Heart Disease Prediction for Data Mining Techniques: A Review

Prerna[1], Er. Ram Singh[2]

[1] *Research Scholar of Computer Science Engineering, Punjabi university Patiala (Punjab)*

[2] *Assistant Professor of Computer Science Engineering, Punjabi university Patiala (Punjab)*

ABSTRACT- From the large historical cardiovascular patient databases, Knowledge Discovery by data mining approaches (KDD) constitutes a process that allows data sets to be modeled and analyzed. To find, evaluate and address the pros and cons of the existing methods in existing literature on the issue, a comprehensive study is produced in this paper. A classification and clustering formulation is incorporated for clinical subjects based on the comparison of parameters associated with cardiovascular risk factors by means of KDD-based algorithms. The data mining is the technology which can mine the useful information from the rough data. The prediction analysis is the technique of data mining which can predict the future situations from the current data. The prediction analysis is the combination of the clustering and classification. The attributes of the dataset have the complex which is used for the heart disease prediction. Heart disease prediction techniques are based upon the methods of classification. In this research work, neural network technique will be applied which can drive relationship between the attributes for the prediction analysis.

KEYWORDS: *Data mining, Classification, Clustering, K-means, SVM*

I. INTRODUCTION

The heart is vital part or an organ of the body. Life is subject to the proficient working of the heart. In the event that operation of heart is not proper, it will influence the other body parts of human, for example, mind, kidney, etc. the Cardiovascular / Heart Disease is a class of disease that affects the heart and blood vessel many of which are related to a process is called atherosclerosis There are the number of elements which build the danger of Heart infection:

- the family history of coronary illness
- smoking
- Poor eating methodology
- high pulse
- cholesterol
- high blood cholesterol
- obesity
- Physical inertia
- Drug abuse
- Stress
- Age (55 or older of women)

For cardiovascular or Heart disease diagnosis, healthcare practitioners face a lot of challenges to detect and predict the correct cause of disease and to provide cost-effective and high-quality patient care. This is critical in particular to enhance the effectiveness of disease treatment and preventions. It becomes more important in case of heart disease that is regarded as the primary reason behind death in adults. Data mining serves as an analysis tool to discover hidden relationships and patterns in medical data. By far, the observations reveal that Neural Networks performed well as compared with the state-of-the-art methods.

A. Data Mining

The process of extraction of interesting knowledge and patterns to analyze data is known as data mining. In data mining, there are various data mining tools available which are used to analyze different types of data. Decision making, market basket analysis, production control, customer retention, scientific discovery and education systems are some of the applications that use data mining in order to analyze the collected information [1]. The customer categorized group and purchasing patterns done by clustering can be used by the marketer to discover their customer's interest. In a city, similar houses and lands area can be identified by employing clustering in geology. To discover new theories, information clustering can be used that classify all documents available on Web. The unsupervised data clustering classification method create clusters, the group of objects in such a way that objects in different clusters are distinct and that are in the same cluster are very similar to each other. In data mining, cluster analysis is considered as

one of the traditional topics that are the first step in the discovery of knowledge.

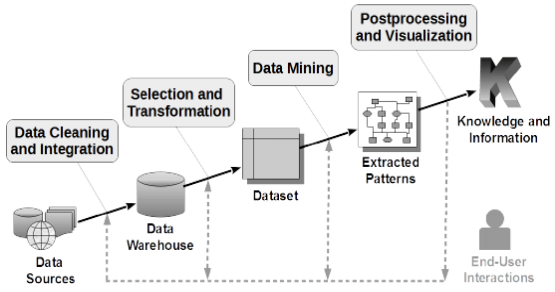

Fig.1: Data Mining Process

II. DATA MINING TECHNIQUES

To analyze large amount of data, data mining came into picture and is also called as KDD process. To complete this process various techniques developed so far are explained in this section,

A. Clustering in Data Mining:

The data objects are grouped into a set of disjoint classes which is known as clusters [2]. Now, objects within a class have a high resemblance to each other and in the meantime objects in separate classes are more unlike. Following are some broader categories into which the clustering methods have been categorized:

a. Partitioning Methods**:** The gathering of samples that are of high similarity in order to generate clusters of similar objects is the basic functioning of this method. Here, the samples that are dissimilar are grouped under different clusters from similar ones. These methods completely rely on the distance of the samples [3].

b. Hierarchical Methods: A given dataset of objects are decomposed hierarchically within this technique. There are two types in which this method is classified on the basis of the type of decomposition involved. They are agglomerative and divisive based methods [4]. A bottom-up technique in which the formation of the separate group is the first step performed is known as agglomerative technique. Further, the groups that are near to each other are merged together.

c. Density-Based Methods: The distance amongst the objects is taken as a base in order to separate the objects into clusters in most of the technique. However, these methods can only be helpful while identifying the spherically shaped clusters. It is difficult to obtain arbitrarily shaped clusters within these techniques.

d. Grid-Based Methods: A grid structure is generated by quantizing the object space into the finite number of cells which is known as the grid-based method. This method has high speed and does not depend on the number of data objects available.

