International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)
Volume No.5 Issue: Special 4, pp: 790-991 20 May 2016

PREDICTION OF CARDIOVASCULAR RISK ANALYSIS AND PERFORMANCE EVALUATION USING VARIOUS DATA MINING TECHNIQUES: A REVIEW

Bindushree D C
Assistant Professor,
School of Computing & IT, REVA University,
Kattigenahalli, Yelahanka, Bengaluru, India
bindushreedc@reva.edu.in

Dr. Udaya Rani V
Senior Associate Professor,
School of Computing & IT, REVA University,
Kattigenahalli, Yelahanka, Bengaluru, India
udayamurthy@revainstitution.org

***Abstract*—In the recent trends of technology implementation, the knowledge discovery in database (KDD) is alarmed with development of methods and techniques for making use of data. This can be achieved by one of the most important step of the KDD: Data Mining. Data mining is the process of pattern discovery and extraction where huge amount of data is involved i.e., process of analyzing enormous sets of data and then extracting the meaning of the data. Both the data mining and healthcare industry have emerged some reliable early detection systems and other various healthcare related systems from the clinical, and diagnosis data. The data generated from this prediction for the heart disease are complex and voluminous to be processed and very difficult to be analyzed using some of the existing traditional methods. The techniques and methodologies available in data mining help to transform this huge amount of data into specific and useful data for decision making. These data mining techniques consume less time for the prediction of the disease with more accuracy, to achieve the same. In this paper we have reviewed various paper involved in terms of algorithms, methodologies used, and results in this field. Results and evaluation methods are discussed and a summary of the finding is presented to conclude the paper. Applying data mining techniques to heart disease data can provide as reliable performance as that achieved in diagnosing heart disease.**

Keywords—KDD, Heart disease, Data mining, Data mining techniques

I. INTRODUCTION

Heart Diseases remain the biggest cause of deaths for the last two decades. Myocardial infarctions (MI) and strokes are the first and fourth leading causes of death. Life is dependent on efficient working of heart because heart is essential part of our body. If functioning of heart is not proper, it will affect the other body parts of human body such as brain, kidney etc. There are number of factors which increases risk of Heart disease. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to the Heart diseases. In 2008, 17.3 million people died due to Heart Disease. Over 80% of deaths in world are because of Heart disease. WHO estimated by 2030, almost 23.6 million people will die due to Heart disease as written in [3].

The main objective of our paper is to emphasize on different techniques of data mining used in prediction of heart disease. Accurate results of the disease can be obtained by prediction using data mining techniques. To extract the hidden knowledge and discover the data associated with heart disease from historical heart disease database IHDPS (Intelligent Heart Disease Prediction System) is used. By diagnosing heart disease it can answer complex queries helping healthcare analysts and practitioners to make intelligent clinical decisions which traditional decision support systems cannot.

Machine learning techniques and computer technology come together to develop a software which assist doctors in making a decision of heart disease in the early stage based on clinical and pathological data. In biomedical diagnosis, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patients suffer from more than one type of disease of the same category. It becomes very difficult for the physicians to diagnose it correctly. Data mining with intelligent algorithms can be used to tackle the said problem of prediction in medical dataset involving multiple inputs. Artificial neural network has been used for complex and difficult tasks in recent days. The neural network is trained from the historical data with the hope that it will discover hidden dependencies and that it will be able to use them for predicting. Feed forward neural networks trained by back propagation have become a standard technique for classification and prediction tasks. Clinicians and patients need reliable information about an individual's risk of developing Heart Disease, ideally, a perfect model to estimate the risk to categorize people with heart disease is achieved only with accurate data. Indeed, the perfect model would even be able to predict the timing of the disease's onset. Heart disease comprises of cardiovascular diseases, heart attack, coronary heart disease and stroke. Stroke is a type of heart disease it is caused by narrowing, blocking, or hardening of the blood vessels that go to the brain or by high blood pressure [20, 21].

