IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 6, Ver. I (Nov.- Dec. 2017), PP 13-19
www.iosrjournals.org

Back Propagation Based K-Medoid Algorithm for Predicting Type-2 Diabetes

Shubhangi Pahwa* and Dr. P.S. Maan
Department of Computer Science DAVInstitute of Engineering and Technology, Jalandhar, Punjab 144008
Department of Computer Science DAVInstitute of Engineering and Technology, Jalandhar, Punjab 144008
Corresponding Author: Shubhangi Pahwa

ABSTRACT*: Clustering is a technique to analyze the data in efficient manner and generate required information. To cluster the dataset, the technique of k-medoid is applied which is a partitioning based method. It uses k as a parameter, divide n objects into k clusters. In K-medoid clustering data points are used as the centers of the clusters. Here, we will work on prediction analysis in which final clustered data will be analyzed and evaluated according to the requirement. In order to analyze the clusters the relationship among the attributes of the dataset must be known which is a difficult task in case of complex dataset is. In this work, a technique will be proposed using back propagation to derive relationship among the attributes of the complex dataset for better analysis of clustered data. This technique will lead to the improvement in accuracy, precision, F-measure and recall of the clusters.*

Keywords*: Clustering, K-medoid clustering, back Propagation, Diabetes, Prediction Analysis*

Date of Submission: 25-10-2017 Date of acceptance: 07-11-2017

I. Introduction

Data mining is the technology of discovering interesting patterns from large amount of data. It is extraction of implicit previously unknown and potentially useful information from data. Data mining is also called as extraction of hidden patterns.It also known as knowledge mining from data, knowledge extraction, It may fully automate or semi-automated process to discover knowledge that is useful for user. [10]

It also a process of finding a hidden information in data base this process may use one or more computer learning techniques to automatically analyses and extract knowledge from data contain within the database, it is part of knowledge discovery process.

Data mining applies algorithms to large data to produce models or patterns interesting to the user and will extract the hidden patterns.

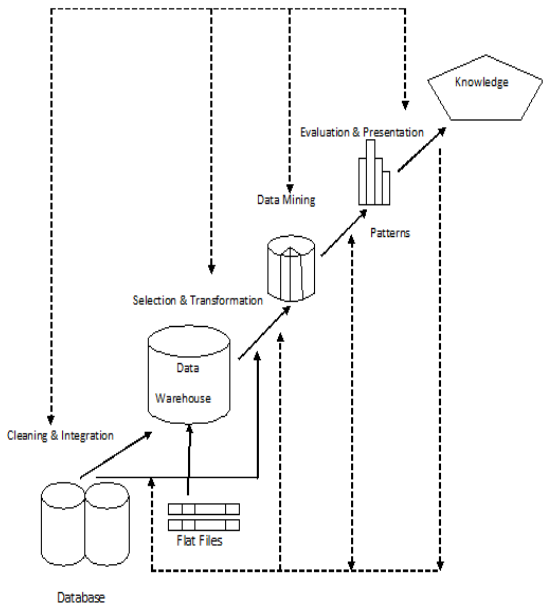

Figure.1.1. Data Mining in KDD process

1.1 CLUSTERING IN DATA MINING: Clustering is an unsupervised learning technique. It is a process of partitioning a set of data into a set of significant sub-classes, called cluster [1]. Data is organized into clusters such that there is high intra-cluster similarity and low inter-cluster similarity. It is a main task of exploratory data mining and a common technique for statistical data analysis [3]. It is implemented in many fields including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

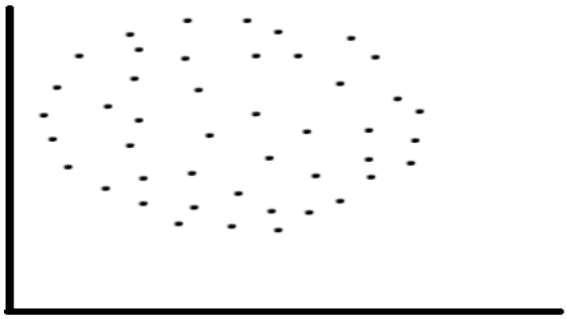

Figure.1.2 Data without clustering

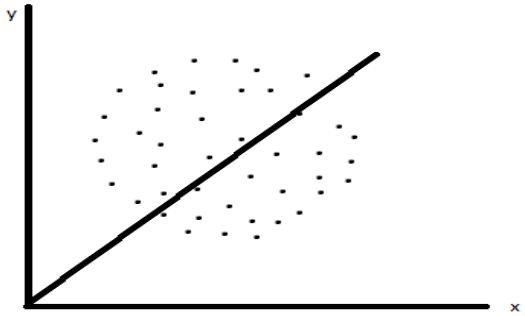

Figure.1.3 K-Medoid Clustering

There are several application of clustering [11] such as data/text mining, image processing, web mining, voice mining. There are several methods of clustering [12].

