Intelligent Manufacturing and Cloud Computing
I.S. Jesus and K. Wang (Eds.)

doi:10.3233/ATDE241346

Performance Analysis of K-Means and Fuzzy C-Means (FCM) Clustering Algorithms for Diabetic Dataset

Thambusamy Velmurugan[a,1] and K. Emayavaramban[b]

[a]*Associate Professor, PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai, India*
[b]*Guest Lecturer, Loganatha Narayanaswamy Government College, Ponneri, India*

Abstract. In the contemporary era, with its incidence rising at an alarming rate, diabetes has become a major global health concern. This work presents a data mining approach that compares the K-Means and Fuzzy C-Means (FCM) clustering algorithms to answer the increasing demand for accurate diabetes management and prediction in a wide range of datasets. As data mining is essential for drawing insightful conclusions from large, complicated datasets, the study concentrates on using these methods to improve the precision and effectiveness of diabetes prediction models. The diabetes dataset is divided into separate clusters using FCM and K-Means after preprocessing steps to manage missing values and outliers. Algorithms often define the clustering process's conclusion and the effectiveness of its domain application. Two significant clustering algorithms centroid-based K-Means and representative object-based FCM (Fuzzy C-Means) clustering algorithms are compared in this research work. These techniques are used, and the effectiveness of the output clustering is to assess performance. The evaluation metrics accuracy, sensitivity, specificity, and AUC-ROC highlight the model's superior performance over individual clustering techniques. This work addresses the urgent problem of diabetes currently, which advances the field of diabetes prediction and highlights the crucial role of data mining, which plays in healthcare analytics over early detection and intervention.

Keywords: Data mining Techniques; diabetic dataset; Clustering Algorithms; k-Means clustering; Fuzzy C-Means clustering

1. Introducton

Since processing big amounts of data is a difficult work, data analysis is seen as a significant and valuable tool in the area of software, and its application has become more and more popular. To be more exact, data mining is the study of observational datasets with the goal of revealing links between datasets that were previously unknown and elegantly summarizing the data so that data users might easily understand and benefit from it [1]. Additionally, it enables data description through sequential analysis, association, and clustering visualization. Primarily, data clustering is a data description approach that is widely employed in the fields of machine learning, data mining, and recognition of patterns, image analysis, and bio-informatics, among other

[1]corresponding author: Thambusamy Velmurugan, velmurugan_dgvc@yahoo.co.in

domains. Another well-known method for classifying data is cluster analysis, which divides a dataset into groups according to similarities within a single cluster and differences across distinct clusters [2].

Fuzzy memberships, or the probability of belonging to distinct clusters, are assigned to data points in the Fuzzy C-Means (FCM) clustering technique, which can be used to predict diabetes. FCM recognizes the inherent ambiguity in medical datasets and enables a more sophisticated portrayal of data points in the setting of diabetic prediction [3]. Clustering algorithms are used by K. Saravananathan et al. [4] described include Fuzzy C-Means and K-Means, which are used for the analysis to compare its execution time range and assess their performance. Based on execution time, the experimental result demonstrates that the k-Means method performs better than the fuzzy C-Means approach. This work utilised the k-Means and Fuzzy C Means algorithms to find the diseases in the diabetes data [5] [6].

2. Materials and Methods

This section discusses the problem definition for this research work. Organizing the many forms of unorganized information in a medical record is the main challenge for healthcare data mining activities. It is necessary to comprehend the patterns and important phrases in the medical record of a person, which can vary greatly, in order to accurately detect disease from medical data. The dataset is preprocessed to remove redundant data, missing data, and unnecessary features. The cleaned diabetes dataset is then used in the prediction process to determine whether or not impacted individuals would be identified using fuzzy C and K means algorithms. The.csv file format for the diabetic dataset can be obtained from the USI repository. The dataset is named diabetic.csv in the file.

2.1. Data Source

The data set provided below has been used to assess and examine the categorization and forecasting performances of different approaches. The UCI repository provides the data set, which is then combined to create a new, customized data set. There are 605 records in all that are taken into account for prediction. There are 19 features in the data set, and the performance analysis has been conducted on 605 occurrences of records. This work calculated the average values from the data set and showed them in Table 1. In table 1, both reflect the characteristics being used and the data set that is utilized for the prediction process. There are 9 qualities per data point, including all the characteristics of diabetes mellitus.

