ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.8

Analysis of Diabetic Patient's Data Using Data Mining Clustering Techniques

G.Samundeeswari[1], G.Silambarasan[2]
M.phil Scholar[1], Assistant professor[2]
Department of Computer Science
Annai Vailankanni of arts & Science College, Tamil Nadu, India

Abstract:
Data mining is Data mining is one of the knowledge discovery steps in database, in which modeling techniques are applied. In this research work, the analysis of K-Means method is applied for dealing with diabetic database for clustering. To increase the efficiency of mining process, some pre-processing needs to be done to the data. Diabetes data mining methods are used to analyze the diabetic data information resources. Diabetic data mining content mining and structure methods are used to analyze the medical data contents. The effort to develop knowledge and experience of frequent specialists and clinical selection data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Diagnose the Diabetes disease is a significant and tedious task in medicine. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of test should be reduced. This reduced test plays an important role in time and performance. This technique has an advantages and disadvantages. This research work analyzes and study about how data mining technique is used for predicting the diabetes disease. This work reviewed the research papers which mainly concentrated on predicting Diabetes. Experimental results showed the good accuracy when applied to the adjust data.

Keywords: Clustering, Diabetes Dataset, K-Means, Performance Measures

I. INTRODUCTION

Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of healthcare information. Knowledge Discovery has the preprocessing, Data mining and Post processing phases. KDD is the iterative or cyclic process that involves sequence of steps of processes and data mining is the core component of the KDD process. Clustering is an important area of application for a variety of fields including data mining, knowledge discovery, statistical data analysis, data compression and vector quantization. Clustering has been formulated in various ways in machine learning, pattern recognition, optimization and statistics literature. Clustering is the most common form of unsupervised learning. The *k*-means problem is to find cluster centers that minimize the intra-class variance, i.e. the sum of squared distances from each data point being clustered to its cluster center (the center that is closest to it). Although finding an exact solution to the *k*-means problem for arbitrary input is NP-hard the standard approach to finding an approximate solution (often called Lloyd's algorithm or the *k*-means algorithm) is used widely and frequently finds reasonable solutions quickly. However, the *k*-means algorithm has at least two major theoretic shortcomings:

- First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size.
- Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering. Diabetes mellitus (DM) or simply diabetes is a group of metabolic diseases in which a person has high blood sugar. This high blood sugar produces the symptoms of frequent urination, increased thirst, and increased hunger. Untreated, diabetes can cause many complications. Acute Complications include diabeticetoacidosis and nonketotic hyperosmolar coma. Seriouslong-term complications include heart disease, kidney failure, and damage to the eyes.

II. LITERATURE REVIEW

Subhagata chattopadhyay, Dilip kumar pratihar, Sanjib chandra de sarkar, A cluster [1] is usually represented as either grouping of similar data points around a center (called centroid) or a prototype data instance nearest to the centroid. In other way, a cluster can be represented either with or without a well-defined boundary. Clusters with well-defined boundaries are called crisp clusters, while those without such feature are called k means clusters. The present work deals with k means clustering only. Bottou, L. and Bengio, Y. The k-means algorithm [2] is by far the most popular clustering tool used in scientific and industrial applications. The name comes from representing each of k clusters Cj by the mean (or weighted average) c of its points, the so-called centroid. While this obviously does not work well with categorical attributes, it has the good geometric and statistical sense for numerical attributes. Rahul Malik , Raj Kumar [3] From all the above calculations we come to the conclusion that the K-Mean algorithm is an excellent algorithm when we are dealing with a small or medium sized data. It simply provides good performance vector every time. A direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets. The main disadvantage of the k-means algorithm is that the number of clusters, K, must be supplied as a parameter. In this

work we present a simple validity measure based on the intra-cluster and inter-cluster distance measures which allows the number of clusters to be determined automatically. Bahman Bahmani, Benjamin Moseley, Andrea Vattani, k-means [4] initialization algorithm achieves this, obtaining an initial set of centers that is provably close to the optimum solution. A major downside of the k-means is its inherent sequential nature, which limits its applicability to massive data: one must make k passes over the data to _nd a good initial set of centers. In this work we show how to drastically reduce the number of passes needed to obtain, in parallel, a good initialization. R.N. Dave, K. Bhaswan, clustering [5] has been widely studied and applied in a variety of key areas and k means cluster validation plays a very important role in k means clustering. K means*C*-Means clustering algorithm on a number of widely used data sets, and make a simple analysis of the experimental results.

