JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF TYPE 2 DIABETIC USING FUZZY C-MEANS CLUSTERING ALGORITHM

Mrs.K.Gandhimathi
Assistant Professor, Department of Computer Science (AI)
PSGR Krishnammal College for Women,
Coimbatore, India.

***Abstract*—** Diabetes is one of the disease increases in humanity across the world-wide. Data mining is a process of extracting the information from a large dataset and transforms it into understandable structure. Medical data mining has been a great capability for finding hidden patterns from the large data sets of the medical dataset. The Data mining techniques used for the prediction of diseases like heart diseases, cancer, kidney stones, EEG etc. Prediction of Diabetes is an emerging and fastest growing technology in the medical analysis data. This research paper concentrates on the clustering method for grouping diabetic data based on cluster head attributes. In this paper the popular clustering algorithm Fuzzy C-Means (FCM) is selected and analyzed based on their fundamentals by using diabetes dataset. The algorithm performance is tested based on its execution time range values. The values are displayed in chart to predict best algorithm.

Keywords: Fuzzy C-Means (FCM), Clustering, and Diabetes dataset.

I. INTRODUCTION

Diabetes is a scientific disease is characterized by the hyperglycemias due to the absolute or relative efficiency of insulin. It causes raise in increasing the blood sugar range. Diabetes is a syndrome that affect slot of peoples in the world. If the diabetes is not recognized at the right time and to treat properly at an early stage for patients, it would affect people in various complications in their body. The main symptoms of the diabetes include number of times urination, severe hunger, and unexpected weight loss etc, The other symptoms are that burning paining legs, feet, arms, fruity smell of breath and sweat etc.

Globally, many chronic diseases are prevalent in the developing and developed in all over countries. Diabetes mellitus (DM) is known as diabetes is one of the metabolic disorders which may cause blood sugar, by producing more significant or a smaller amount of insulin. Diabetes affects the different parts of the human body parts such as eyes, kidneys, heart, and nerves. The prevention and detection of disease in the early stages could likely save human lives. Most common diabetes are Type-I diabetes and Type-II diabetes. In the Type I disease, insulin will not be produced by the human body and 10% of diabetes caused by this type. Type II diabetes is the non-insulin-dependent diabetes caused because of not enough production of insulin by the pancreas or body cells are resistant to absorb it. Therefore, there is a need for to prevent disease and diagnose the disease to save human life from an early death. The conventional method for turning data into information relies on manual data analysis. As data volumes are grow rapidly, this form of the data analysis is slow, expensive, and subjective. The conventional method is becoming completely not practical in many fields and could not meet the need of data analysis [1].

Data mining, also known as the knowledge discovery in databases (KDD) could meet this need by providing tools to discover knowledge from data. Data mining is the development of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web and other information repositories, or data that are streamed into the system dynamically [2]. During the

precedent decades, data mining has been applied to a variety of areas, such as marketing, finance, fraud detection, manufacturing, telecommunications and many scientific fields, including the analysis of medical data. As medical data volumes grow dramatically, there is a growing pressure for efficient data analysis to extract useful, task-oriented information from the enormous amounts of data [3]. Such information may play an important role in the future medical decision-making.

II. LITERATURE SURVEY

Norul Hidayah Ibrahim et.al[4] has presented a new hybrid model by exploring Agglomerative Hierarchical Clustering and Decision Tree Classifier on Pima Indians Diabetes dataset. The experiments result compared performance accuracy of the Decision Tree Classifier against the same classifier augmented with Hierarchical Clustering. The results showed in the hybrid model achieved higher accuracy with 80.8% as compared to 76.9% of the standard model.

Yihong Dong et al., presented an algorithm that applies fuzzy set theory to hierarchical clustering so as to find out clusters with arbitrary shape. Their algorithm can be performed in highdimensional data sets. And also their algorithm creates superior quality clusters than traditional algorithms and scales up well for huge datasets [5]. P.Padmaja et al., found that the overall performance of clustering depends on identifying the quality clusters. Thus most of the clustering algorithms are concerned with efficiently determining the set of clusters in the given dataset. Here, different algorithms were used which are targeted to generate quality clusters [6].

Ashish Ghosh et al., presented an experimental results with justify the potentiality of the APC algorithm both in terms of the clustering quality and execution time compared to other algorithms for a large number of data sets [7]. Adil M.Bagirov, was proposed a new version of the global k-Means algorithm. The starting point for the kth cluster center in their algorithm is computed by minimizing an auxiliary cluster method. The results of numerical experiments on 14 data sets demonstrate the superiority of their algorithm. It requires more computational time than the global k-means algorithm [8].

