International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 135-139
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

A CRITICAL EXAMINATION OF DATA MINING CLASSIFICATION TECHNIQUES IN MEDICAL IMAGE DATABASES

Kavita
Research Scholar
Computer Science and Engineering
SVIET
Banur, Punjab, India

***Abstract*— Modern medicine generates a great deal of information stored in the medical database. Extracting useful data and making scientific decision for diagnosis and treatment of disease from the database increasingly becomes necessary. We propose a Heart diseases Prediction System for the society to prevent the cause of the death. We have applied Naïve-Bayes and ID3 algorithms and have compared both results. So we are analyzing heart disease patient to identify which treatment is most effective one and provide better result.**

Keywords— **Heart disease, Data mining, KDD, Classification, Classifier**

I. INTRODUCTION

Data can be an incredible advantage to various organizations of healthcare; however they must be initially changed into information. More requests are put on by utilizing this information for building knowledge that empowers the technique of organizations of healthcare: minimize cost and maximize care of patient. The environment of healthcare is seen as being rich in information and poor in knowledge. There is an abundance of administrative and clinical information accessible within systems of healthcare; in any case there is an absence of effective tools of analysis for discovering knowledge present in the databases of these systems. Knowledge Discovery in database (KDD) alludes to the "non trifling extraction of previously implicit unknown and conceivably helpful information about data". The central part of KDD is the data mining which is characterized as "a procedure of selection, investigation and determination, investigation and displaying of expansive amounts of information to find regularities or relations that are at first obscure with the point of getting clear and valuable results for the proprietor of database". The found information in social insurance databases can be utilized by human services overseers to move forward operations and nature of administration. It can be additionally utilized by social insurance experts to enhance their restorative practice and patient consideration. [1]

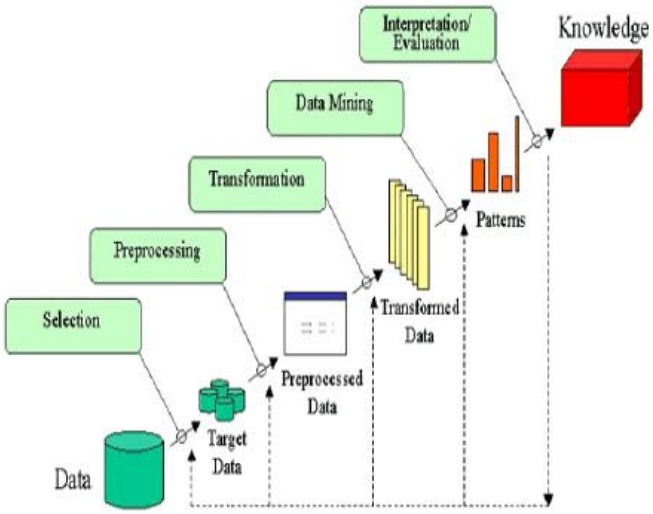

Fig. 1. KDD Process

II. MEDICAL DATA MINING

Medicinal Data Mining is an area of test which includes grouping of uncertainty and imprecision. Procurement of quality services at reasonable expense is the significant test confronted in the organization of health care. The decision of poor clinic may prompt tragic outcomes. Medicinal services information is huge. Clinical choices are regularly made in view of specialist's experience as opposed to on the learning rich information covered up in the data base. This now and again will bring about errors, over the top therapeutic cost which influences the nature of administration to the patients. Medicinal history information includes various tests essentials to analyze a specific infection. It is conceivable to pick up the benefit of Data mining in health care by utilizing it as a tool for diagnosis which is intelligent. The analysts in the field of medical distinguish and anticipate the sickness with the guide of Data mining systems. [2]

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 135-139
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

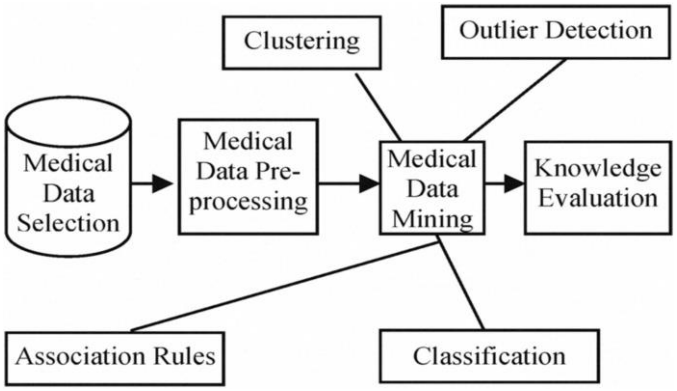

Fig. 2. A framework for medical data mining [4]

1) Association Rules: The major technique of data mining is mining of association rule, and is a most generally utilized patterns as a part of a data set. It has been broadly utilized as a part of analysis of data of medical. A case of the utilization of associations is the examination of the case frames put together by patients to a company of medical insurance. Each case structure contains an arrangement of therapeutic methodology that were performed on a given patient amid one visit. By characterizing the arrangement of things to be the gathering of every single medicinal technique that can be performed on a patient and the records to compare to every case frame, the application can discover, utilizing the function of association, connections among methods of medical that are frequently together performed.

