www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 8 Issue 8 August 2019, Page No. 24829-24833
ISSN: 2319-7242 DOI: 10.18535/ijecs/v8i08.4354

Comparative study of Data Mining Techniques used to Predict Risk of Heart Disease

Tejali Mhatre[1], Satishkumar Varma[2]

Department of Information Technology, Pillai College of Engineering, New Panvel, India

Abstract
Heart disease is considered as one of the major causes of death throughout the world. Classification of heart disease can be valuable for medical practitioners to predict disease at early stage. To get accurate result of prediction of disease medical diagnosis system very useful. There are different data mining techniques available to classify medical data related to heart disease. In this paper, survey is carried out to find which are different data mining techniques can be used for heart disease prediction system.

Keywords – Heart disease, Data Mining, Support vector machine, ANN, Decision Tree algorithm, K nearest neighbor, heart disease prediction system.

I. Introduction

Heart disease refers to the class of diseases that involve the heart or blood vessels which also known as Cardiovascular disease. Cardiovascular disease technically refers to any disease that affects the cardiovascular system. It is usually used to refer to those related to atherosclerosis. Cardiovascular diseases include coronary heart disease, cerebrovascular disease, raised blood pressure, peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. In practice, cardiovascular disease is treated by cardiologists, thoracic surgeons, vascular surgeons, neurologists, and interventional radiologists, depending on the organ system that is being treated. The heart is the organ that pumps blood to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidney suffer and if the heart stops working, death occurs within minutes. The World Health Organization (WHO) has estimated that 17.9 million deaths occur worldwide, every year due to heart disease.

Data mining is a nontrivial extraction of implicit, previously unknown potential useful information called as knowledge from the medical data using complex algorithms. Big data (BD) can be referred as huge record of information set. Big data and data mining are two different things. The task carried out by these two methods are similar focusing on collecting the huge amount of data, handling them and preparing report on the data by taking out the information which is knowledgeable. Data mining is basically an activity of observing the patterns in the data which is relevant and with particular information by using big data. The useful patterns with hidden patterns, unknown correlations are analytically handled for making knowledgeable decision through this BD analytics process.

Medical diagnosis is an important yet complicated task that needs to be done accurately and efficiently. The automation of this system is very much needed to help the physicians to do better diagnosis and treatment. The representation of medical knowledge, decision making, choice and take into consideration. Medical progress is

always supported by data analysis which improves the skill of medical experts and establishes the treatment technique for diseases. The purpose of medical diagnosis system is to assist physicians in defining the risk level of an individual patient.

The heart disease dataset found in University of California, Irvine Machine Learning Repository is used for training and testing the system [10]. The purpose of using this dataset is to provide a complex, real world data example where the relationships between the features are not easily discovered by casual inspection. In this proposed system, the advantages of genetic algorithm and neural network are combined to predict the risk cardiovascular disease. Genetic algorithm is an optimization algorithm that mimics the principles of natural genetics. It finds acceptably good solutions to problems acceptably quickly. In many applications, knowledge that describes desired system behavior is contained in datasets. When datasets contain knowledge about the system to be designed, a neural network promises a solution because it can train itself from the datasets. Neural networks are adaptive models for data analysis particularly suitable for handling nonlinear functions. By combining the optimization technique of genetic algorithm with the learning power of neural network, a model with better predictive accuracy can be derived.

II. Literature Survey

In [1], M. Anbarasi, E. Anupriya and N. Ch. S. N. Iyengar used Genetic algorithm to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient. It is used to reduce thirteen attributes to 6 attributes using genetic search. Subsequently, they use three classifiers like Naive Bayes, Classification by clustering and Decision Tree are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. Also, the observations exhibit that the Decision Tree data mining technique out performs other two data mining techniques after incorporating feature subset selection with relatively high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering performs poor compared to other two methods.

