International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Diagnosis of Diabetic Disease using Patient's Data with Data Mining Validity Measures

M. Gayatri [a], D. Swetha [b*]

[a, b] Assistant Professor, Department of Computer Applications, Idhaya College for Women, Sarugani.

A B S T R A C T

Data mining is one of the steps of knowledge discovery in databases where modeling techniques are applied. In this research work, the K-Means analysis method is used to cluster the diabetes database. In order to improve the efficiency of the mining process, it is necessary to preprocess the data. Diabetes data extraction methods are used to analyze diabetes data information resources. Content extraction and structuring methods for diabetes data mining are used to analyze the content of medical data. Efforts to develop the knowledge and experience of frequent specialists and clinical selection of patient data collected in databases to facilitate the diagnostic process are considered a valuable option. Diagnosing diabetes is an important and tedious task in medicine. To detect the disease, a series of tests must be performed on the patient. But when using data mining techniques, the number of tests should be reduced. This reduced testing plays a big role in terms of time and performance. This technique has advantages and disadvantages. This research work analyzes and studies how data mining techniques can be used to predict diabetes diseases. This article reviews research papers that mainly focus on diabetes prediction. Experimental results show good accuracy when applied to fitted data.

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

1. Introduction

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, reduce 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 an environment for developing the tools necessary to handle the massive amounts of data facing organizations that rely on ever-growing databases of healthcare information. Knowledge discovery has three stages: preprocessing, data mining and post processing. KDD is an iterative or cyclic process involving a series of process steps, and data extraction is a core component of the KDD process.

Clustering is a major application area in 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 the literature on machine learning, pattern recognition, optimization, and statistics. Clustering is the most common form of unsupervised learning.

The k-means problem consists of finding the cluster centers that minimize the within-class variance, which is the sum of the squares of the distances from each data point of a cluster to its cluster center (the center closest to it). Although finding an exact solution to the k-means problem for arbitrary inputs is NP-hard, standard methods for finding approximate solutions (often called Lloyd's algorithm or k-means algorithm) are widely used and often find reasonable solutions.

However, the k-means algorithm has at least two major theoretical disadvantages:

- First, it has been shown that the worst-case execution time of the algorithm is hyper polynomial in the size of the input.
- Second, the approximation found may be arbitrarily poor in terms of 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 elevated blood sugar levels. This high blood sugar level produces symptoms of frequent urination, increased thirst and hunger.

Diabetes can cause many complications if left untreated. Complications include diabetic acidosis and no kenotic hyperosmolar coma. Serious long-term complications include heart disease, kidney failure and eye damage.

2. Literature Review

Alpan, Kezban, and Galip Savaş İlgi., A cluster [1] is usually represented as a group of similar data points around a center (called a centroid) or a prototypical data instance closest to a centroid. Otherwise, a group can be represented with or without well-defined boundaries. Clusters with well-defined boundaries are called sharp clusters, while clusters without such characteristics are called k-means clusters. This article deals only with k-means grouping.

Eyasu, Kedir, Worku Jimma, and Takele Tadesse 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 the k clusters Cj by the mean (or weighted mean) c of its points, the so-called centroid. While this obviously doesn't work for categorical attributes, it makes good geometric and statistical sense for numeric attributes.

Raja, J. Beschi, and S. Chenthur Pandian [3] From all the above calculations, we conclude that the K-Mean algorithm is an excellent algorithm when dealing with small to medium data. It delivers a great performance vector every time. The time required for the straightforward algorithm of the k-means method is proportional to the product of the number of modes and the number of groups per iteration. This is computationally expensive, especially for large datasets. The main disadvantage of the k-means algorithm is that the number of groups K must be provided as a parameter. In this paper, we propose a simple effectiveness measure based on within-cluster and between-cluster distance measures that automatically determines the number of clusters.