2) Classification: The fundamental techniques for classification are Bayesian networks, neural networks, support vector machines and decision tree analysis. The probabilistic classifiers see the value of the status of categorization as the likelihood by the utilization of Bayes theorem.

$$P\,(c|d) = \frac{P\,(d|c)P(c)}{P\,(d)}$$

3) Clustering: Clustering and segmentation are the procedures of partition making so that every one of individuals from every arrangement of the partition are comparable as indicated by some metric. A cluster is an arrangement of objects gathered together due to their proximity or similarity. This procedure is utilized for finding similarities or structure in data. The most well-known metric in algorithm of clustering is the standard Euclidean distance:

$$D(Xi, Xj) = \sqrt{\Sigma}(xik - xjk)\text{^}2$$

Which is a particular case with p=2 of Minkowski metric

$$D_p(X_i, X_j) = \left(\sum_k \left(x_{ik} - x_{jk}\right)^p\right)^{1/p}$$

4) Outlier Detection: In few circumstances, rear case of medical might be more significant than expected case of medical. It can assist diagnosis by technology of clustering on data related to medical. Be that as it may, the patients' symptoms might don't have a place in any cluster. In this circumstance, doctor ought to take legitimate advice from experts of medical to keep away from treatment which is incorrect. [4]

III. HEART DISEASE

Heart is indispensable part or an organ of the body. Life is liable to capable working of heart. If operation of heart is not legitimate, it will impact the other parts of body of human, for instance, kidney, mind and so forth. Heart is basically a pump, which pumps the blood through the body. If blood in body is lacking then numerous organs like cerebrum endure and if heart stops working by, death happens within few minutes. There are number of components which assemble the risk of Heart disease [3]:

- Physical inertia
- Obesity
- High blood cholesterol
- High pulse
- Poor eating methodology
- family history of coronary illness

The diagnosis which is initial of a heart attack is made by a blend of changes of characteristic electrocardiogram (ECG) and symptoms related to clinical changes. An ECG is a recording of the electrical action of the heart. Affirmation of a heart attack must be made hours after the fact through identification of raised CPK (creatinine phosphokinase) in the blood. CPK is an enzyme of muscle protein which is discharged into the circulation of blood by dying the muscles of heart when their encompassing dissolves. [2]

IV. CLASSIFICATION TECHNIQUES

Classification model (Classifier) can be worked through the process of learning. It is depicting a foreordained arrangement of concepts or classes of data. The model is built from accessible data i.e. examples which are classified. These examples comprise of set of variables (features). This model is incited through the process of supervised learning; the classified examples would be processed as data of training.

A. Naïve Bayes Classifier

Naïve classifiers are well known and old sort of classifiers. They utilize a probabilistic approach, i.e, they attempt to process probabilities of conditional class and after that predicate the class which is most probable. Naïve classifiers utilize as their name shows off- Bayes rule and an arrangement of an assumption which is independent conditionally. Applying probabilistic approaches to techniques of classification particularly include modeling the conditional

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 135-139
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

probability distribution P(C |D), where C ranges over classes and D over data of descriptions, in some language, of objects of which classification is to be done. Given a description d of a specific object, we appoint the class arg max P(C = c|D = d). A Bayesian approach divides this posterior distribution into a prior distribution P(C) and a likelihood P(D |C):

$$\underset{c}{argmax}\, P(C = c|D = d) = \underset{c}{argmax} \frac{P(C = c|D = d)P(C = c)}{P(D = d)}.$$

B. Decision Tree

Decision trees are a prevalent consistent technique for classification. A decision tree is a progressive structure that partitions information into some groups which are disjoint in light of their distinctive trait values. The decision tree leaves contain records of one or about one class, thus it has been utilized for characterization. Leverage of strategies of decision tree is that decision trees can be changed over into rules which are understandable. A most generally utilized framework of decision tree is C4.5, its progenitor ID3, and a business adaptation C5.0 .Decision trees have been primarily used to construct determination models for medicinal information. When it is utilized for investigating designs as a part of restorative information, work in demonstrates that it is lacking for such investigation. In most of the cases, ID3 algorithm is used which utilizes the measure of information gain to choose among the applicant traits at every progression while developing the tree. Information gain is essentially the normal diminishment in entropy brought about by dividing the case as per this trait. The information gain, Gain (S , A) of A attribute, with respect to an accumulation of case S, is characterized as

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(S) = \sum_{i=1}^{c} p_i \log_2 p_i$$

where *pi* is the proportion of *S* belonging to class *i*.