The major primary clustering methods can be classified into following categories [13]:

PARTITIONING METHODS**:** The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between separate clusters. Most partitioning methods are distance-based. Given k, the number of partitions to create, a partitioning method creates an initial partitioning and then uses an iterative relocation technique that attempts to improve the partitioning by transferring objects from one group to another [5].

HIERARCHICAL METHODS**:** In this method hierarchical breakdown of the given set of data objects is created. It can be classified into two approaches agglomerative and divisive. Agglomerative approach is the bottom up approach. This approach starts with each object forming a separate group [7].

DENSITY BASED METHODS**:**Generally partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and come across difficulty in discovering clusters of random or arbitrary shapes. So for arbitrary shapes new methods are used known as density-based methods which are based on the notion of density. In these methods the cluster is continue to produce as long as the density in the neighbourhood cross some threshold [6].

GRID BASED METHODS**:** Grid based methods quantize the object space into a fixed number of cells that form a grid structure. It is a fast method and is independent of the number of data objects and depends only on the number of cells in each dimension in the quantized space [8].

II. Related Work

The relationship of various indicators of body fat distribution including SAD (supine sagittal abdominal diameter) is investigated with CVD (Cardiovascular Diseases) risk factors [1]. The study was performed on 146 women and 83 men aged between 67 to 78 years. Indicators used in this study were waist circumference, BMI, waist to hip ratio, SAD, SAD to thigh ratio and waist to height ratio. The study showed that waist and SAD are more closely related with risk factors of CVD in old age.

Aastha Joshi, Rajneet kaur et.al.[2] provided the comparison between various clustering techniques like partitioning method hierarchical method, density based method, grid method. Clustering algorithm are mainly

used to manage data, categorized data for data compression, model creation and also used for outlier discovery etc. Main motive each clustering technique is to find cluster center that represent each cluster. Then input data is compared with each cluster center, and then based on these cluster centers defined which cluster is nearest or similar one. Partitioning method like k-mean clustering algorithm is used for large datasets, as number clusters is increased its performance is also increased. But its use is limited to numeric values. Hierarchical algorithms are used for categorical data. DBSCAN is adopted to find cluster of arbitrary shapes.

It was estimated in paper [3] thatthe number of people with diabetes of all age groups and the prevalence of diabetes for the year 2000 and 2030. For this they have used the data from a limited number of countries on prevalence of diabetes of all age groups. This data was extrapolated to 191 member states of WHO and was applied to the population estimates of United States for year 2000 and 2030.

Various techniques to measure the body fat distribution have been discussed [4] as it is believed that body fat is hazardous for mortality and morbidity. According to the studies in this paper obesity, especially the abdominal obesity is increasing continuously which to the higher incidence of type 2 diabetes and cardiovascular diseases. The techniques discussed in this are BMI, WHR, waist circumference, SAD, CT, MRI, DXA, dilution techniques and so on.

A prediction model is proposed [5] for medical data with missing value imputation techniques, then analyzing these techniques by using K-means algorithm and choosing the best among them. Thus this model improves the quality of data by using the best imputation technique. Methods such as case deletion, most common method, concept most common, K-means clustering imputation, k-nearest neighbor etc are applied to fill the missing data values in the data. The efficiency is calculated on three data sets namely Hepatitis, Wisconsin Breast Cancer and Pima Indians Diabetes from the UCI repository. This model achieved accuracy of 99.82% for Diabetes data set, 99.39% for Breast Cancer and 99.08% for Hepatitis data set. For Diabetes and Hepatitis data sets Concept Most Common (CMC) is chosen as the best method, and for Breast Cancer Case deletion is selected as best missing value imputation method.

Algorithms like genetic algorithm, PSO, ANN can be used in predicting heart disease [6]. Combining these algorithms with the data mining techniques such as clustering, classification etc. or by combining these algorithms with one another will give better performance and accuracy.

A study was carried out on the diabetes diagnosis [7] dataset. The experiment was performed using back propagation network trained by LM algorithm. The results of this study were compared with the previous results and the values clearly showed that accuracy was remarkably improved by carrying out this experiment.