B. Classification in Data Mining

The group membership for data instances can be predicted with the help classification technique within the data mining. In order to predict the data for example classification can be utilized by the applications on a specific day to identify the weather which can be either “sunny”, “rainy” or “cloudy”. Two steps are followed within this process. They are [5]:

a**.** Model Construction**:** Model construction describes the set of predetermined classes. The class label attribute determined each tuple/sample which is assumed to belong to a predefined class.

b**.** Model usage: the second step used in the classification is model usage. In order to classify the future and unknown objects, model usage is widely used as this model estimates the accuracy of the model. The classified result from the model is used to compare with the known label of the test sample. The test set is not dependent on a training set.

In this prediction process, the classification techniques utilized are,

- Artificial neural network
- k-nearest neighbor algorithm (KNN)
- Naïve- Bayes
- Decision Tree
- Support vector machine

C. Artificial Neural network

An ANN network of many simple process or unit that are connected by communication channel that carry numeric data. It is a network of simple unit that can be real- value input and output. In Neural network prediction accuracy is generally high.

D. K-nearest neighbor algorithm (KNN)

The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. KNN can be used for classification- the output is a class membership (predicts a class discrete value). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K-

nearest neighbors measured by a distance function. It can be used a Euclidean distance.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

E. Naïve- Bayes

It is one of the popular classification techniques of algorithms used in data mining. It is a probability classifier. It links the attributes mutually & is dependent on the number of parameters.

F. Decision Tree

In this type of classification the knowledge is represented in a tree diagram. Schematic representation will be in the form of tree to depict the decisions. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too.

G. SVM classifier

SVM stands for support vector machine. It is a binary classifier that maximizes the margin. The best hyperplane which separates all the data points of an individual class can be identified through the classification provided by SVM. The largest margin between the two classes describes the best hyperplane for an SVM [6]. The maximum width between the slabs parallel to the hyperplane is known as margin which has no interior data points. The SVM algorithm is used to separate maximum margin in a hyperplane. The margin planes determined using the point from each class are called support vectors (SVs). It has many applications such as bioinformatics, text, image recognition etc. It becomes popular due to its success in handwritten digit recognition. A huge and varied community works on them like machine learning, optimization, statistics, neural networks, functional analysis, etc.

III. LITERATURE REVIEW

Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang [7] proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. The data was gathered from a hospital which included within it both structured as well as unstructured types of data. In order to make predictions related to the chronic disease that had been spread within several regions, various machine learning algorithms were streamlined here. 94.8% of prediction accuracy was achieved here along with the higher convergence speed in comparison to other similar enhanced algorithms.

Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal [8] presented, that different analytic tool has been used to extract information from large datasets such as in medical field where a huge data is available. The proposed algorithm has been tested by performing different experiments on it that gives excellent result on real data sets. In real-world problem enhanced results are achieved using a proposed algorithm as compared to existing simple k-means clustering algorithm.

Sanjay Chakraborty, Prof. N.K Nigwani and Lop Dey [9], that powerful tool clustering is used as different forecasting tools. The weather forecasting has been performed using proposed incremental K-mean clustering generic methodology. The weather events forecasting and prediction becomes easy using modeled computations. In the last, the authors have performed different experiments to check the proposed approach correctness.

Chew Li Sa, Bt Abang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, M [10] particular university student results have been recorded to keep a track using Student Performance Analysis System (SPAS). The design and analysis have been performed to predict student's performance using proposed project on their results data. The data mining technique generated rules that are used by the proposed system to gives enhanced results in predicting student performance. The student's grades are used to the classy existing student using classification by data mining technique.

Qasem A. Al-Radaideh, Adel Abu Assaf and Eman Alnagi [11] that data analysis prediction is considered as import subject for forecasting stock return. The data analysis future can be predicted through past investigation. The past historical knowledge of experiments has been used by stock market investors to predict better timing to buy or sell stocks. There are different available data mining techniques out of all a decision tree classifier has been used by authors in this work.

K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin [12] presented a study related to medical fast-growing field authors. In this field, every single day a large amount of data has been generated and to handle this much of a large amount of data is not an easy task. The medical line prediction based systems optimum results are produced by medical data mining. The K-means algorithm has been used to analyze different existing diseases. The cost-effectiveness and human

effects have been reduced using proposed prediction system based data mining.

Bala Sundar V, T Devi and N Caravan [13] examined real and artificial datasets that have been used to predict the diagnosis of heart diseases with the help of a K-mean clustering technique results to check its accuracy. The clusters are partitioned into k number of clusters by clustering which is the part of cluster analysis and each cluster has its observations with the nearest mean. The first step is the random initialization of whole data then a cluster k is assigned to each cluster. The proposed scheme of integration of clustering has been tested and its results show that highest robustness, accuracy rate can be achieved using it.