Table 1. Details of Dataset

Patient ID	Blood glucose	Pre Breakfast blood glucose	Post breakfast blood glucose	HbA1c	Age	BMI	TriGly	BP
1	141	112	150	6	38	25.2	154	82
2	143	113	152	6.1	41	25.3	155	83
3	148	138	155	6.3	43	25.6	160	82
4	131	118	169	6.3	38	25.1	152	80
5	132	123	159	6.1	41	25.5	153	82
6	161	151	186	6.5	44	25.8	161	85
7	156	133	180	6	43	25.4	163	83
8	131	116	158	6	42	25.2	160	81
9	151	143	181	6	37	25.2	153	80
10	152	112	192	6.1	40	25.4	156	90

Each of the parameters has been determined to be significant in the diabetes prediction based on the values found in the data set. Diabetes has been identified if the blood glucose level exceeds the reference range of 145. Similarly, if a person's pre-breakfast blood glucose level is higher than 120, they may be diagnosed with diabetes. Additionally, if the blood glucose level after breakfast exceeds 160, it has been deemed diabetes. Conversely, diabetes may be diagnosed if the HbA1c score is greater than 6.5. In light of these limitations, the method's accuracy has been assessed using the following metrics.

2.2. Methodology

The analysis is conducted using two different unsupervised clustering techniques, K-Means and fuzzy C-means, depending on the distance between the various input data points. Figure 1 shows the overall methodology of the research work. The clustering algorithms are used after the preprocessing of the data, then the performance of the algorithms are measured based on the performance metrices and the results are analyzed.

Figure 1. Research Methodology

Clustering algorithms plays a vital role in all fields of research not only in medical area. Also, it includes many divertible data bases and yields very effective results. The results are analysed based on the number of clusters in which the clustering methods produced.

3. Experimental Results

In this section, the existing unsupervised machine learning methods are used to evaluate the diabetic data set and the produced results are compared with the other methods. K means clustering and Fuzzy C means clustering are used in this research to find the optimal method.

3.1. K-Means Clustering

One of the most simple and well-liked unsupervised data mining algorithms is K-means clustering. The technique keeps the centroids as minimal as feasible by first determining the k amount of centroids, after which it assigns every point of data to the closest cluster. The term "means" describes the process of averaging the data, or locating the centroid [7]. The first set of randomly chosen centroids, which serve as the

starting points for each cluster, are the first step in the learning process of the K-means algorithm in data mining

K-Means clustering algorithmic steps:

Step 1: Set K: Select K as the appropriate number of clusters.

Step 2: Initialization: Selecting k beginning locations to serve as rough approximations of the cluster the centroids. They are considered to be the original beginning points.

Step 3: Classification: Analyze every point on the data set and place it in the cluster with the closest centroid.

Step 4: Centroid computation: The new k centroids must be computed once every point on the data set has been allocated to a cluster.

Step 5: Convergence criteria: The steps (3) and (4) must be continued until either the centroids stop moving or no point modifies its cluster assignment.

The number of clusters to be found and the values of the first beginning point are the parameters that are entered of the clustering algorithm [8]. Equation is used to find the distance between each of the sample data point and each initial starting value once the initial values for starting have been determined [9]. Next, every data point is positioned within the cluster linked to the closest beginning point [10]. The algorithm's steps (3) and (4) are then repeated using the updated centroids for the new initial values.

Table 2. Results of k-Means Clustering by Attributes

Cluster No.	Attributes	No of Patients
1	Fasting + PB + HbA1c	243
2	Fasting + PB	95
3	Only Fasting	103
4	Only PB	76
5	Only HbA1c	37
6	Not Affected	51
	Total	605

Table 3. Results of K means Clustering by Age

Cluster No.	Age	No of Patients
1	<40	52
2	40 - 50	99
3	51 - 60	158
4	>60	245

Table 2 shows the results obtained by k-Means Clustering based on attributes of clustering phase. The parameters are calculated by using the attributes. The No of Patients affected are calculated by Fasting + PB + HbA1c, Fasting + PB, Only Fasting, Only PB, Only HbA1c and Not Affected are given respectively 243, 95, 103,76, 37 and 51.

