DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5872

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY

Amandeep Kaur
M. Tech, Dept.CSE
Desh Bhagat University, Punjab, India

Jyoti Arora
Assistant professor, Dept.CSE
Desh Bhagat University, Punjab, India

Abstract: Heart disease is a common cause of death for people around the world. The overall examination on reasons for death because of coronary illness has been watched that it is the real reason for death. Analysis of these issues at beginning period helps the doctors in treating it at starting stage and to enhance the patient's wellbeing. In this manner the need to treat coronary illness that is found in individuals which precise entangled issues, if overlooked at beginning time. Different Data Mining Techniques can be used to analyze heart related issues. The essential point is an analysis of the Data Mining technique which is generally exact. There are different types of Data Mining Techniques such as Decision Tree, Naïve Bayesian, Support Vector Machine (SVM), K-NN classifier,Hybrid Approach, ArtificialNeural Network ANN). In this paper, we analyze different classification algorithms.

Keywords: Heart disease, Data Mining Techniques, Decision Tree, Naïve Bayesian, Support Vector Machine (SVM), K-NN classifier, Hybrid Approach, Artificial Neural Network (ANN).

1. INTRODUCTION

Data Mining is the way toward extracting interesting patterns and knowledge from huge amount of information. The Data Mining process is a combination of choosing, analyzing, planning, interpreting and evaluating the outcomes [1]. Numerous clinical finding achievement in the data mining techniques for prediction and clustering. Data mining comprises of the different specialized approaches including machine learning, database system and statistic [2]. The healthcare industry assembles immense measure of healthcare data which are not abundant to discover hidden information for effective decision making. Utilizing distinctive medical profiles such as sex, blood pressure, age, hypertension, lack of physical activity, blood sugar it can find the probability of patients getting a coronary illness [3]. Diagnosing machines or frameworks are quiet useful in this procedure because not every doctor must have the learning of each and every kind of problem of disease. In this manner automated diagnosing machine is used by them to diagnose the problem accurately. The WHO consortium has shared this information that ten a great many passing happen in this world is a consequence of coronary illness. so it was an extremely dangers problem in world. These systems typically create huge amounts of information which appear as numbers, charts and images. There are numerous sort of heart disease such as coronary heart disease, cardiomyopathy disease and cardiovascular sickness. Cardiovascular disease is an illness which specifically impacts the blood circulation in the body and blood vessels which are associated to the heart. Increment in heart disease is because of many facts like great BP, smoking, Family history etc. some unique factors that likewise causes heart illnesses are elevated cholesterol level, hyper solidness, improper diet etc[4].

2. LITRATURE SURVEY

Bala Sundar V et.al examined in this paper real and artificial datasets that have been used to predict diagnosis of heart diseases with the help of a K-mean clustering technique results to check its accuracy [1]. Sayali D. Jadhav et.al proposed that the issue goes for taking in the relationship between an arrangement of feature variables and a target variable of interest [2]. Mr. P Sai Chandrasekhar Reddy he proposed work separates proposed system in two sections such as execution model and prediction model. Performance model is intended to assess the general performance of the application [4].Dursun Delen, et.al [5] presented prescient models so as to investigate the immense databases of patients across the nation. . The performance results of mechanisms that include decision trees and neural networks are less exact that the ones with SVM calculations. David L. Olson, et.al [6] amongst the normal accuracy and decision tree size a tradeoff is given among this paper. Minimum parameters can be used keeping in mind the end goal to control the decision tree's size. Akhilesh Kumar Yadav, et.al [7] proposed algorithm has been tested by performing distinctive experiments on it that gives excellent outcome on essential data sets. In real world problem enhanced outcomes are accomplished utilizing foggy k-mean clustering algorithm as compared to existing simple k-means clustering algorithm. Daljit Kaur et.al [8]explained in this paper that data contained similar objects has been divided using clustering. The proposed algorithm has been tested and results shows that it is able to reduce efforts of numerical calculation, complexity along with maintaining an easiness of its implementation and also able to solve dead unit problem. Sanjay Chakrabotryet.al [9] stated that powerful tool clustering is used as different forecasting tools. The weather determining has been performed utilizing proposed incremental K-mean clustering generic technique. K. Raja lakshmi et.al[10]stated that K-means algorithm has been utilized to study distinctive extant illness. The cost

effectiveness and human effects has been reduced using proposed prediction system based data mining. Mustafa A. Al-Fayoumi[11]proposed an Associative Classification based on Incremental Mining (ACIM) algorithm in order to maintain the huge amount of information. Sajida Perveen, et.al[12] presented that J48 decision tree was utilized in order to apply adaboost and bagging ensemble methods in order to differentiate patients that are suffering from diabetes mellitus based on different elements that can cause diabetes. Basma Boukenze, et.al[13] studied the prediction of kidney disorder byusing numerous machine learning methods is the prior aim of this research. The algorithms that are included within this study are SVM, Decision Tree (C4.5), and Bayesian Network (BN). Nancy. P, et.al[14]investigated that performance of around fifteen data mining classification algorithms utilized within the data mining systems. Johan Holmgren, et.al [15]presented that utilization of support vector regression such that the numbers of bicycles that are being registered are predicted. Min Chen, et.al[16]proposed a novel Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP) algorithm and 94.8% of prediction accuracy was achieved here along with the higher convergence speed in comparison to other similar enhanced algorithms.