III. PROBLEM DESCRIPTION

PIMA Indian diabetes dataset are from UCI repositories which consist of 338 dataset. PIMA Indian dataset are provide diabetes database which used for easily diagnosis diabetes.
In training all below function are performed.
a)Read all input data set.
b)All activation function and derivatives selection are performed by system.
c) Particular algorithm performed such as back propagation, feed forward neural network, naive bayes, svm etc…
d) Create such type of function that generates network error.
e) Implement the train function

A. Testing Data:

Testing system means to check according to symptoms system Diagnosis of disease properly or not. User gives query that is testing instance is created. And that test instances gives to testing module. Considering symptoms testing module diagnosis of diabetes. They diagnosis of diabetes in Yes or No format. If diabetes occurs then give yes otherwise no.

B. Cluster Analysis

The objective of cluster analysis is the classification of objects according to similarities among them, and organizing of data into groups. Clustering techniques are among the unsupervised methods, they do not use prior class identifiers. The main potential of clustering is to detect the underlying structure in data, not only for classification and pattern recognition, but for model reduction and optimization. Different classifications can be related to the algorithmic approach of the clustering techniques. Partitioning, hierarchical, graph-theoretic methods and methods based on objective function can be distinguished. In this work we have worked out a toolbox for the partitioning methods, especially for hard and fuzzy partition methods.

IV. METHODOLOGY

A. Data Mining

a. Overview

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

b. Continuous Innovation

Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years..

c. Data, Information, and Knowledge

Data mining consists of five major elements

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available

- ***Artificial neural networks:*** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- ***Genetic algorithms*****:** Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- ***Decision trees*****:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome.
- ***Nearest neighbor method:*** A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- ***Rule induction:*** The extraction of useful if-then rules from data based on statistical significance.
- ***Data visualization:*** The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

B. K-means Algorithm

The *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.

Algorithm

Given the data set X, choose the number of clusters $1 < c < N$.
Initialize with random cluster centers chosen from the data set.
Repeat for $l = 1; 2; : :$
Step 1 Compute the distances

$$D_{ik}^{2} = \left(x_{k} - v_{i}\right)^{T} \left(x_{k} - v_{i}\right), \qquad 1 \le i \le c, \quad 1 \le k \le N.$$

Step 2 Select the points for a cluster with the minimal distances, they belong to that cluster.
Step 3 Calculate cluster centers

$$v_i^{(l)} = \frac{\sum_{j=1}^{N_i} x_i}{N_i}$$

Until

$$\prod_{k=1}^{n} max \left| v^{(l)} - v^{(l-1)} \right| \neq 0$$

Ending Calculate the partition matrix

K means Algorithm Steps

Input:
D = {d1, d2,.....,dn}
D is n data items set.
k = No. of desired clusters
Output:
A set of k clusters.
Steps:
1. Select k data-items from set D as initial centroid;
2. Repeat Assign each item di to the cluster which has the closest centroid;
Calculate new mean for each cluster;
Until convergence criteria is met

C. Euclidean Distance

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric The Euclidean distance, data vector p and centroid q is computed as

$$d(p,q) = \sqrt{\sum_{k=1}^{n} (q_{ik} - p_{ik})^2}$$

D. Cluster Validity Measure

Cluster validity refers to the problem whether a given k means partition fits to the data all. The clustering algorithm always tries to find the best fit for a fixed number of clusters and the parameterized cluster shapes. However this does not mean that even the best ¯t is meaningful at all.

- Starting with a sufficiently large number of clusters, and successively reducing this number by merging clusters that are similar (compatible) with respect to some predefined criteria. This approach is called *compatible cluster merging*.
- Clustering data for different values of *c*, and using *validity measures* to assess the goodness of the obtained partitions

1. Partition Coefficient (PC): measures the amount of "overlapping" between clusters.

$$PC(c) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^2$$

Where μ_{ij} is the membership of data point *j* in cluster *i*. The disadvantage of PC is lack of direct connection to some property of the data themselves. The optimal number of cluster is at the maximum value.