In extremely fast growing field of medical, a huge amount data has been generated by this field every day. To handle this data is very difficult, so there is a need of a technology to handle this data [8]. To turn these data into useful pattern, there is a need of a data to be mined. The medical data mining are useful to produce optimum results on prediction based system of medical line. This paper analyzes various disease predictions techniques using K-means algorithm. This data mining based prediction system are reduces the human effects and cost effective one. The characteristics that help to determine the presence of diabetes and to track maximum number of people suffering from diabetes are analyzed [9]. For this purpose they have used data mining technique clustering to find out the characteristic. The data for this study was acquired from National Institute of diabetes, digestive and kidney diseases.

III. Material Andmethods

3.1. K-MEDOID CLUSTERING**:** The K-medoid clustering algorithm is the basic algorithm which is based on partitioning method which is used for many clustering tasks especially with low dimension datasets. K-medoid minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of the cluster.

K-medoid is a variant of k-mean clustering algorithm. In K-medoid algorithm instead of using mean as the center of clusters an actual point in the cluster is used to represent the center.

It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to each other but dissimilar to other objects in other clusters. In K-medoid algorithm data points are used as the centers of the clusters. This algorithm is related to both k-means and medoid shift algorithms. K-medoid is based on the basic concept of partition based clustering that it clusters n objects into k clusters.

The algorithm attempt to find the initial medoid and then each dataset is associated with the closest medoid. The following procedure summarizes the K-medoid algorithm.

1) Select the initial medoid.
2) Associate data set to its closest medoid.
3) Iterate while the configuration cost decreases

a) In every cluster find the point which minimizes the sum of distance within the cluster and the medoid.
b) Now reassign each point to the cluster defined by closest mediod determined in step a.

The most common realization of K-medoid algorithm is PAM (Partitioning Around Medoids) algorithm. In Partitioning Around Medoids algorithm greedy search is employed which may not find the optimum solution but yet it a faster method.

3.2. BACK PROPAGATION ALGORITHM:Back propagation algorithm is a popular form of training neural network. It is abbreviated for backward propagation. It is used in conjunction with an optimized method (e.g. gradient descent). Back Propagation algorithm helps to minimize the errors. The advantages of back propagation algorithm are accuracy and versatility. Back Propagation is an important tool for prediction in data mining and machine learning. It is a common method in which initial system output is compared to the desired output and the systems is adjusted until difference between the two is minimized.

3.3 DATASETS**:** Data sets are organized into some type of data structure. In a database for instance, a data set can contain a collection of business data with attributes like names, salaries, contact, information, and so on.
In this study three diabetes dataset has been used that comes from the UCI repository [12].

IV. Proposed Methodology

The prediction based algorithm is to analyze the clusters using k-medoid algorithm. In k-medoid algorithm dataset will be taken as input and the relationship between various attributes of the dataset will be examined. On the basis of derived relationship the central point will be calculated. The final clusters will be obtained using the Euclidian distance from the central point. These clusters can be analyzed according to the requirement. Medical datasets are extremely complex so finding the relationship among the attributes can be a difficult job. In this work we will use back propagation algorithm to derive relationship between various attributes which will make the final clusters easy to analyze. At the end we will compare the results of proposed and existing techniques to show the improvement.

INPUT: Dataset

OUTPUT: Clustered Data

1. Read dataset and dataset has number of rows “r” and number of columns “m”
2. Apply Back Propagation to identify the relationship among various attributes.
3. Then select the medoid point from the given dataset

a) Calculate the Euclidian distance from the medoid to the rest of the objects.
b) The objects are associated to their closest medoids.

4. Repeat step 3 while the cost is decreased.
5. The steps from 3 and 4 are repeated until all points get clustered.
6. The clusters are analyzed on the basis of various performance measures and the results are compared with traditional k-medoid algorithm.