R.Nithya et al., the performance metrics is evaluated for the clustering algorithms such as Hierarchical Clustering, Density Based Clustering and Simple k-Means clustering algorithms. The algorithms are analyzed by using the trained set parameter based on its class attribute. From the experimental results it is inferred that the k-Means algorithm gives better performance when comparing with the other two algorithms by using the Diabetes dataset [9].

Zeynel Cebeci and Figen Yildiz described that, k-Means algorithm was always extremely faster than Fuzzy C-Means algorithm in all datasets containing the clusters scattering in regular or irregular patterns. Fuzzy C-Means is an algorithm based on more iterative fuzzy calculations, so its execution was found comparatively higher as it is expected [10].

Usha G Biradar and Deepa S Mugali concluded their research it is inferred that the k-Means algorithm gives better performance when comparing with the other two algorithms by using the Diabetes dataset in different tools [11]. Jianpeng Qi et al., described k initial centers powerfully get better the possibility of obtaining best local optima, and multi-round top-n nearest clusters combining approaches the optimal result step by step [12].

K. Saravananathan et.al[13] presented a clustering algorithms k-Means and Fuzzy C-Means are applied for the analysis in order to test their performance by comparing the execution time range. The experimental result shows that k-Means algorithm is better than the Fuzzy C-Means algorithm by their execution time.

III. CLUSTERING METHODS

The unsupervised clustering method Fuzzy C-Means is examined to analyze based on the distance between the different input data points. Cluster centers are arranged for each cluster and the clusters are arranged to the distance between data points. The data points in every cluster are showed by verities of colours and the execution time is calculated in seconds.

Fuzzy C-means clustering:

FCM is a data clustering algorithm in which each data point belongs to a cluster to a degree specified by a relationship grade. Fuzzy C-Means

still uses a cost function that is to be minimized while trying to partition the data set. Clustering analysis the involves assigning data points to clusters such that items in the same cluster are as similar, while items belonging to the different clusters are as dissimilar as possible. Clusters are identified via similarity measures. These similarity measures include distance, connectivity, and intensity. Figure-1 shows the FCM cluster methodology.

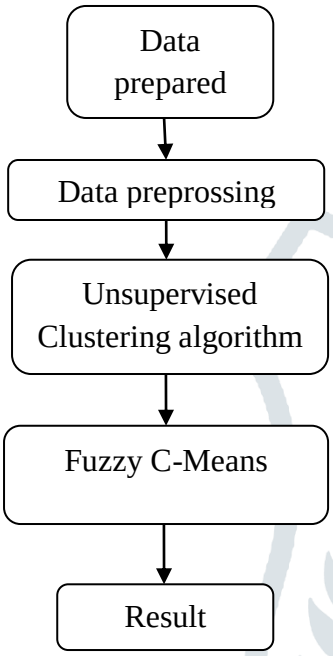

Figure-1: Proposed Framework Description of Clustering

The fuzzy *c*-means algorithm:

- Choose a number of clusters.
- Assign the coefficients randomly to each data point for being in the clusters.
- Repeat until the algorithm has converged.
- Compute the centroid for each cluster (shown below).
- Compute for each data point and compute its coefficients of being in the clusters.

IV. EXPEREMENT RESULTS

In this section the dataset description and clustering algorithm is discussed. The following clustering algorithms Fuzzy C-Means (FCM) is taken for this analysis .Matlab are used for grouping the attributes values in the diabetes dataset.

4.1. Model Description

The model consists of fuzzy C-Means clustering is examined to analyze for remove incorrectly clustered data and optimized data.

4.1.1.Dataset Description

The Dataset consists of information on 800 patients. Tested positive and tested negative indicates whether the patient is diabetic or not, respectively. Each instance is comprised of 10 attributes, which are all numeric.

Attribute details are listed below

- Number of times pregnant (preg)
- Plasma glucose concentration at 2 hours shedule in an oral glucose tolerance test (plas)
- Diastolic blood pressure (pres)
- Triceps skin fold thickness (skin)
- 2-hour serum insulin (insu)
- Body mass index (bmi)
- Diabetes pedigree function (pedi)
- Age (age)
- Diabetic level(Dia_lel)
- Glucose level(Glu_lel)
- Class variable (class)

4.1.2.Data Preprocessing

The quality of the data, to a large extent, affects the result of prediction. This means that data preprocessing is an important role in the model [14]. The number of pregnancies has been little connection with DM [15]. The value 0 indicates the non-pregnant and 1 indicates the pregnant.