2. Classification Entropy (CE): it measures the k means of the cluster partition only, which is similar to the Partition Coefficient.

$$CE(c) = -\frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij} log(\mu_{ij}),$$

3. Partition Index (PI): is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the k means cardinality of each cluster.

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{j=1}^{N} (\mu_{ij})^m ||x_j - v_i||^2}{N_i \sum_{k=1}^{c} ||v_k - v_i||^2}$$

SC is useful when comparing different partitions having equal number of clusters. A lower value of *SC* indicates a better partition.

4. Separation Index (SI): on the contrary of partition index (SC), the separation index uses a minimum-distance separation for partition validity.

$$S(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^2 ||x_j - v_i||^2}{N min_{i,k} ||v_k - v_i||^2}$$

5. Xie and Beni's Index (XB): it aims to quantify the ratio of the total variation within clusters and the separation of clusters.

$$XB(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^m ||x_j - v_i||^2}{N min_{i,j} ||x_j - v_i||^2}$$

The optimal number of clusters should minimize the value of the index.

6. Dunn's Index (DI): this index is originally proposed to use at the identification of "compact and well separated clusters". So the result of the clustering has to be recalculated as it was a hard partition algorithm

$$DI(c) = min_{i \in c} \{ min_{j \in c, i \neq j} \{ \frac{min_{x \in C_i, y \in C_j} d(x,y)}{max_{k \in c} \{ max_{x,y \in C} d(x,y) \}} \} \}$$

The main drawback of Dunn's index is computational since calculating becomes computationally very expansive as *c* and *N* increase.

7. Alternative Dunn Index (ADI): the aim of modifying the original Dunn's index was that the calculation becomes more simple, when the dissimilarity function between two clusters ($min_{x \epsilon Ci, y \epsilon Cj}$ *d*(*x*, *y*)) is rated in value from beneath by the triangle-non equality:

$$d(x,y) \geq |d(y,v_j) - d(x,v_j)|$$

where *vj* is the cluster center of the *j*-th cluster.

$$ADI(c) = min_{i \in c} \{ min_{j \in c, i \neq j} \{ \frac{min_{x_i \in C_i, x_j \in C_j} |d(y,v_j) - d(x_i,v_j)|}{max_{k \in c} \{ max_{x,y \in C} d(x,y) \}} \} \}$$

Partitional Clustering

Non-hierarchical, each instance is placed in exactly one of K non-overlapping clusters, Since only one set of clusters is output, the user normally has to input the desired number of clusters K.

D. K Means Clustering

K-means clustering is the simplest and the most commonly used partitioning method for splitting a dataset into a set

of kgroups (i.e. clusters). It requires the analyst to specify the number of optimal clusters to be generated from the data.
The algorithm is composed of the following steps:
1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

V. EXPERIMENTAL RESULTS

A. Pima Indian Diabetes Dataset

World Health Organization (WHO) report had shown a marked increase in the number of diabetics and this trend is expected to grow in the next couple of decades. People with type 1 diabetes must take daily insulin injections to survive. This form of diabetes usually develops in children or young adults, but can occur at any age. Type 2 (also called adult-onset or non insulin-dependent) diabetes results when the body doesn't produce enough insulin and/or is unable to use insulin properly (insulin resistance).

This form of diabetes usually occurs in people who are over 40, overweight, and have a family history of diabetes, although today it is increasingly occurring in younger people, particularly adolescents. Type II Diabetes (not depending on insulin) is the most common form of diabetes (90 to 95 per cent) and occurs primarily in adults but is now also affecting children and young adults. Type I Diabetes (insulin-dependent) affects predominately children and youth, and is the less common form of diabetes (5 to 10 percent). The major risk factors for diabetes include obesity, high cholesterol, high blood pressure and physical inactivity. The Pima Indian diabetes data set is taken from the UCI machine learning repository [18].The data set has 768 instances with two class problems to test whether the patient is positive or negative for diabetes. The patients in this dataset are Pima Indian Women who lives near Phoenix Arizona, USA. This data set consists of 9 attributes as shown in Table1

