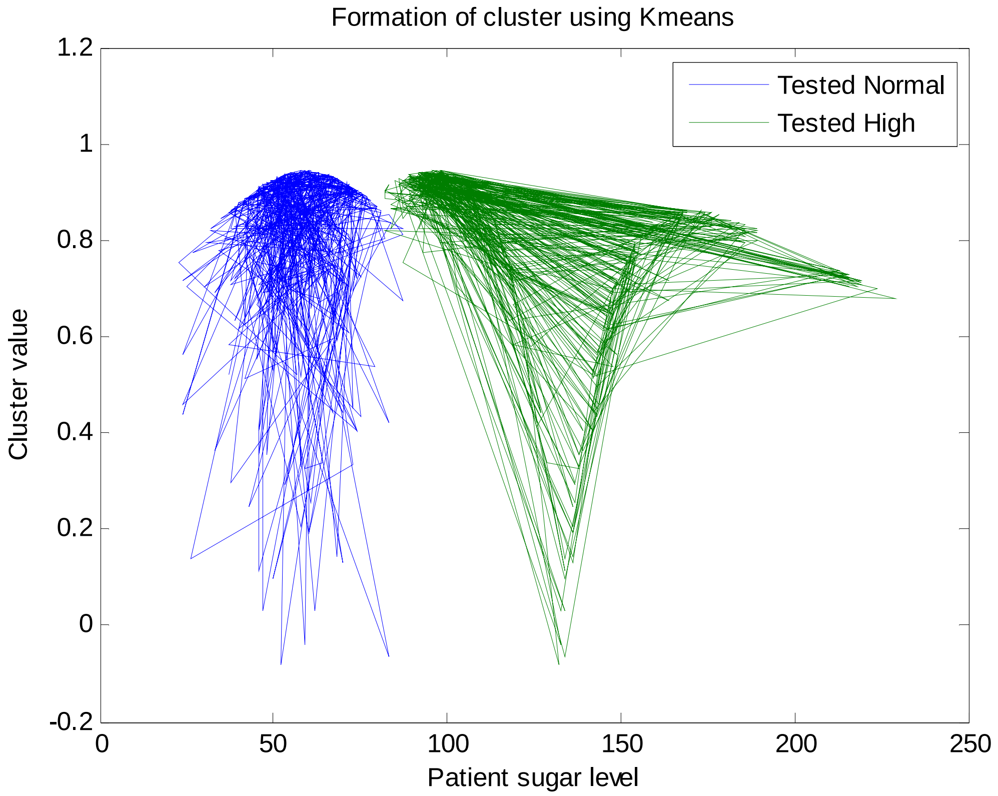

Figure 2: Patient Test value (Fasting sugar) and Kmeans cluster.

Above figure(2) illustrate that patient sugar value in x-axis and cluster value in y-axis apply kmeans clustering algorithm to form clusters, it use Euclidean distance to find minimum distance from initial point to other data points in cluster group. Normally it chooses farthest point and takes average mean to find centroid value and assign closest point to the cluster.