Daljit Kaur and Kiran Jyot [14] explained that data contained similar objects has been divided using clustering. A data of similar objects are in the same group and in case dissimilar objects occur then it will be compared with other group's objects. The proposed algorithm has been tested and results show that it is able to reduce efforts of numerical calculation, complexity along with maintaining an easiness of its implementation. The proposed algorithm is also able to solve the dead unit problem.

IV. PROBLEM FORMULATION

The prediction analysis is the technique which can predict the future possibilities from the existing data. The prediction analysis techniques are based on the clustering and classification. The CNN-MDRP modal for the prediction analysis is based on the neural networks. In which the clustering algorithm called k-means clustering is applied which can categorize the data into a certain number of classes. The clustered data is given as input to the classification algorithm which can divide the dataset into two parts testing and training. The SVM classifier is used to classify the data into a certain number of classes. In the k-mean clustering algorithm, the central points are calculated by taking arithmetic mean of the whole dataset which can reduce the accuracy of prediction analysis. When the dataset is complex, it is difficult to establish a relationship between the attributes of the dataset. In this research work, improvement in the k-mean clustering will be applied which can select centered points in an efficient manner to categorize input data. The proposed improvement directly increase the accuracy of clustering and reduce execution time.

V. OBJECTIVES

- The study of different prediction based algorithms and then analyzes it for data mining.
- To propose improvement in multimodal disease risk prediction (CNN-MDRP) algorithm.
- The proposed improvement will be based on back propagation algorithm which calculate relation between attributes of the dataset
- The existing and proposed algorithm will be compared in terms of time and accuracy.

VI. CONCLUSION

KDD can be used to extract relevant data from clinical databases, which are strongly correlated with well-known cardiovascular risk markers. In this paper, it is concluded that prediction analysis is the technique of data mining which is used to predict future from the current data. The prediction analysis is the combination of clustering and classification. The clustering algorithm group the data according to their similarity and classification algorithm can assign a class to the data. In this paper, various prediction analysis algorithms are reviewed and analyzed in terms of various parameters. The literature survey is done on the various techniques of prediction analysis from where the problem is formulated. The formulated problem can be solved in future to increase the accuracy of prediction analysis.

VII. REFERENCES

[1]. N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques over Heart Disease Data base" [IJESAT] international journal of engineering science & advanced technology ISSN: 2250–3676, Volume-2, Issue-3, 470 – 478

[2]. Asha Rajkumar, G.Sophia Reena, Diagnosis Of Heart Disease Using Datamining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[3]. K.Srinivas, B.Kavihta Rani, A.Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, (IJCSE) International Journal on Computer Science And Engineering Vol. 02, No. 02,2010,250-255.

[4]. Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S (2012), "*Reducing the Time Requirement of K-Means Algorithm*" PLoS ONE, vol. 7, 2012, pp-56-62.

[5]. K.Srinivas, Dr.G.Ragavendra and Dr. A. Govardhan," A Survey on prediction of heart morbidity using data mining techniques",International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, no.3, pp.14-34, May 2011.

[6]. Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network",International Conference on Computer and Communication Technology (ICCCT), DOI:10.1109/ICCCT.2010.5640377, 17-1

[7]. Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "*Disease Prediction by Machine Learning over Big Data from Healthcare Communities*", 2017, IEEE, vol. 15, 2017, pp- 215-227

[8]. Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal (2014), "*Clustering of Lung Cancer Data Using Foggy K-Means*", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.

[9]. Sanjay Chakraborty, Prof. N.K Nigwani and Lop Dey (2014), "*Weather Forecasting using Incremental K-means Clustering*", vol. 8, 2014, pp. 142-147.

[10]. Chew Li Sa, Bt Abang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, M. (2014), "*Student performance analysis system (SPAS)*", in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, vol.15, 2014, pp.1-6.

[11]. Qasem A. Al-Radaideh, Adel Abu Assaf and Eman Alnagi (2013), "*Predicting Stock Prices Using Data Mining Techniques*", The International Arab Conference on Information Technology (ACIT'2013), vol. 23, 2013, pp. 32-38.

[12]. K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin (2015), "*Comparative Analysis of K-Means Algorithm in Disease Prediction*", International Journal of Science, Engineering, and Technology Research (IJSETR), Vol. 4, 2015, pp. 1023-1028.

[13]. Bala Sundar V, T Devi and N Caravan, "*Development of a Data Clustering Algorithm for Predicting Heart*", International Journal of Computer Applications, vol. 48, 2012, pp. 423-428.

[14]. Daljit Kaur and Kiran Jyot (2013), "*Enhancement in the Performance of K-means Algorithm*", International Journal of Computer Science and Communication Engineering, vol. 2 2013, pp. 724-729