3. ANALYSIS OF THE DATA MINING TECHNIQUES

Various data mining techniques are usable and still lot of research is going on to find new techniques that can produce exact outcomes.

3.1 Decision Tree

The Decision tree is a classification strategies in which classification is done by the dividing criteria. The decision tree is a schema like a tree structure that gatherings instances by sorting them in perspective of the feature values. Each and every node in a decision tree depicted the features in a case to be classified. Decision tree makes the rule for the classification of the data set. The three fundamental algorithms are extensively utilized that are ID3, CART and C4.5.

A.ID3:ID3 stands for Iterative Dichotomiser 3 is an algorithm introduced by Ross Quinlan utilized to make a decision tree. ID3 algorithm is a classification algorithm in view of Information Entropy, its basic idea is that all examples are summarized to distinctive categories according to distinctive values of condition attribute set; its base is to decide the finest classification attribute form condition attribute sets.

B.C4.5: is the latest adaption of ID3 (Iterative Dichotomiser 3) induction algorithm. This assembles a decision tree like Iterative Dichotomiser 3.It develop a decision tree from planning dataset utilizing Information Entropy idea. With the objective that C4.5 is routinely called as Statistical Classifier. This C4.5 is a broadly utilized free data mining tool.

C. CART: It remains for Classification and Regression Trees. It was invented by Breiman in 1984. The classification tree development by CART is depends on binary separating of the properties. CART also based on Hunt's algorithm and can be executed serially. Gini index is utilized as splitting measure in picking the splitting attribute. CART is not quite same as other Hunt's based algorithm since it is additionally used for assistance of the regression trees.

3.2 Naïve Bayesian: Naive Bayes: The Bayesian Classification express as supervised learning strategy and statistical technique for classification. Assumes a hidden probabilistic model and it empower us to capture vulnerability about the model in a right way by deciding probabilities of the results. It can take care of diagnostic and predictive issues.

Naive Bayes algorithm is depends on Bayesian Theorem.

Bayesian Theorem:

Given training data X, back likelihood of a theory, H, P(H|X), takes after the Bayes hypothesis

$P(H|X)=P(X|H)P(H)/P(X)$

3.3 SVM: Support Vector Machine algorithm make good judgement for data points that are outside the preparing set. There are two classes of information in SVM. The data points are separated such that they could draw a horizontal line on the figure. The line is made in a way that it isolates every one of the focuses on one side of one class and every one of the focuses on the reverse side of alternate class. When such circumstance happens, then the data are linearly separable. The line used to isolate the dataset is known as a separating hyperplane. The points nearest to the isolating hyperplane are called as support vectors. Kernels are utilized to extend SVMs to a bigger number of datasets. Mapping of one feature space to another is finished by kernel. Kernel method, maps the information (in some cases likewise called as nonlinear information) from a little dimensional space to an extensive dimensional space. In a bigger measurement, it decides straight issue that is nonlinear in smaller-dimensional space. The Radial Bias Function (RBF) is a prominent kernel that measures the separation among two vectors.

3.4 K-Nearest Neighbor classifier (KNN): KNN is a basic, lazy and nonparametric classifier. KNN is preferred when every one of the features are persistent. KNN is likewise called as case-based reasoning and has been utilized in numerous applications like statistical estimation, pattern recognition. Classification is distinguishing the closest neighbor to decide the class of an unknown sample. KNN is favored over other classification algorithms due to its high merging velocity and straightforwardness. KNN characterization has two phases:

a) Find the k number of examples in the dataset that is nearest to instance S

b) These k number of examples at that point vote to decide the class of instance S

The Accuracy of KNN relies upon separate metric and K value. Different methods for estimating the separation between two instances are cosine, Euclidian distance. To evaluate the new unidentified sample, KNN figure out its KNN and assign a class by dominant part voting.

3.4 Hybrid Approach: An efficient prediction method to resolve and extract the unidentified knowledge of heart disease using hybrid combination of K-means clustering algorithm and Artificial Neural Network (ANN). The primary goal of using K-means clustering procedure is that it sorts out the data into classes The main aim of this clustering is to discover the positions μi, i=1...k within group to decrease sum of squares separate from the centroid. K-means algorithm calculate on k clusters, and it may fast for various solutions. So to discard such dependency,

modified or enhanced k-means was proposed. K-means is accompanied with Lloyd's algorithm to dispose of condition.In this paper Neural Network and K-mean clustering utilized as hybrid approach to increase accuracy. Utilizing this strategy the outcomes show the quality of clusters is not compromised. Steps for K-means algorithm are:

1. Compute the intermediate of the clusters from n data points xi, i=1...n that must be divided in k clusters
2. Attribute the nearest cluster to every data point utilizing Euclidean distance
3. Set the position of each cluster to domain of each data points directing toward that cluster
4. Repeat stages 2-3 until merging in our structure K-means algorithm assume an essential role in order to collect the related number of data clusters. Using this algorithm along with Euclidean distance centroids are figured for distinctive patient attribute.