C. K-Nearest Neighbor Classifier

Nearest neighbor classifiers depend upon analogy learning. The samples of training are portrayed by numeric attributes which are of n dimensions. Every sample speaks to a point in a space which is n-dimensional. Along these lines, all of the samples of training are put away in a pattern space which is of n-dimensions. At the point when given a sample which is unknown, a k-nearest neighbor classifier searches the space of pattern for the k samples of training that are closest to the sample which is unknown. "Closeness" is characterized as far as Euclidean distance, where the Euclidean distance between two points, X=(x1,x2,……,xn) and Y=(y1,y2,….,yn) is

$$d(X,Y) = \sqrt{\sum_{i=}^{n} (xi - yi)\text{^}2}$$

The algorithm of k-nearest neighbor is amongst the most straightforward of all algorithms of machine learning. The classification of object is done by a larger part of its neighbor's vote, with the object being allocated to the class most common amongst its *k* nearest neighbors. *k* is a positive integer, typically small. If *k*= 1, then the object is essentially allocated to the class of its nearest neighbor. In problems of binary (two class) classification, it is useful to pick *k* to be an odd number as this maintains strategic distance from tied votes. [6]

D. Support Vector Machine

Support vector machine guarantees a technique of machine learning on the premise of theory of statistical learning. It makes a hyperplane which is discrete in the descriptor space of data of training and classification of compounds is done taking into account the side of located hyperplane.
The benefit of SVM is that, by utilization of supposed "kernel trick", the separation between hyperplane and molecule can be estimated in a feature space which is non linear, lacking of transformation which is explicit of the descriptors which are original one. The kernel function is stated as follows [7]:

$$K(\bar{x}, \bar{x}_i) = \exp\left(-\frac{\| \bar{x} - \bar{x}_i \|^2}{2\alpha^2}\right)$$

V. Related Work

This section reviews the existing work on the prediction of heart diseases by using classification techniques.

Paper [8] presented an empirical study on prediction of heart disease using classification techniques of data mining. In this paper, the utilization of strategies of data mining and pattern recognition into models of risk prediction in the clinical area of cardiovascular pharmaceutical is proposed. The data is to be displayed and classified by utilizing characterization information mining system. A portion of the impediments of the ordinary restorative scoring frameworks are that there is a nearness of natural direct blends of variables in the information set and subsequently they are not adroit at displaying nonlinear complex collaborations in medicinal areas. This restriction is taken care of in this examination by utilization of characterization models which can verifiably identify complex nonlinear connections amongst needy and autonomous variables and additionally the capacity to recognize every conceivable association between variables of predictor.

In Reference [9], authors discussed about decision trees for early diagnosis of heart disease. Recent study demonstrates that

disease of heart is a main source of death in India and also in whole world. Critical life investment funds can be accomplished, if an auspicious and financially savvy clinical choice framework is produced. Unfavorable responses happen if an illness is not analyzed legitimately. A clinical decision emotionally supportive network can help medicinal services experts for early analysis of disease of heart from patient's medicinal information. Machine learning what's more, present day information digging techniques are helpful for anticipating what's more, grouping coronary illness. In this paper compelling exchanging approach of decision tree for ahead of schedule determination of coronary illness is presented. Substituting decision tree is another kind of rule of classification. It is a speculation of decision trees, voted decision trees and voted decision stumps. This methodology is connected on records of patients of heart disease gathered from different healing facilities in Hyderabad. Enhancement of elements enhances productivity of gaining calculation. PCA is utilized to decide fundamental components of coronary illness information.

In Paper [10], authors presented techniques of data mining in diagnosis and treatment of heart diseases. The accessibility of tremendous measures of medicinal information prompts the requirement for intense information examination apparatuses to extricate helpful learning. Analysts have for quite some time been worried with applying tools of statistics and data mining to enhance information examination on substantial information sets. Infection conclusion is one of the applications where information mining devices are demonstrating fruitful results. Heart disease is the main source of death everywhere throughout the world in the previous ten years. A few specialists are utilizing factual and information mining apparatuses to help human services experts in the determination of coronary illness. Utilizing technique of single data mining as a part of the conclusion of heart disease has been completely researched appearing worthy levels of exactness. As of late, analysts have been exploring the impact of hybridizing more than one system indicating upgraded results in the conclusion of disease of heart. Be that as it may, utilizing information mining methods to distinguish an appropriate treatment for coronary illness patients has gotten less consideration. This paper distinguishes crevices in the examination on conclusion and treatment of disease of heart and proposes a model to methodicallly close those holes to find if applying procedures of data mining to treatment of heart disease data can give as solid execution as that accomplished in diagnosing disease of heart.