The complexity of the dataset was reduced by this process .There are some missing and incorrect values in the dataset due to errors or deregulation. Most of the inaccurate experimental and results were caused by these meaningless values. For example, in the original dataset, the values of diastolic blood pressure and the body mass index could not be 0, which indicates that the real value was missing. To reduce the influence of meaningless values, we used the means from the training data to replace all missing values. The unsupervised normalize filter for attribute was used to normalize all the data.

4.2. Results

The minimum and maximum data point assigned by the Fuzzy C-Means algorithm are shown respectively.

Figure-2 shows the mentioned after clustering the data points are distributed by the Fuzzy C-Means algorithms.

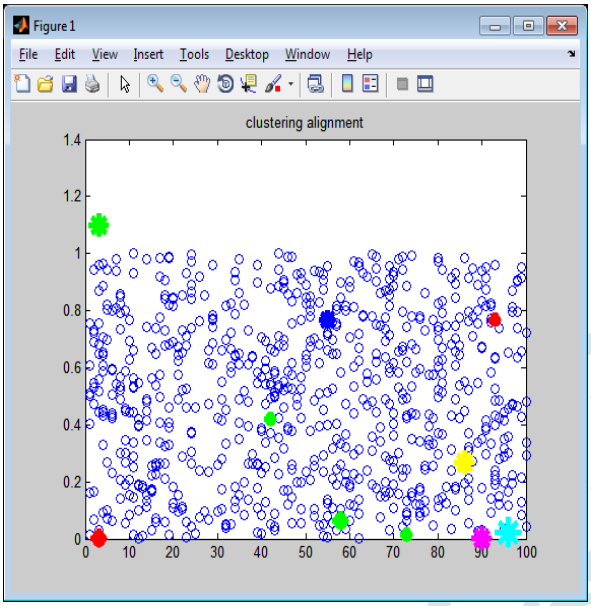

Figure-2: Fuzzy C-Means algorithms

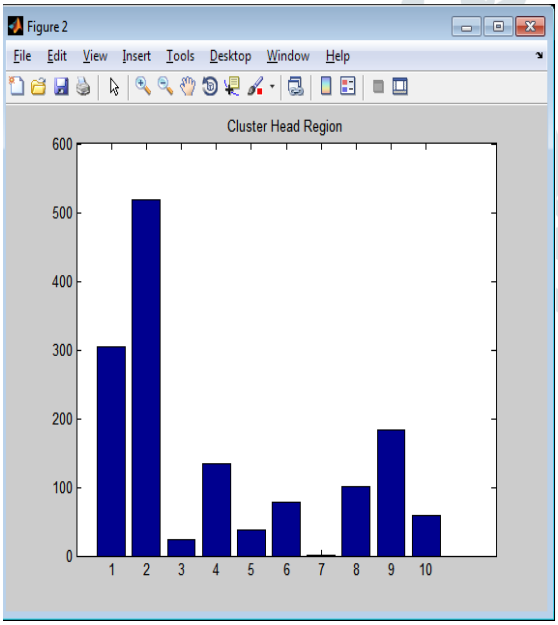

Figure-4: Cluster head region (FCM)

Figure-4 shows the cluster head region chart of after clustering the data points are distributed by Fuzzy C-Means algorithms.

Table-1 shows the comparison value for execution range value of FCM algorithms after clustering the data points.

Attributes	FCM
preg	23.9662
plas	517.6724
pres	100.8576
skin	182.8168
insu	58.4704
bmi	134.865
pedi	77.5418
age	1.1486
Dia_lel	37.2862
Glu_lel	304.1134