Table.1. Pima Indian Dataset

No.	Attribute	Description	Missing Values
1	pregnant	Number of times pregnant	110
2	glucose	Plasma glucose concentration (glucose tolerance test)	5
3	pressure	Diastolic blood pressure (mm Hg)	35
4	triceps	Triceps skin fold thickness (mm)	227
5	insulin	2-Hour serum insulin (mu U/ml)	374
6	mass	Body mass index (weight in kg/(height in m)^2)	11
7	pedigree	Diabetes pedigree function	0
8	age	Age (years)	0
9	diabetes	Class variable (test for diabetes)	0

Class Distribution: Class value 1 is interpreted as "tested positive for diabetes"
Class Value: 0 - Number of instances - 500
Class Value: 1 - Number of instances – 268

Table.2. Description of Dataset

Dataset	Number of Objects	Number of Attributes	Number of Clusters
Pima Indian Diabetes	768	8	2

B. Validity Measure

Table.3. Clustering Validity Measure for Diabetes Dataset and Algorithm

Dataset	Algorithm	PC	CE	SC	S	XB	DI	ADI
Diabetes	**K-Means**	1	NaN	0.72655	0.00144	3.17784	0.64816	0.5636

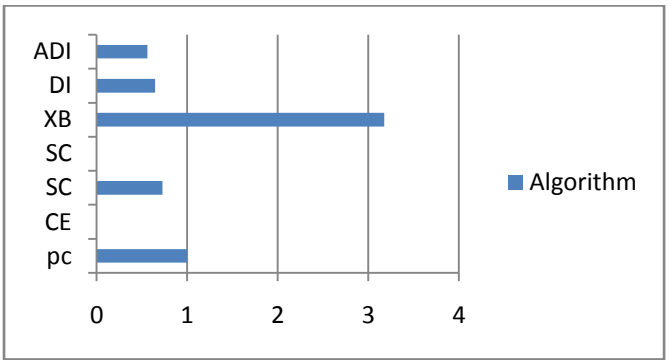

Figure.1. Cluster Validity Measure

VI. CONCLUSION

Cluster analysis is one of the major tasks in various research areas. The clustering aims at identifying and extract significant groups in underlying data. Thus based on a certain clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters. Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed and are available in literature. In the proposed system to analysis the major clustering algorithm such as K-Means with Euclidean distance measure by using UCI dataset. It illustrates the efficiency of clustering algorithm with its validity measures. It shows the K-Means clustering algorithm had better than other clustering algorithms. The experimental result shows the performance of the K-Means algorithm was improved significantly.

VII. REFERENCE

[1]. Huy Nguyen Anh Pham and Evangelos Triantaphyllou "Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization" Department of Computer Science, 298 Coates Hall, Louisiana State University, Baton Rouge, LA 70803

[2]. Ms.S.Sapna, Dr.A.Tamilarasi "Data mining – K means Neural Genetic Algorithm in predicting diabetes" Department of Computer Applications (MCA), K.S.R College of

Engineering “BOOM 2K8” Research Journal on Computer Engineering, March2008.

[3]. Subhagata chattopadhyay, Dilip kumar pratihar, Sanjib chandra de sarkar , “A comparative study of k meansc-means Algorithm and entropy-based k means Clustering algorithms”.

[4]. BOTTOU, L. and BENGIO, Y. 1995. Convergence properties of the K-means algorithms. In Tesauro, G. and Touretzky, D. (Eds.) Advances in Neural Information Processing Systems 7, 585-592, The MIT Press, Cambridge, MA.

[5]. David Arthur, Sergei Vassilvitskii, “k-means++: The Advantages of Careful Seeding “

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

[7]. Bahman Bahmani, Benjamin Moseley, Andrea Vattani, “Scalable K Means++“

[8]. R.N. Dave, K. Bhaswan, Adaptive k meansc-shells clustering and detection of ellipses, IEEE Trans. Neural Networks 3 (5) (1992) 643–662.

[9]. Asha Gowda Karegowda , M.A. Jayaram, A.S. Manjunath, Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients

[10]. WeinaWang, Yunjie Zhang, “On k means cluster validity indices “