In [2], Sa delma Banu N.K, Suma Swamy proposed heart disease prediction system using two techniques, Support Vector Machine and Artificial Neural Network. In their work they compare result between two techniques by measuring accuracy. As a result, they stated that SVM gives more accuracy than ANN.

In [3], Salma Banu N.K and Suma Swamy conducted survey from 2004 to 2015 which gives the idea of different models available and the different data mining techniques used. The accuracy obtained with these models is also mentioned. It is observed that all the technique available have not used big data analytics. Use of big data analytics along with data mining will give promising results to get the best accuracy in designing the prediction model.

Hai H. Dam and Hussein A. Abbass in [4] proposed a system which incorporate NNs into UCS. Proposed system was named as NLCS which is neural based learning classifier system.

In [5], they studied the problem of constraining and summarizing different algorithms of data mining. They focused on using different algorithms for predicting combinations of several target attributes. In paper, they presented an intelligent and effective heart attack prediction methods using data mining.

III. Techniques And Methods For Prediction

To predict the risk disease there are mainly four techniques available namely support vector machine, artificial neural network, decision tree algorithm, K nearest neighbor algorithm.

A. Support Vector machine

SVM is machine learning technique that realize the idea outlined above. SVM used to find optimal

hyperplane as a solution to learning problem. This is supervised machine learning model of classification.to predict the class label, it uses hyperplane. Within n-dimensional space each data forms coordinates. The hyperplane divides the data into different class labels using maximum margin. Data points near hyperplane are called as a support vectors. This classification process generates non-linear decision boundaries and classify new records. The formula for hyperplane is

$$f(x) = w.x + bias$$

Where w is weights and x is input data.

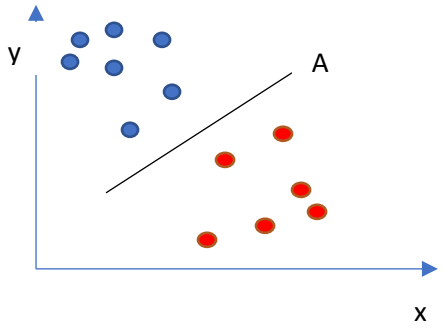

Fig 1. Support vector machine

B. Decision tree

Decision tree is classification technique which form tree like structure. It is one of predictive model which uses statistics to predict the class label. It consists of leaf node, decision node and root node. Root node is topmost node where decision node is resulting node. Each node has to calculate entropy and information gain. Entropy is measure of randomness of unpredictability of dataset. Information gain shows how well a given attribute separates training dataset. This technique used to find largest information gain and assign the node with largest information gain as final output.

C. Neural Network

Neural network works exactly same as biological neuron system. The human brain is a highly complex structure viewed as a massive, highly interconnected network of simple processing elements called neurons. Artificial neural network is very useful for predictive model. Based on input data it changes its structure. It calculates error for each input attribute and minimize it by adjusting weights of the network.

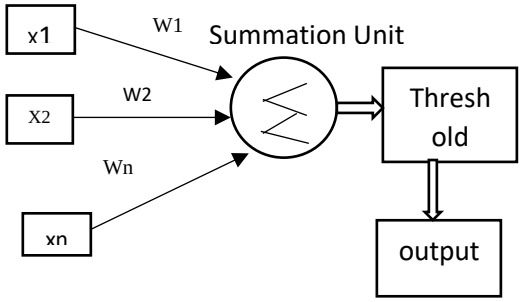

Fig 2. Neural network

D. K nearest neighbor

K-nearest neighbor classification algorithm is a well-known method for classifying an unseen instance. It is done by classifying the instances closest to it. KNN classification algorithm works by finding K training instances that are close to the unseen instance. This is done by using distance measurements such as Euclidean, Manhattan, maximum dimension distance, and others. Finally, the algorithm decides the class for the unseen instance by taking most common class in the nearest K instances.