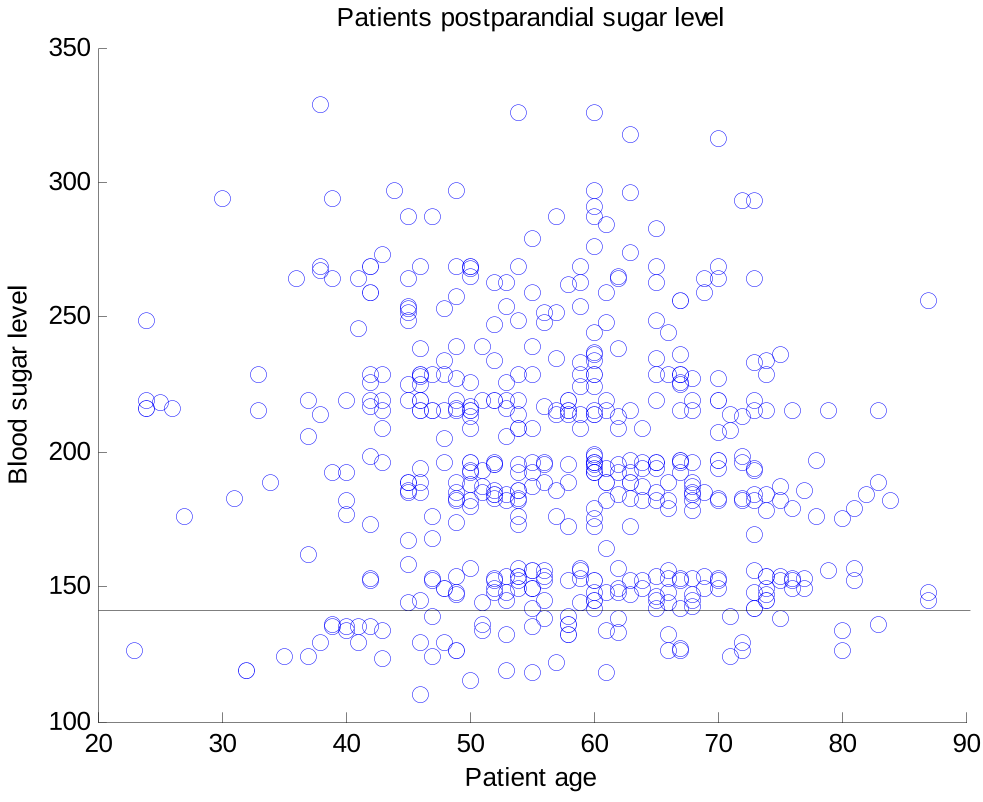

Figure 3: Patients age and postprandial glucose level.

Above figure(3) illustrate patient age value in x-axis and blood sugar (postprandial) value in y- axis Normally the postprandial glucose level are taken after one and half hours from breakfast in the mean time avoid liquid food. In general, this sugar value ranges from 100 to 140. If patient sugar value lies in between this range means treated as normal, above this range means treated as high in sugar level.