Table-1: FCM Execution Range Table

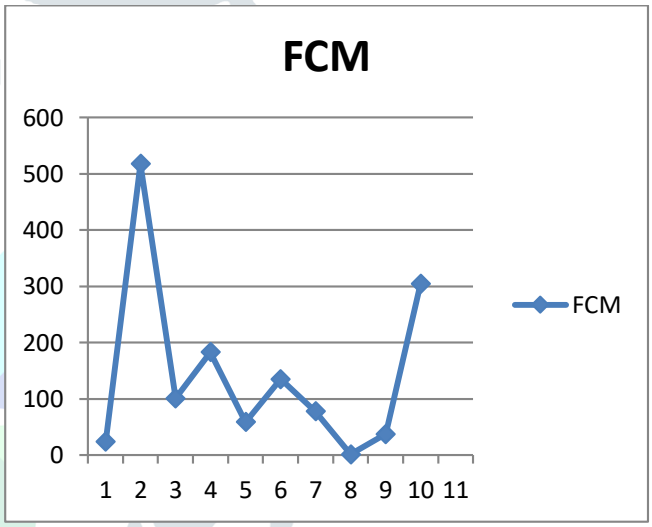

Figure-6: FCM chart

Figure-6 shows the FCM chart for execution range value of FCM algorithm after clustering the data points.

V. CONCLUSION

The clustering algorithms Fuzzy C-Means is applied for the analysis in order to test their performance by the execution range based on attributes. This analysis helps the prediction of diabetes disease for the physicians and the medical experts. The performance of the clustering algorithms differs by its approach and the input data set and also the system dependent. The result shows the Fuzzy C-Means algorithm execution time range. The future plan is optimized data was used as input to classification algorithms for create a model and prediction.

REFERENCES:

[1] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2012.

[2] Sumathi S, Sivanandam S N. Introduction to Data Mining and its Applications[J]. Studies in Computational Intelligence, 2006.

[3] Hasim N, Haris N A. A study of open-source data mining tools for forecasting[C]// International Conference on Ubiquitous Information Management and Communication. ACM, 2015.

[4] Norul Hidayah Ibrahim1 , Aida Mustapha2 , Rozilah Rosli3 , Nurdhiya Hazwani Helmee4 "A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients" International Journal of Engineering and Technology (IJET)2013.

[5] Yihong Donga, Yueting Zhuanga, Ken Chenc, Xiaoying Taib, "A hierarchical clustering algorithm based on fuzzy graph connectedness", Fuzzy Sets and Systems, Vol. 157 (13), pp. 1760– 1774.

[6] P. Padmaja, V. Srikanth, N. Siddiqui, D. Praveen, B. Ambica, V. B. V. E. Venkata Rao, and V.J.P.Raju Rudraraju, "Characteristic evaluation of diabetes data using clustering techniques"International Journal of Computer Science and Network Security, Vol. 8 (11), 2008, pp. 244, 251.

[7] Ashish Ghosh, Anindya Halder, Megha Kothari, and Susmita Ghosh, "Aggregation pheromone density based data clustering", Information Sciences, Vol. 178 (13), 2008, pp. 2816 –2831.

[8] Adil M. Bagirow, "Modified global k- Means algorithm for minimum sum-of-squares clustering problems", Pattern Recognition, Vol. 41 (10), 2008, pp. 3192–3199.

[9] R.Nithya, P.Manikandan, and D.Ramyachitra, "Analysis of clustering technique for the diabetes dataset using the training set parameter", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4 (9), 2015, pp. 166–169.

[10] Zeynel Cebeci and Figen Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures", Journal of Agricultural Informatics, Vol. 6 (3), 2015, pp. 13–23.

[11] Usha G Biradar and Deepa S Mugali, "Clustering Algorithms on Diabetes Data: Comparative Case Study", International Journal of Advanced Research in Computer Science, Vol. 8 (5),2017, pp. 550–552.

[12] Jianpeng Qi, Yanwei Yu, Lihong Wang, Jinglei Liu and Yingjie Wang "An effective and efficient hierarchical k-Means clustering algorithm", International Journal of Distributed Sensor Networks, Vol. 13 (8), 2017, pp.1–17.

[13] K. Saravananathan and T. Velmurugan: Cluster based performance analysis for Diabetic data, International Journal of Pure and Applied Mathematics,2018.

[14] Karim M. Orabi1 , Yasser M. Kamal, and Thanaa M. Rabah. Early Predictive System for Diabetes Mellitus Disease. ICDM 2016, LNAI 9728, pp. 420– 427, 2016.

[15] B.M. Patil, Hybrid prediction model for Type-2 diabetic patients.Systems with Applications 37 (2010) 8102–8108.

[16]. Qinpei Zhao, Mantao Xu, and Pasi Fränti "Sum-of-Squares Based Cluster Validity Index and Significance Analysis.