IV. Heart Disease Prediction System

Heart disease prediction system is important in medical field. This system can predict the disease at early stage so that patients can start treatment as soon as possible. The prototype for heart disease prediction system consist of mainly four parts data collection, preprocessing, classification technique and output.

Dataset can be collected from standard dataset library. Data preprocessing is stage where data cleaning, removing noise, data transformation carried out.in the classification technique phase algorithm like SVM, ANN, Decision tree i.e. ID3, C4.5 and KNN can be applied. After completing all stages result of risk prediction is analyzed

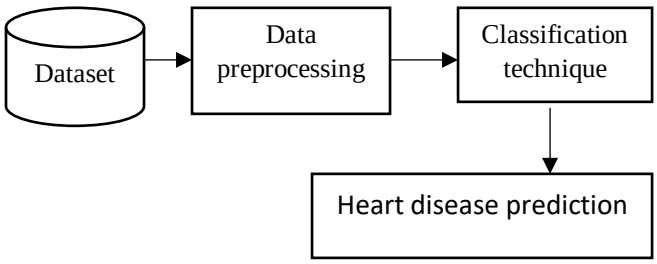

Fig 3. Heart disease prediction system

V. Comparative Study

S. No.	Reference paper	Technique used	Accuracy	Advantages	Disadvantages
1	M. Anbarasi et al. [1]	Naïve Bayes	96.5%	This technique is more robust	It takes more time
2	M. Anbarasi et al. [1]	Decision Tree	99.2%	Easy to understand	Sensitive to noise
3	Sa delma Banu N.K et al.[2]	SVM	84.7%	This technique is more robust	It takes more time
4	Suma Swamy [2]	ANN	81.8%	Suitable for nonlinear data	It takes more time to train data
5	Mrs. S. Radhimeenakshi [4]	NCL	81%	It is self-organized network	Huge training time required
6	Yanwei Xing, Jie Wang and Zhihong Zhao [7]	Decision tree	89.6%	It is very compact and very fast	Greedy algorithm
7	Yanwei Xing, Jie Wang and Zhihong Zhao [7]	SVM	92.1%	Works efficiently for large number of attributes	Wrong choice of kernel can affect result

VI. Conclusion

In this paper, various techniques for heart disease prediction system is focused. These techniques are classification methods of data mining. Each method is studied individually. Comparative study gives idea about each algorithm. Advantages and disadvantages of each algorithm is described along with accuracy.

References

[1] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar," Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 2010.

[2] Sa delma Banu N.K, Suma Swamy," Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey", 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT).

[3] Mrs. S. Radhimeenakshi," Classification and Prediction of Heart Disease Risk Using Data Mining Techniques of Support Vector Machine and Artificial Neural Network",2016 International Conference on Computing.

[4] Hai H. Dam, Hussein A. Abbass," Neural-Based Learning Classifier Systems", IEEE Transactions On Knowledge And Data Engineering, VOL. 20, NO. 1, JANUARY 2008.

[5] S.Florence, N. G. Bhuvaneswari Amma, G. Annapoorani, K.Malathi,"Predicting the Risk of Heart Attacks using Neural Network and Decision Tree", Delhi Business Review,Vol.8, No.1, pp.99-101,2007.

[6] Shantakumar B.Patil and Y.S.Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack prediction", International Journal of Computer Science and Network Security ,Vol.9, No.2, pp.228-235, 2009.

[7] Yanwei Xing, Jie Wang and Zhihong Zhao," Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease",

International Conference on Convergence Information Technology,2007.

[8] S.Rajasekaran and G.A.Vijayalakshmi Pai, "Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications", Prentice Hall of India, 2007.

[9] http://www.ics.edu, UCI Repository of Machine Learning Data bases, Cleveland Heart Disease Dataset.

[10] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufman Publishers, 2009

[11] Monika Gandhi and Dr. Shailendra Narayan Singh, "Predictions in Heart Disease Using Techniques of Data Mining", International Conference on Futuristic trend in Computational Analysis and Knowledge Management,2015.