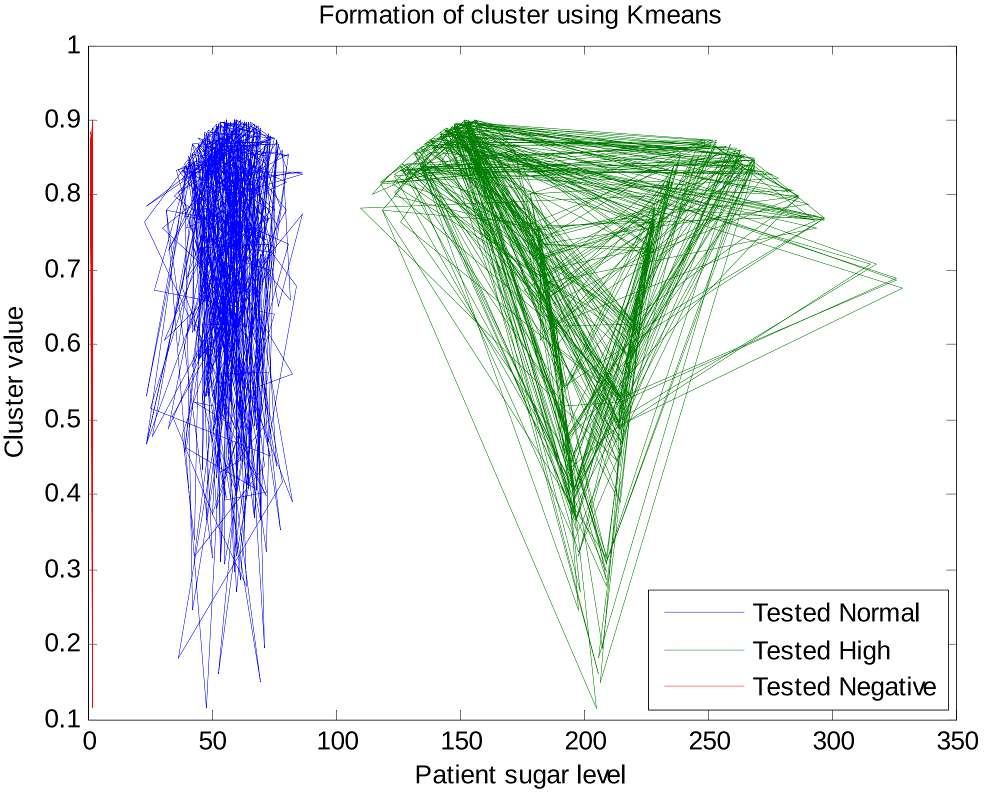

Figure 4: Patient Test value (postprandial) and Kmeans cluster.

Above figure (4) illustrate that patient postprandial sugar value in x-axis and cluster value in y-axis apply kmeans algorithm on sugar value to form the cluster with tested negative, normal and high. Some patient have above 300 sugar level also that range people's need more concentration on health because that high level affects nerves which causes eye disease, kidney and heart problem.

The data values are applied in data mining software tool and form clusters then same values are applied in modified algorithm it produce some difference in centroid value also number of iteration is reduced. In proposed algorithm some sequence steps are carry out to find the centroid value. Once data points are arrange in order means then centroid value are quickly obtain then it reduce the iteration shown below.

	Fasting Sugar levels	
Cluster Centroid	Software Tool	Proposed Algorithm
1	98.8873	99.4527
2	103.8023	150. 2901
3	161.7297	196.1951
Number of Iteration	11	10
	Postprandial Sugar values	
Cluster Centroid	Software Tool	Proposed Algorithm
1	164.8041	144.7125
2	174. 2179	203.0357
3	244.3156	267.2873
Number of Iteration	15	10

Table 1: Evaluate centroid value after cluster formation.

S.No		**Fasting sugar level.**		**Postprandial sugar level.**	
1.	Maximum difference	147.0		219.0	
2.	Cluster length	49.0		73.0	
3.	Mid value	131.0		183.0	
4.	Cluster Centroid	Initial	Final	Initial	Final
	1	100.0066	99.4527	151.3908	144.7125
	2	152.5333	150.2901	211.7295	203.03572
	3	201.8437	196.1951	277.36206	267.28735
5.	Number of iteration	10		10	

Table 2: comparison of proposed algorithms test results

Above table shows the comparison fasting sugar level and postprandial sugar levels with initial and final centroids. Difference between initial and final centroids value is minimum, so once initial centroid are chosen best means then assign data points with minimum distance and number of iteration is also reduced, it shown in below figure5.

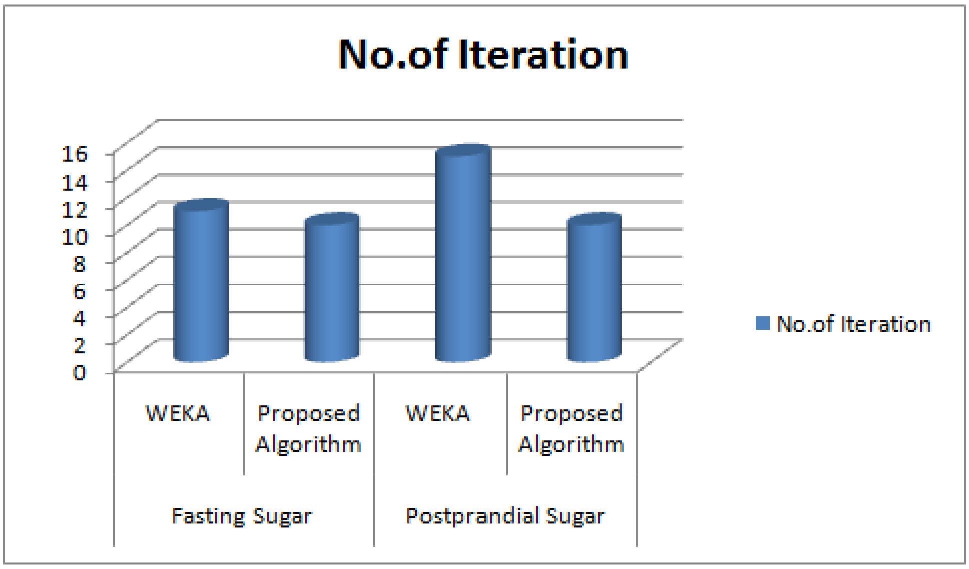

Figure 5: Compare Kmeans with its Processing steps.