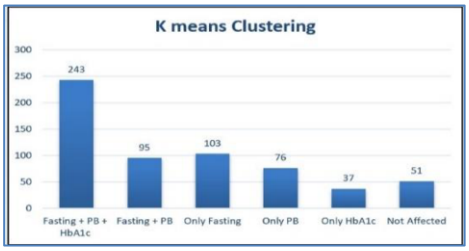

Figure 2. Results of K means Clustering by Attributes

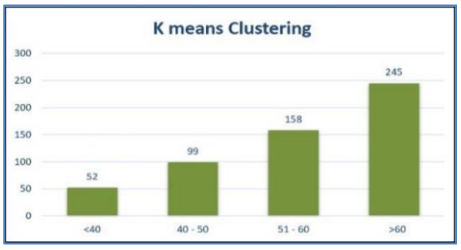

Figure 3. Results of K means Clustering by Age

In figure 2, Results of K means clustering by attributes are shown in graphical representation, The clustered attributes Fasting + PB + HbA1c, Fasting + PB, Only Fasting, Only PB, Only HbA1c and Not Affected with data labels are shown in the figure. Table 3 shows the results obtained by K means Clustering based on age of clustering phase. The number of patients is calculated by the age group. The results are

<40 is 52, 40 – 50 are 99, 51 – 60 are 158 and >60 is 245. In figure 3, Results of K means clustering by age are shown in graphical representation, The clustered attributes are <40 is 52, 40 – 50 are 99, 51 – 60 are 158 and >60 is 245 with data labels are shown in the figure. This process keeps going until either the centroids stop moving or there are no further data point changes [11].

3.2. Fuzzy C-Means Clustering

The Ruspini Fuzzy Clustering Theory served as the foundation for the 1980s proposal of the FCM clustering algorithm, which was made possible by the advancement of fuzzy theory [12]. FCM is a method for data clustering [13, 14] that groups a data set into n cluster with every point of data in the dataset associated with each cluster and it will be highly connected to both the cluster itself and another data point [15].

Fuzzy C-Means clustering algorithmic steps:

Fixing c, where c is (2<=c<n), choosing a value for parameter 'm' and initializing the partition matrix $U^{(0)}$ come next [16]. In this algorithm, each step will be denoted by the letter 'r' where r = 0, 1, 2,...

For every step, we must compute the c center vector, or $\{V_{ij}\}$.

$$v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_{kj}}{\sum_{k=1}^{n} (\mu_{ij})^m} \tag{1}$$

Determine the $D_{[c,n]}$ distance matrix.

$$D_{ij} = \left(\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right)^{1/2} \tag{2}$$

For the r^{th} step, modify the partition matrix $U^{(R)}$ as

$$\mu_{ij}^{r-1} = \left(1 \Big/ \sum_{j=1}^{c} (d_{ik}^{r} / d_{jk}^{r})^{2/m-1} \right) \tag{3}$$

Table 4. Results of Fuzzy C means Clustering by Attributes

Cluster No.	Attributes	No of Patients
1	Fasting + PB + HbA1c	235
2	Fasting + PB	92
3	Only Fasting	106
4	Only PB	78
5	Only HbA1c	41
6	Not Affected	53
	Total	605

Table 5. Results of Fuzzy C means Clustering by Age

Cluster No.	Age	No of Patients
1	<40	41
2	40 - 50	104
3	51 - 60	152
4	>60	255

If $\| U^{(k+1)} - U^{(k)} \| < \delta$, we can end the process; if not, we must go back to step 2 and update the cluster centers and membership grades for each data point iteratively [17]. Table 4 shows the results obtained by Fuzzy C means Clustering based on attributes of clustering phase. The parameters are calculated by using the attributes. The No of Patients affected are calculated by Fasting + PB + HbA1c, Fasting + PB, Only Fasting, Only PB, Only HbA1c and Not Affected are given respectively 235, 92, 106,78, 41 and 53.

In figure 4, Results of Fuzzy C means clustering by attributes are shown in graphical representation, The clustered attributes Fasting + PB + HbA1c, Fasting + PB, Only Fasting, Only PB, Only HbA1c and Not Affected with data labels are shown in the figure. Table 5 shows the results obtained by Fuzzy C means Clustering based on

age of clustering phase. The number of patients is calculated by the age group. The results are <40 is 41, 40 – 50 are 104, 51 – 60 are 152 and >60 is 255.