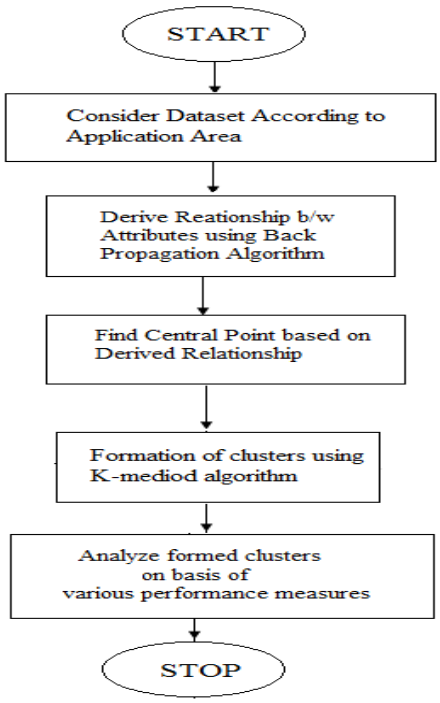

Figure.4.1 Flow Chart of Methodology

V. Experimental Results

The proposed idea is implemented in MATLAB which is widely used in all areas of research universities, and also in the industry. The data this paper studies comes from UCI repository. Here the experiment is run using three different datasets for diabetes and the corresponding results are noted.The results are compared on the basis of four parameters Accuracy, Precision, Recall and F- Measure.

Table 5.1: Experimental Results for Accuracy

	K-Medoid	Improved K-Medoid
Dataset 1	79.80	93.71
Dataset 2	76.31	90.71
Dataset 3	75.94	89.76

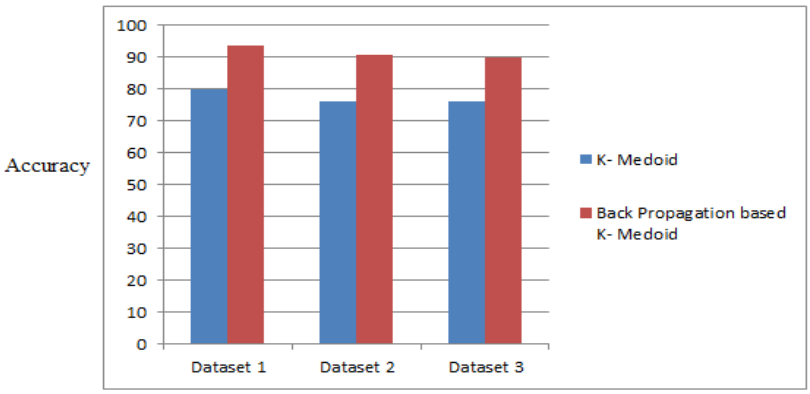

Figure 5.1: Accuracy Results

Table 5.2: Experimental Results for Precision

	K-Medoid	Improved K-Medoid
Dataset 1	39.8	50.5
Dataset 2	39.6	52.4
Dataset 3	39.9	50.8

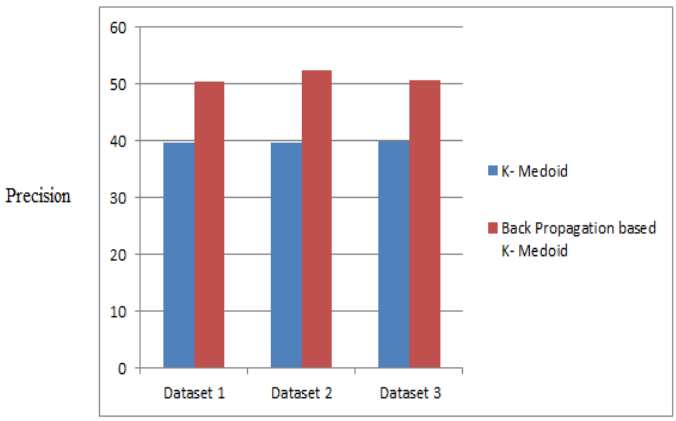

Figure 5.2: Precision Results

Table 5.3: Experimental Results for Recall

	K-Medoid	Improved K-Medoid
Dataset 1	0.167	0.214
Dataset 2	0.165	0.217
Dataset 3	0.161	0.215

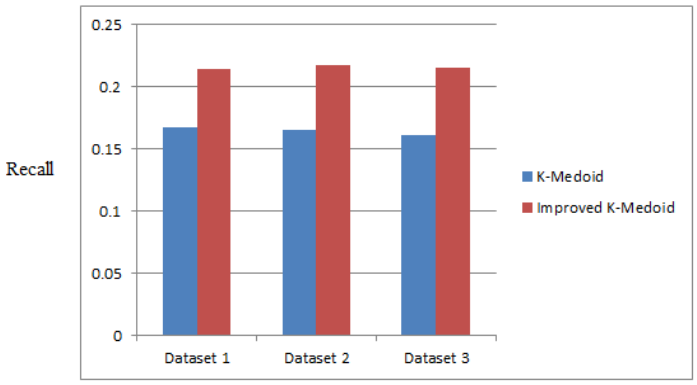

Figure 5.3: Recall Results

Table 5.4: Experimental Results for F-measure

	K-Medoid	Improved K-Medoid
Dataset 1	0.332	0.391
Dataset 2	0.334	0.357
Dataset 3	0.328	0.365