V Conclusion

Clustering algorithms have a powerful technique for organizing objects in that Kmeans is most special and powerful algorithm to form clusters. Mostly this algorithm chooses random points for their first initialization to find the centroid value then using Euclidian distance for arrange the data points. But in this new method some sequence method are used to find initial centroids and form the cluster which produce optimization result. The data values are applied in software tool apply Kmeans algorithm to produce patients sugar level like, Tested negative, Tested Normal, Tested High. Number of Iteration is reduced when compared to the previous algorithm.

REFERENCES.

[1] Huang Z, “Extensions to the k-means algorithm for clustering large data sets with categorical values,” *Data Mining and Knowledge Discovery*, (2):283–304, 1998.

[2] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, “An Efficient enhanced k-means clustering algorithm,” *Journal of Zhejiang University*, 10(7):1626–1633, 2006.

[3] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, “A New Algorithm to Get the Initial Centroids,” *Proc. of the 3rd International Conference on Machine Learning and Cybernetics*, pages 26–29, August 2004.

[4] K. A. Abdul Nazeer, M. P. Sebastian,”Improving the Accuracy and Efficiency of the k-means Clustering Algorithm”, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K. ISBN: 978-988-17012-5-1

[5] Madhu Yedla, Srinivasa Rao Pathakota,, T M Srinivasa “Enhancing K-means Clustering Algorithm with Improved Initial Center” International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, 121-125.

[6] Koheri Arai and Ali Ridho Barakbah, “Hierarchical K-means: an algorithm for Centroids initialization for k-means,” department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University,Vol. 36, No.1, 2007.

[7] Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit TM,K Ranjith Kumar, Diagnosis of Diabetes Mellitus based on risk Factors”,International Journal of Computer Applications, Vol.10, Issue No.4, November.2010

[8] M.Franciosi and M.Sacco, "Use of the diabetes risk score and impaired glucose tolerance", Diabetes care Vol.28,no.5, pp 1187-2005.

[9] S. Z. Selim and M. A. Ismail, "K-means type algorithms: a generalized convergence theorem and characterization of local optimality," in *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 6, No. 1, pp. 81--87, 1984.

[10] S. S. Khan and A. Ahmed, "Cluster center initialization for Kmeans algorithm," in *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1293-302, 2004.

[11] S. Anuradha, P. Jyothirmai, Y. Tirumala, S. Goutham, V. HariPrasad," Comparative Study of Clustering Algorithms on Diabetes Data" in International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 3 Issue 6, June – 2014.

[12] http://www.emedicinehealth.com/diabetes .

AUTHORS PROFILE

G.G.Gokilam received her MTech(Computer science and Engineering) in 2010 from PRIST University. Now doing PhD in PRIST University.

Dr. K. Saravanan received his M.Sc. in Computer Science from A.V.C. College (Autonomous), Mayiladuthurai in 1992, M.S. in Software Systems from B.I.T.S. Pilani in 1998, M. Phil. in Computer Science from M.S. University, Tirunelveli in 2003, Ph.D. in Computer Science from PRIST University, Thanjavur in 2011 and M.Tech., in CSE from PRIST University, Thanjavur in 2013. He is having 24+ years of teaching experience and 13+ years of research experience. He guided more than 70 candidates at M.Phil level and guiding 8 candidates at Ph.D. level. His research interest includes in the areas of Big Data Analytics, Data Mining, Cloud Computing, Wireless Sensor Networking and Computational algorithms.