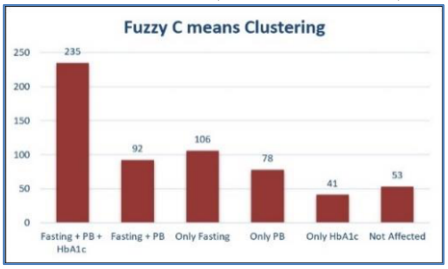

Figure 4. Results of Fuzzy C means Clustering by Attributes

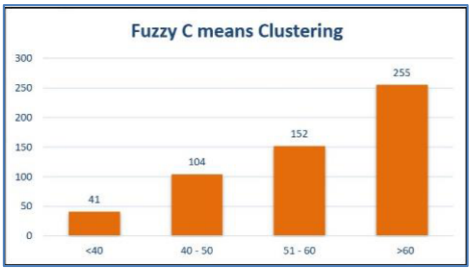

Figure 5. Results of Fuzzy C means Clustering by Age

In figure 5, Results of Fuzzy C means clustering by age are shown in graphical representation, The clustered attributes are <40 is 41, 40 – 50 are 104, 51 – 60 are 152 and >60 is 255 with data labels are shown in the figure. This algorithm functions and has a structure that is essentially comparable to the K-Means algorithm [18, 20].

3.3. Performance Metrices

Performance metrics provide information about the dataset's performance. The evaluation criteria used to evaluate the recommended scheme's presentation are Precision, Recall, and F-measure. Here, conventional count values are taken advantage of, including True Positive (Tp), True Negative (Tn), False Positive (Fp), and False Negative (Fn).

3.4. Results and Discussions

The techniques are used in conjunction with a prediction process, and the outcomes are tracked and compared. The following outcomes are the product of the procedure.

Table 6. Performance of K-Means and Fuzzy C- Means Clustering by Attributes

Cluster No.	Attributes	K means	Fuzzy C means
1	Fasting + PB + HbA1c	243	235
2	Fasting + PB	95	92
3	Only Fasting	103	106
4	Only PB	76	78
5	Only HbA1c	37	41
6	Not Affected	51	53

Table 7. Performance of K means and Fuzzy C means Clustering by Age

Cluster No.	Age	K means	Fuzzy C means
1	<40	52	41
2	40 - 50	99	104
3	51 - 60	158	152
4	>60	245	255

Table 6 shows the results obtained by K means clustering and Fuzzy C means Clustering based on attributes of clustering phase. The parameters are calculated by using the attributes. By using the K means clustering the No of Patients affected are calculated by Fasting + PB + HbA1c, Fasting + PB, Only Fasting, Only PB, Only HbA1c and Not Affected are given respectively 243, 95, 103, 76, 37 and 51. Then displayed values using the Fuzzy C means clustering, the No of Patients affected are calculated by Fasting + PB + HbA1c, Fasting + PB, Only Fasting, Only PB, Only HbA1c and Not Affected are given respectively 235, 92, 106,78, 41 and 53.

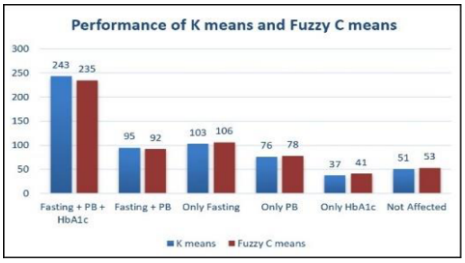

Figure 6. Performance of Clustering by Attributes

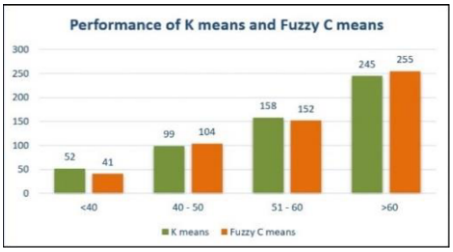

Figure 7. Performance of Clustering by Age

Figure 6 shows the Performance of K means and Fuzzy C means Clustering by Attributes in graphical representation with data labels. All the attributes are clustered by both K means and Fuzzy C means clustering algorithms. Table 7 displays the Performance of K means and Fuzzy C means Clustering by Age. All the attributes are clustered by both K means and Fuzzy C means clustering algorithms with respect to <40, 40 – 50, 51 – 60 and >60. Figure 7 shows the Performance of K means and Fuzzy C means Clustering by Age in graphical representation with data labels. The number of patients are clustered by both K means and Fuzzy C means clustering algorithms with respect to <40, 40 – 50, 51 – 60 and >60.