The k-means initialization algorithm [4] of Rghioui, Amine, Assia Naja, and Abedlmajid Oumnad achieves this, obtaining an initial set of centers that are likely to be close to the optimal solution. A major disadvantage of k-means is its inherent sequential nature, which limits its applicability to large data: It is necessary to run k through the data to find a good set of initial centers. In this work, we show how to substantially reduce the number of passes required to obtain good initialization in parallel.

Ameena, R. Rifat, and B. Ashadevi Clustering [5] has been extensively researched and applied in various key fields, and cluster validation of k-means plays a very important role in clustering k-means. K represents the C-Means clustering algorithm on a series of widely used datasets, and a simple analysis of the experimental results.

3. Problem Description

The PIMA India Diabetes dataset is from the UCI repository consisting of 338 datasets. PIMA Indian Dataset provides a diabetes database for easy diagnosis of diabetes.

In training all below function are performed.

Read all input data set.

a. All activation function and derivatives selection are performed by system.

b. Particular algorithm performed such as back propagation, feed forward neural network, naive bayes, svm etc…

c. Create such type of function that generates network error.

d. Implement the train function

3.1. Testing Data

The test system refers to checking whether it is correct or not according to the disease diagnosis symptom system. The user gives a query to create a test instance. These test cases give the test modules. . Consider the symptoms of the Diabetes Diagnostic Test Module. They make a diagnosis of diabetes in Yes or No format. If diabetes occurred, answer yes, otherwise no.

3.2. Cluster Analysis

The goal of cluster analysis is to classify objects based on their similarity and organize data into groups. Bundling techniques are unsupervised methods, they do not use previous class identifiers. The main potential of clustering is to detect underlying structure in data, not only for classification and pattern recognition, but also for model reduction and optimization. The different classifications may be related to the algorithmic approach of the clustering technique. A distinction can be made between partitioned, hierarchical, graph-theoretic methods and methods based on objective functions. In this work, we develop a toolbox for partitioning methods, especially hard and fuzzy partitioning methods.

4. Methodology

4.1. Data Mining

4.1.1. Overview

In general, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and aggregating it into useful information—information that can be used to increase revenue, reduce costs, or both. Data mining software is one of several analytical tools used to analyze 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.

4.1.2. Continuous Innovation

While data mining is a relatively new term, the technology is not. For years, companies have used powerful computers to sift through reams of data from supermarket scanners and analyze market research reports.

4.1.3. Data, Information, and Knowledge

- Data mining consists of five major elements
- Extract, transform and load transactional data into data warehouse systems.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyse 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 are structurally similar to biological neural networks.
- Genetic Algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in designs based on concepts of natural evolution.
- Decision Tree: A tree structure representing a decision set. These decisions generate rules for classifying datasets. Specific decision tree methods include classification and regression trees (CART) and chi-squared automated interaction detection (CHAID). CART and CHAID are decision tree techniques for classifying datasets. 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 the class combination of the k records most similar to the record in the historical dataset (where k = 1). It is sometimes called the k-nearest neighbor technique.
- Rule induction: extract useful if-then rules from data based on statistical significance
- Data Visualization: Visual interpretation of complex relationships in multidimensional data. Graphical tools are used to illustrate data relationships.

4.2. K-means Algorithm

The k-means algorithm takes the input parameter k as a parameter, and divides a set of n objects into k groups, so that the similarity within the group is high and the similarity between groups 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

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

$1 \le k \le N$.

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

4.3. Euclidean Distance

In mathematics, a Euclidean distance or Euclidean metric is the "ordinary" distance between two points measured with a ruler, given by the Pythagorean formula. By using this formula for distances, Euclidean space (or even any becomes a metric space. The associated norm is called the Euclidean norm. The oldest literature refers to the metric as a Pythagorean metric.inner product space) becomes a metric space. The associated norm is called the Euclidean norm. The oldest literature refers to the metric as a Pythagorean metric.