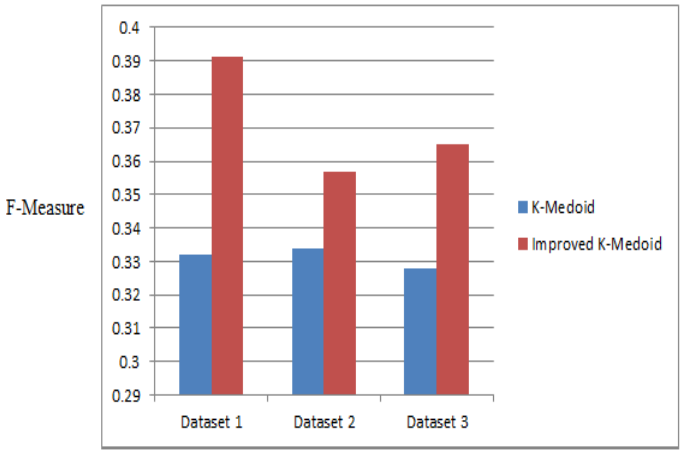

Figure 5.3: F- Measure Results

From above Tables and Figures we can see that accuracy, precision, recall and F-measure clusters have been enhanced with the proposed methodology.

VI. Conclusion

This paper first described partition clustering algorithm K-medoid and analyzed its characteristics. It then proposed method to improve the algorithm using back propagation method and carried out experiment verification by implementing the proposed algorithm to diabetes data. It demonstrated that the performance measuring parameters including accuracy, precision, recall and F- Measure rise up considerably.

Acknowledgement

The author would like to thank the department of computer science and engineering for their encouragement and support towards this research work.

References

[1]. E. Turcato, O. Bosello, V. Di Francesco, T. B. Harris, E. Zoico, L. Bissoli, E. Fracassi, and M. Zamboni, "Waist circumference and abdominal sagittal diameter as surrogates of body fat distribution in the elderly: their relation with cardiovascular risk factors," Int. J. Obes. Relat. Metab. Disord., vol. 24, no. 8, pp. 1005–1010, Aug. 2000.

[2]. Aastha Joshi, Rajneet kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, March 2013.

[3]. S.Wild,G.Roglic, A. Green,R. Sicree, and H.King, "Global prevalence of diabetes: estimates for the year 2000 and projections for 2030," Diabetes Care, vol. 27, no. 5, pp. 1047–1053, May 2004.

[4]. M. B. Snijder, R. M. Van Dam, M. Visser, and J. C. Seidell, "What aspects of body fat are particularly hazardous and how do we measure them?" Int. J. Epidemiol., vol. 35, no. 1, pp. 83–92, Feb. 2006.

[5]. Purwar, A., & Singh, S. K., " Hybrid prediction model with missing value imputation for medical data. *Expert Systems with Applications*, *42*(13), 5621-5631, 2015
[6]. Kumari, V. A., & Chitra, R., " Classification of diabetes disease using support vector machine", International Journal of Engineering Research and Applications, 3(2), 1797-1801.
[7]. M.Durairaj, G.Kalaiselvi," Prediction of Diabetes using Back Propagation Algorithm", International Journal of Emerging Technology and Innovative Engineering, Vol. 1, August 2015
[8]. K.Rajalakshmi, Dr.S.S.Dhenakaran, N.Roobini," Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, July 2015
[9]. P.Padmaja, Srikanth Vikkurty, Nilofer Inaz Siddiqui, Praveen Dasari. Bikkina Ambica, V.B.V.E.Venkata Rao, Mastan Vali Shaik, V.J.P. Raju Rudraraju," Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network S 244 ecurity, VOL.8 No.11, Nov2013
[10]. Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in *K*-Means Clustering and Application in Colour Image Segmentation", School of Computer Science and Software EngineeringMonash University, Wellington Road, Clayton, Victoria, 3168, Australia, 2015. https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Shubhangi Pahwa Back Propagation Based K-Medoid Algorithm for Predicting Type-2 Diabetes." IOSR Journal of Computer Engineering (IOSR-JCE) , vol. 19, no. 6, 2017, pp. 13-19