Diabetes Diagnosis using Clustering & Classification

Yogesh Rasam[1] Saish Shet[2] Narayan Chari[3] Pawan Toralkar[4] Akash Patil[5]
[1,2,3,4,5]Agnel Institute of Technology & Design, India

Abstract— In modern times a lot of data is being collected all over the world particularly in the field of medical sciences. All this data needs to be stored in an organised manner so that its retrieval is efficient enough. As the data increases it becomes more and more challenging to organise it. Clustering and classification are two major methods primarily used to organise the data. Clustering is the process of grouping similar objects and classification is the process of categorising the data. Usually these methods are used independently but in the proposed method we are combining both this methods to obtain better results. It uses patient's data where they are diagnosed for presence or absence of diabetes. The data is first normalized using Min-Max normalization method. Normalized data is then clustered using Bisecting K-means, K-Medoids and DBSCAN. The output obtained from clustering is then given to Naïve Bayes classifier. The output determines which of the combination for data processing the best is.

Key words: Diabetes, Clustering & Classification

I. INTRODUCTION

Data mining is one of the important field in today's world. Data mining has become a field of interest to many due to ever-increasing data in logistics, transport, aviation and medical sectors. Data mining offers new and emerging techniques to handle such a huge amount of data in a way, so that ease of handling the data and retrieving related information from it which is needed by the user is improved. Data mining offers different algorithms. Each of these algorithms has special characteristics and thus can be used for different applications in order to make data processing task relatively simpler as well as efficient. Data mining reveals patterns and relationship within a huge amount of data which in turn makes its analysis easier.

Two of the major and well-known fields of data mining are clustering and classification. Generally in the normal practice both these techniques are used independently. In the proposed method we want to combine these two methods with combination of clustering and classification algorithm in order to get the best combination possible to get maximum efficiency for the given data. Bisecting K-means, K-medoids and DBSCAN are the clustering algorithms which will be used in combination with Naïve Bayes classifier. The combination of both these methods will lead to discovery of a more robust combination to handle data processing.

II. RELATED WORK

1) The working principle of the related system is as follows. It consists of 3 steps:
- Data pre-processing is done by replacing the missing values by mean. [1]
- The pre-processed data is given as an input to K-Means to form clusters and to remove outliers, inconsistent and noisy data and the reduced data is used to select the optimal features with genetic algorithms. [1]
- This reduced data set is then used to classify using SVM classifier to achieve better accuracy when compared to existing available methods.10 fold cross validation is used in order to increase the reliability of the classification algorithm. [1]

2) The main idea of the proposed method, is to sample the given data set into equal parts, and to compute the means of these initial clusters as the seeds of the method. More details are presented in the following pseudo-code:

Input: A data set X whose cardinality is n and an integer k
Output: k seeds cj
p=round(n/k)
For j=1:k
c(j,:)=mean(X(1+(j-1)*p:j*p,:))
end For [I,C]=kmeans(X,k,'start',c)
Like kkz, this method has a complexity of O(nkd), but practically our experimental results showed that it is often faster than kkz. [2]

III. PROPOSED ARCHITECTURE

The proposed concept of the diabetes diagnoses system is shown in figure 1.1. The input that is given to the proposed system is the diabetes data set. Data pre-processing techniques are then applied on the data set to normalize the record's present in the data set. The processed data is then given to the clustering and classification algorithms for further processing. Clustering is carried out using K-medoid clustering algorithm, Bisecting K-means clustering algorithm, DBSCAN clustering algorithm and Classification is carried out using Naïve Bayes classifier. Hence the accuracy can be improved using a hybrid approach.

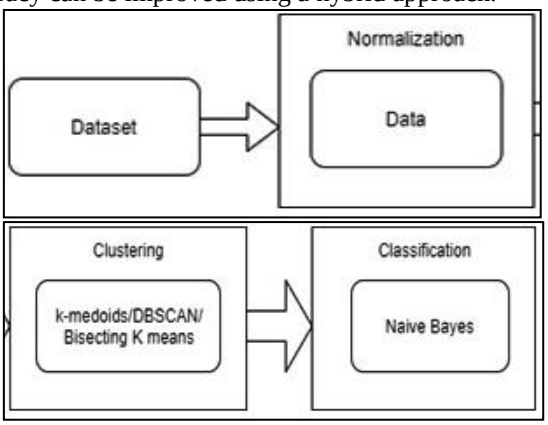

Fig. 1.1: Block Diagram of Proposed System

In the proposed method, we first apply Min-Max normalization technique to data set to bring it in the range of 0 to 1. Next the preprocessed data is given as an input to clustering algorithms. We use KKZ initial centroid selection method to get first two initial points and not choosing them randomly. Further, the output obtained from clustering is given to classifier along with the training data to determine the class of each data tuple.

IV. IMPLEMENTATION

A Graphical User Interface (GUI) is the best way to visualise all of these operations in an organised manner. Here we can select the data file from the system along with the parameters needed from user for various algorithms. A series of buttons are available for each operation as shown in figure 1.2.