Reference [11] presented a survey on classification algorithms for data mining. The concept of data mining is developing very quickly in fame, it is a technology that including techniques at the convergence of (database system, Statistics, Machine learning and artificial intelligence), the fundamental objective of procedure of data mining is to concentrate data from an extensive information into structure which could be justifiable for further utilize. A few algorithms of data mining are utilized to offer answers for arrangement issues in database. In this paper an examination among three arrangement's calculations will be concentrated on, these are (K-Nearest Neighbor classifier, Decision tree and Bayesian system) calculations. The paper will exhibit the quality what's more, precision of every calculation for arrangement in term of execution effectiveness and time unpredictability required. For model acceptance reason, twenty-four-month information investigation is led on a false up premise.

In Paper [12], authors demonstrated survey on the techniques of classification in data mining. Data mining is the step of analysis of KDD or Knowledge Discovery in Database. It is an interdisciplinary subfield of software engineering and the computational procedure of finding examples in vast information sets including techniques at the convergence of counterfeit intellectual prowess, machine learning, figures and pertinent information and database frameworks. Characterization is an information mining (machine learning) strategy used to foresee bunch participation for information example. In this paper, it bargains about the study of the few arrangement systems. Illustrations are a few methods for characterization technique, for example, fuzzy logic, k- nearest neighbor classifier, Bayesian networks and decision tree induction techniques.

V. CONCLUSION

Day by day healthcare data is increasing and having this huge amount of data that is being difficult to manage so mining techniques apply on it. We proposed a Heart disease prediction system that provides the important tool for physicians to take decisions from this huge and mined data for analysis based on previous data. The research undertakes an experiment on application of various data mining algorithms to predict the heart attack and to compare the best method of prediction. Different classification algorithms are used to analyze on heart disease patient data, it will check all the symptoms to predict the presence of heart disease and also measure the accurate result based on the performance of the algorithm. The predictive accuracy determined by Naïve Bayes, ID3 algorithms are measured and then we have compared both results and found out that Naïve Bayes algorithm is better . For the future study, analysis of heart disease patient based on the treatment and medicine provided by the doctors to find the best and effective treatment for the risky patient.

VI. REFERENCES

[1] Eman AbuKhousa, Piers Campbell, "Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems", *International Conference in Information Technology (IIT), IEEE,* 2012.

[2] Sivagowry. S, Dr. Durairaj. M, Persia. A, "An Empirical Study on Applying Data Mining Techniques for the Analsysis and Prediction of heart Disease", *IEEE.*

[3] Monika Gandhi, Dr. Shailendra Narayan Singh, "Predictions in Heart Disease using Techniques of Data

Mining", *1st International Conference on Futuristic Trend in Computational Analysis and Knowledge Management, IEEE*, 2015.

[4] Jitao Zhao, Ting Wang, "A General Framework for Medical Data Mining", *International Confernce on Future Information Technology and Management Engineering, IEEE*, 2010.

[5] Lamia AbedNoor Muhammed, "Using Data Mining Technique to diagnosis heart disease", *IEEE*.

[6] Thair Nu Phvu, "Survey of Classification Techniques in Data Mining", *Proceedings of International Multi Confernce of Engineers and Computer Scientists, Vol. 1,* March 2009.

[7] Dr. S. Vijyarani, Mr. S. Dhayanand, "Data Mining Classification Algorithms for Kidney Disease Prediction", *International Journal on Cybernetics & Informatics (IJCI), Vol. 4, No. 4,* August 2015.

[8] T. John Peter, K. Somasundaram, "An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques", *IEEE- International Confernce on Advances in Engineering, Science and Management (ICAESM),* March 2012.

[9] M.A.Jabbar, B.L. Deekshatulu, Priti Chndra, "Alternating decision trees for early diagnosis of heart disease", *Proceedings of International Confernce on Circuits, Communication, Control and Computing (I4C),* 2014.

[10] Mai Shouman, Tim Turner, Rob Stocker, "Using Data Mining Technqiues in Heart Disease Diagnosis and Treatment", *IEEE*, 2012

[11] Delveen Luqman Abd Al-Nabi, Shereen Shukri Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)", *Computer Engineering and Intelligent Systems, Vol. 4, No. 8*, 2013.

[12] M. Soundarya, R. Balakrishnan, "Survey on Classifiation Techniques in Data Mining", *International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7,* July 2014.

[13] Srinivas, K., B. K. Rani, A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", *International Journal on Computer Science and Engineering, Vol. 2, No. 2, pp. 250-255*, 2010.