The Euclidean distance, data vector p and centroid q is computed as

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

4.4. Cluster Validity Measure

Cluster validity refers to the question of whether a given k means that the partition fits all the data. Clustering algorithms always try to find the best fit for a fixed number of parameterized clusters and cluster shapes. However, that doesn't mean that even the best is unimportant.

- 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 μij is the degree of membership of data point j in group i. The downside of the PC is the lack of a direct connection to any property of the data itself. The optimal number of clusters is 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 with the same number of clusters. Lower SC values indicate better partitioning.

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 disadvantage of Dunn's index is computational, as it becomes very computationally expensive as c and N increase.

7. Alternative Dunn Index (ADI): The purpose of modifying the original Dunn index is to make the calculation easier, when the dissimilarity function (minxϵCi,yϵCj d(x, y)) between the two groups is classified as value-inequality from below through the triangle :

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

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

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

5. Experimental Results and Discussion

5.1. Pima Indian Diabetes Dataset

The World Health Organization (WHO) reports that the number of people with diabetes has increased significantly, and this trend is expected to increase over the next two decades.

People with type 1 diabetes must take daily insulin injections to survive. This form of diabetes usually occurs in children or young adults, but it can occur at any age. Type 2 diabetes (also called adult-onset or non-insulin-dependent) occurs when the body doesn't produce enough insulin and/or doesn't use it properly (insulin resistance). This form of diabetes usually occurs in people over the age of 40 who are overweight and have a family history of diabetes, although it is now increasingly occurring in younger people, especially teenagers. Type 2 diabetes (non-insulin-dependent) is the most common form of diabetes (90% to 95%), mainly occurring in adults but now also affecting children and young adults. Type I (insulin-dependent) diabetes mainly affects children and young adults and is the least common form of diabetes (5% to 10%). Major risk factors for diabetes include obesity, high cholesterol, high blood pressure and lack of physical activity.

The Pima Indian Diabetes dataset is taken from the UCI Machine Learning Repository [18]. The dataset has 768 instances with two types of questions testing whether a patient has diabetes. The patients in this dataset were Pima Indian women living near Phoenix, Arizona, USA. The dataset contains 9 attributes, as shown in Table 1

Table 1- Pima Indian Dataset

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

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

5.2.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	0.7265	0.0014	3.1784	0.6481	0.5636

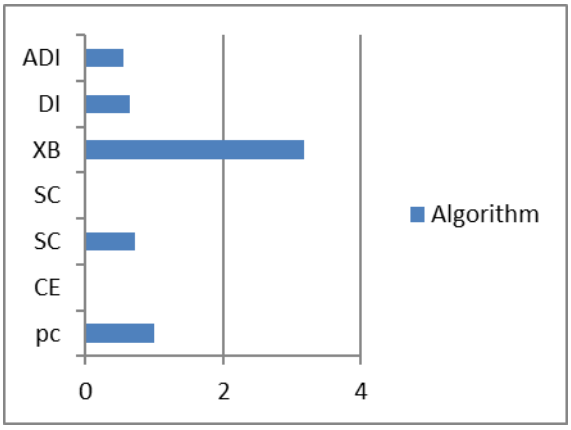

Fig 1- Cluster Validity Measure

6. Conclusion

Cluster analysis is one of the main tasks in several research fields. Clustering aims to identify and extract important groups in the underlying data. Therefore, based on certain grouping criteria, data is grouped such that data points in one group are more similar to each other than points in different groups. Since clustering is applied in many fields, various clustering techniques and algorithms have been proposed and available in the literature. In the proposed system, major clustering algorithms such as K-Means with Euclidean distance metric are analyzed by using the UCI dataset. It illustrates the efficiency of clustering algorithms and their effectiveness measures. It shows that K-Means clustering algorithm is superior to other clustering algorithms. Experimental results show that the performance of K-Means algorithm has been significantly improved.