Table 8. Comparative Performance

Algorithms	Precision	Recall	F-measure
K means	80.34	81.48	85.82
Fuzzy C means	73.05	75.33	79.67

Figure 8. Comparative performance of algorithms

Table 8 shows the performance analysis of K means algorithm and Fuzzy C means algorithm. The two approaches are evaluated for effectiveness using the following metrics: f-measure, recall, accuracy, and precision. The precision, recall, & f-measure rates for each of the two algorithms are shown in Figure 8. The K means algorithm yields an 80.34% precision, 81.48% recall, and 85.82% f-measure value. The Fuzzy C means Algorithm obtains a precision of 73.05%, a recall score of 75.33%, and F-measure score of 79.67%. It is clear from the data that the k-Means algorithm performs better than the Fuzzy C Means algorithm.

4. Conclusion

The study compares the results range along with various analyses based on features from the dataset to examine the performance of the algorithms for clustering K-Means and Fuzzy C-Means For doctors and other medical professionals, this analysis aids in the prognosis of diabetic disease. The technique, the input data set, and the system dependence all affect how well the clustering algorithms function. The number of the final cluster (k) must be specified in advance when using the K-Means partitioning-based clustering technique. We can infer from the findings that the K-Means method performs better than FCM algorithm. Because fuzzy measures calculations are involved,

FCM requires more computing time than K-Means clustering even though it yields results that are similar to K-Means clustering within the algorithm. It can also yield approximation answers more quickly. In addition to picture retrieval, its primary applications have been in the identification of association rules and functional relationships. In summary, it appears that the K-Means method outperforms the Fuzzy C-Means approach.

References

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

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

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

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

[5] Norul Hidayah Ibrahim, Aida Mustapha, Rozilah Rosli, Nurdhiya Hazwani Helmee, "A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients" *International Journal of Engineering and Technology (IJET)*, 2013.

[6] Yihong Donga, Yueting Zhuanga, Ken Chenc, Xiaoying Taib, "A hierarchical clustering algorithm based on fuzzy graph connectedness", *Fuzzy Sets and Systems*, Vol. 157, No.13, pp. 1760– 1774.

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

[8] B.M. Patil, "Hybrid prediction model for Type-2 diabetic patients", *Expert Systems with Applications*, Vol.37, 2010, pp. 8102–8108.

[9] A. Rakhlin and A. Caponnetto, "Stability of K-Means clustering", *Advances in Neural Information* Processing *Systems*, MIT Press, Cambridge, MA, 2007, pp. 216–222.

[10] Rui and J. M. C. Sousa, "Comparison of fuzzy clustering algorithms for Classification", *International* Symposium *on Evolving Fuzzy Systems*, 2006, pp. 112-117.

[11] 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, No.11, 2008, pp. 244, 251.

[12] Ashish Ghosh, Anindya Halder, Megha Kothari, and Susmita Ghosh, "Aggregation pheromone density-based data clustering", *Information Sciences*, Vol. 178, No.13, 2008, pp. 2816 –2831.

[13] Adil M. Bagirow, "Modified global k- Means algorithm for minimum sum-of-squares clustering problems", *Pattern Recognition*, Vol. 41, No.10, 2008, pp. 3192–3199.

[14] 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, No.9, 2015, pp. 166–169.

[15] Zeynel Cebeci and Figen Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures", *Journal of Agricultural Informatics*, Vol. 6, No.3, 2015, pp. 13–23.

[16] 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, No.5, 2017, pp.550–552.

[17] 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, No.8, 2017, pp.1–17.

[18] L. Hui, "Method of image segmentation on high-resolution image and classification for land covers", *Fourth International Conference on Natural Computation*, Vol. 5, 2008, pp. 563-566.

[19] S. Borah and M. K. Ghose, "Performance analysis of AIM-K-Means and K-Means in quality cluster generation", *Journal of Computing*, Vol. 1, 2009.

[20] T. Kanungo and D. M. Mount, "An Efficient K-means Clustering Algorithm: Analysis and Implementation", *Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No.7, 2002.