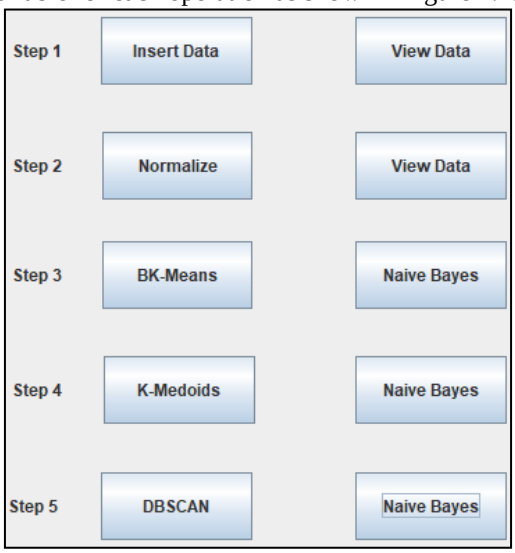

Fig. 1.2: Graphical User Interface (GUI)

Browse button is used to access the directories to insert the data into database as shown in figure 1.3

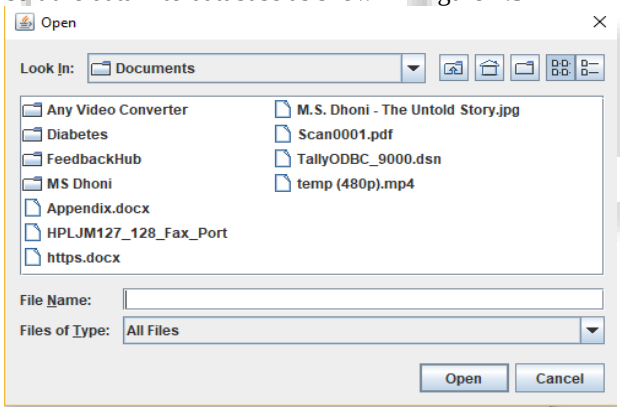

Fig. 1.3: Showing System Directory

We can also insert the necessary data for execution via an input panel as shown in figure 1.4.

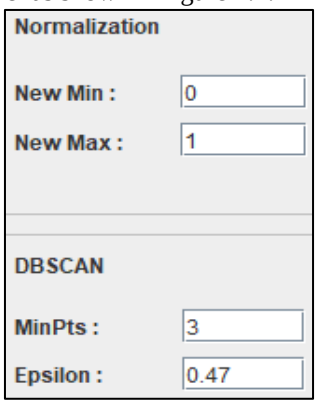

Fig. 1.4: Data Insertion Panel

Once the operations are complete we can see two buttons visualize & accuracy which can be used to see the final results shown in figure 1.5.

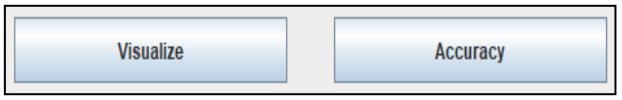

Fig. 1.5: Visualize & Accuracy Button

V. RESULT ANALYSIS

We have compared three clustering algorithms and combined each of them with one classifier. The three clustering algorithms are K-Medoid, Bisecting K-means and DBSCAN and the classifier used is Naive Bayes.

The clustering results are as shown in figure 1.6.

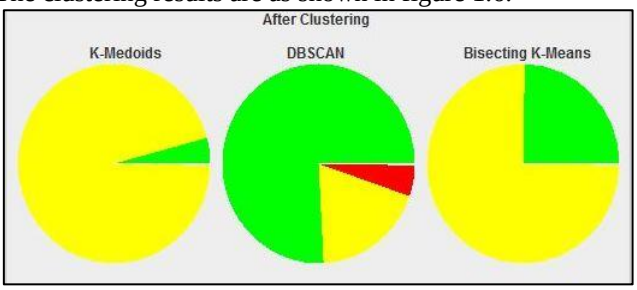

Fig. 1.6: Clustering Results

The classification results are as shown in figure 1.7.

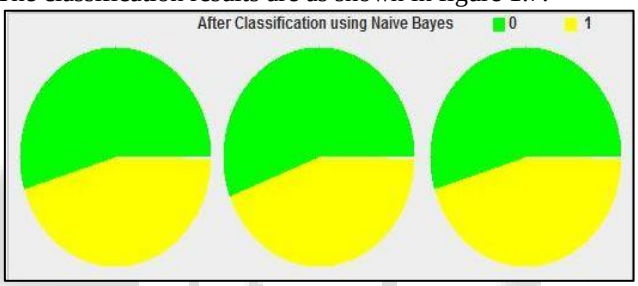

Fig. 1.7: Classification Results

Algorithm	After Clustering	After Classification
K-medoids	56%	81%
Bisecting K-means	45%	81%
DBSCAN	48%	79%

Table 1.1: Accuracy of the Algorithms After Clustering & Classification

From the results shown in table 1.1 we can say that combination of K-medoids along with Naïve Bayes classifier is the best combination among all tested combinations.

VI. CONCLUSION

The hybrid approach of combining clustering and classification indeed shows promising results. The accuracy of classifying the tuples correctly is tremendously increased with this approach. The overall performance is much better than the traditional approach of using clustering and classification solely. The combination of K-Medoids clustering along with Naïve Bayes classifier shows promising results as compared to Bisecting K-Means or DBSCAN. The major drawback of this method is that the performance of clustering algorithm will change depending on the amount of input data. Further improvements can be done to get the results more efficiently and accurately.

REFERENCES

[1] T. Santhanam a, M.S Padmavathi b. "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis". Procedia Computer Science, vol.47, pp.76-83, 2015.
[2] Omar Kettani, Faical Ramdani. "A Fast Deterministic Kmeans Initialization". International Journal of Applied Information Systems, vol.12-No.2, pg. 6-11, 2017.