3.5 Artificial Neural Network: Artificial Neural Network (ANN) is a mathematical structure in view of biological neural networks. Artificial Neural Network is depends on perception of a human brain .Human brain is extremely web of neurons. Analogically artificial neural network is arrangement of three simple units specifically input, hidden and output unit. The parameters that are passed as input to the following structure a first layer. In medical finding patients hazard factors are treated as input to the neural network.

4. ANALYSIS RESULTS

The accuracy of different classification algorithms:

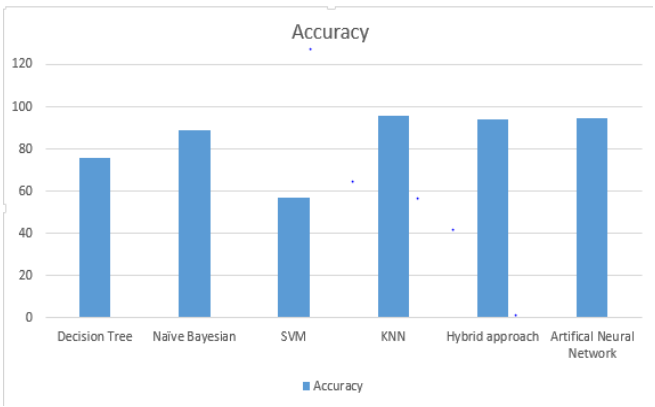

Figure:1 Accuracy Comparison

5. CONCLUSION

Heart Disease is an incurable disease by its nature. This disease makes a dangerous complexities such as heart attack and death. The relevance of Data Mining in the Medical field is acknowledged and steps are produced to apply applicable strategies in the Disease Prediction. The different research works with some compelling procedures done by various people were studied. Though, various classification techniques are widely used for Heart Disease Prediction.

6. FUTURE SCOPE

In future, we are planning to introduce an efficient disease prediction system to predict the heart disease with better accuracy utilizing different data mining classification techniques such as Decision Tree,Naïve Bayes, and Support Vector Machine(SVM).

REFERENCES

1. BalaSundar V, T Devi and N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications, vol. 48, pp. 423-428,2012
2. Sayali D. Jadhav, H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques", International Journal of Science and Research (IJSR) , Volume 5, Issue 1 , Paper ID: NOV153131,2016
3. SellappanPalaniappan, RafiahAwang "Intelligent Heart Disease Prediction System Using Data Mining Techniques" IEEE, pp.978-1-4244-1968,2008
4. Mr. P Sai Chandrasekhar Reddy, Mr. PuneetPalagi, Ms. Jaya, "HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING"IJCSMC, Vol. 6, Issue. 4, pg.168 – 172,2016
4. DursunDelen, AsilOztekin, Leman Tomak, "An analytic approach to better understanding and management of coronary surgeries", Decision Support Systems vol. 52, pp.698–705,2012
5. David L. Olson, DursunDelen, YanyanMeng, "Comparative analysis of data mining methods for bankruptcy prediction", Decision Support Systems vol.52, pp.464–473, 2012
6. Akhilesh Kumar Yadav, DivyaTomar and Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, pp.121-126,2013
7. Daljit Kaur and KiranJyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, vol. 2, pp. 724-729,2013
8. Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey , "Weather Forecasting using Incremental K-means Clustering", vol. 8, pp. 142-147,2014
9. K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin , "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, pp. 1023-1028,2015
10. Mustafa A. Al-Fayoumi, "Enhanced Associative classification based on incremental mining Algorithm (E-ACIM)", IJCSI International Journal of Computer Science Issues, Volume 12, Issue 1, 2015
11. SajidaPerveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Procedia Computer Science vol.82, pp.115 – 121,2015
12. BasmaBoukenze,HajarMousannifandAbdelkrimHaqiq, "PERFORMANCE OF DATA MINING TECHNIQUESTO PREDICT IN HEALTHCARE CASE STUDY:CHRONIC KIDNEY FAILURE DISEASE", International Journal of Database Management Systems (IJDMS) Vol.8, No.3,2016
13. MonireNorouzi,AlirezaSouri, and Majid SamadZamini, "A Data Mining Classification Approach for Behavioral Malware Detection",Hindawi Publishing CorporationJournal of Computer Networks and Communications,2016
14. Nancy.P, Sudha.V, Akiladevi.R, "Analysis of feature Selection and Classification algorithms on Hepatitis Data", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 6, Issue 1,2017

15. Johan Holmgren, Sebastian Aspegren, Jonas Dahlström, "Prediction of bicycle counter data using regression", Procedia Computer Science 113,pp. 502–507,2017

16. Min Chen, YixueHao, Kai Hwang, Fellow,Lu Wang, and Lin Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, pp. 215-227,2017