References

Alpan, Kezban, and Galip Savaş İlgi. "Classification of diabetes dataset with data mining techniques by using WEKA approach." *2020 4th International Symposium on Multidisciplinary Studies and* Innovative *Technologies (ISMSIT)*. IEEE, 2020.

Eyasu, Kedir, Worku Jimma, and Takele Tadesse. "Developing a prototype knowledge-based system for diagnosis and treatment of diabetes using data mining techniques." Ethiopian journal *of* health sciences 30.1 (2020).

Raja, J. Beschi, and S. Chenthur Pandian. "PSO-FCM based data mining model to predict diabetic disease." *Computer Methods and Programs in Biomedicine* 196 (2020): 105659.

Rghioui, Amine, Assia Naja, and Abedlmajid Oumnad. "Diabetic patients monitoring and data classification using IoT application." *2020 International Conference on Electrical and* Information Technologies *(ICEIT)*. IEEE, 2020.

Ameena, R. Rifat, and B. Ashadevi. "Predictive analysis of diabetic women patients using R." *Systems Simulation and Modeling for Cloud Computing and Big Data Applications*. Academic Press, 2020. 99-113.

Singh, L., P. Singh, and Naveen Dhillon. "Diagnosis and prediction of diabetes patient data by using data mining techniques." *Int. Res. J. Eng. Technol* 7.10 (2020): 1564-1568.

Sharma, Gajendra, and Umesh Hengaju. "Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes." *International Journal of Advanced Networking and Applications* 12.1 (2020): 4509-4518.

Jayasri, N. P., and R. Aruna. "Big data analytics in health care by data mining and classification techniques." *ICT Express* 8.2 (2022): 250-257.

Abdulqadir, Hindreen Rashid, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari. "Data mining classification techniques for diabetes prediction." *Qubahan Academic Journal* 1.2 (2021): 125-133.

Birjandi, Saba Maleki, and Seyed Hossein Khasteh. "A survey on data mining techniques used in medicine." *Journal of Diabetes & Metabolic Disorders* 20.2 (2021): 2055-2071.

Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." *ICT Express* 7.4 (2021): 432-439.

Liu, Yang, Zhaoxiang Yu, and Yunlong Yang. "Diabetes risk data mining method based on electronic medical record analysis." *Journal of Healthcare Engineering* 2021 (2021).

Li, Xiaohua, Jusheng Zhang, and Fatemeh Safara. "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm." *Neural Processing Letters* (2021): 1-17.

Khorshid, Shler Farhad, Adnan Mohsin Abdulazeez, and Amira Bibo Sallow. "A comparative analysis and predicting for breast cancer detection based on data mining models." *Asian Journal of Research in Computer Science* 8.4 (2021): 45-59.

Patel, Sina, Vijayshri Khedkar, and Sonali Kothari Tidke. "Analyses of Feature Selection and Classification Techniques for Diabetes Prediction." *ICT Analysis and Applications*. Springer, Singapore, 2022. 427-435.

Jader, Rasool, and Sadegh Aminifar. "Fast and accurate artificial neural network model for diabetes recognition." *NeuroQuantology* 20.10 (2022): 2187-2196.

Jayasri, N. P., and R. Aruna. "Big data analytics in health care by data mining and classification techniques." *ICT Express* 8.2 (2022): 250-257.

Pati, Abhilash, Manoranjan Parhi, and Binod Kumar Pattanayak. "IADP: An Integrated Approach for Diabetes Prediction Using Classification Techniques." *Advances in Distributed Computing and Machine Learning*. Springer, Singapore, 2022. 287-298.

Chakraborty, Sanjay, S. K. Islam, and Debabrata Samanta. "Real-Time Application with Data Mining and Machine Learning." *Data Classification and Incremental Clustering in Data Mining and Machine Learning*. Springer, Cham, 2022. 129-157.

Shrikhande, Santosh P., and Prashant P. Agnihotri. "Comparative Study of Various Data Mining Techniques for Early Prediction of Diabetes Disease." (2022).