1.1. Leading Risk factors for Heart Disease

The risk factors for heart disease can be divided into modifiable and non modifiable. Modifiable risk factors include:

i. *Obesity*: The term obesity is used to describe the health condition of anyone significantly higher than one's ideal healthy weight. Being obese puts anybody at a higher risk for health problem such as heart disease, stroke, hypertension, type 2 diabetes and more.

ii. *Smoking*: Smoking serves the major cause of heart attack, other peripheral arterial disease, and stroke. Nearly 40% of people who die from smoking tobacco, due to heart and blood vessel diseases. Once a smoker stop smoking, the risk of heart attack reduces rapidly after only one year.

iii. *Cholesterol*: Improper levels of lipids (fats) in the blood are risk factor of heart diseases. Cholesterol is a soft, waxy substance in the lipids of the bloodstream and in body's cells. High level of triglyceride (most common type of fat in body) combined with high levels of LDL (low density lipoprotein) cholesterol speed up atherosclerosis increasing the risk of heart diseases.

iv. *High blood pressure*: High blood pressure also known as HBP or hypertension is a widely mis-conceptualized medical condition. High blood pressure increase the risk of the walls of our blood vessels walls becoming widened and damaged. It increases the risk of having heart attack or stroke and of developing heart failure, kidney failure and peripheral vascular disease.

Feature Name	Median (IQR) or N (%)	% Missing	Description
Gender			
Female	105,234 (60.98)	0	
Male	67,337 (39.02)	0	
Age (Years)	42.00 (30.00 - 55.00)	0	Age at the end of the baseline period
SBP (mm Hg)	118.00 (110.00 - 128.00)	0	Mean systolic blood pressure during baseline period
BMI (kg/m^2)	26.72 (23.60 - 31.06)	10	Body mass index
LDL (mg/dL)	118.00 (96.00 - 142.00)	66	Final low density lipoprotein during baseline period
HDL (mg/dL)	48.00 (39.00 - 58.00)	55	Final high density lipoprotein during baseline period
TRG (mg/dL)	108.00 (77.00 - 158.00)	66	Final tryglyceride during baseline period
Smoking			Smoking status in EMR
Never or Passive	126,463 (73.28)	0	
Quit	16,898 (9.79)	0	
Current	29,210 (16.93)	0	
Comorbidity			Presence of comborbidities related to cardiovascular disease
Yes	14,547 (8.43)	0	
No	158,024 (91.57)	0	
SBP Meds		0	Number of SBP medication classes filled during baseline period
0	118,685 (68.77)	0	
1	20,613 (11.94)	0	
2	13,883 (8.04)	0	
3+	19,390 (11.24)	0	
LDL Meds		0	
0	154,972 (89.80)	0	Number of LDL medication classes filled during baseline period
1+	17,599 (10.20)	0	

Table 1: Summary measures of the risk factors included in our prediction models in the entire study cohort.

The non modifiable risk factors for heart disease are age, gender and family history of heart diseases. That if anybody has a family history of heart disease, one may be at greater risk for heart attack, stroke and other heart diseases, age, and gender. Many people have at least one heart disease risk factor.

II. LITERATURE SURVEY

In today's competitive world, the ability to extract useful knowledge hidden in the large amount of data and to act on this data has become a challenge as the data size is growing day to day. The need to understand large, complex, information enriched data sets has now increased in all the varied fields of technology, business and science. The process of applying computer based information system (CBIS), including new techniques, for discovering knowledge from data is called data mining [6]. This paper aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction.

2.1. Various data mining techniques

i. *Association*: In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in heart disease prediction as it tell us the relationship of different attributes used for analysis and sort out the patient with all the risk factor which are required for prediction of disease. It is the best known technique in data mining.

ii. *Classification*: Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. It is one of the classic models of data mining.

iii. *Clustering*: Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. For example in prediction of heart disease by using clustering we get cluster or we can say that list of patients which have same risk factor, this makes the

Author	*Year*	*Technique Used*	*attribute*
Carlos et al	2001	association rules	25
Dr. K. Usha Rani	2011	Classification	13
		Neural Networks	
Jesmin Nahar , et al	2013	Apriori	14
		Predictive Apriori	
		Tertius	
Latha et al.	2008	genetic algorithm	14
		CANFIS	
Majabbar et al	2011	*Clustering*	14
		Association rule mining.	
		Sequence number.	
Ms. Ishtake et al.	2013	Decision Tree	15
		Neural Network	
		Naive Bayes	
Nan-Chen et al	2012	(EVAR)	
		Machine learning	
		Markov blanket	
Oleg et al.	2012	artificial neural network	
		genetic polymorphis ms	
Shadab et al	2012	Naive bayes	15
Shantakumar et al	2009	MAFIA	13
		Clustering	
		K-Means	

Table 2: Shows different data mining techniques used in the diagnosis of Heart disease over different Heart disease datasets

iv. separate list of patients with high blood sugar and related risk factor and so on.

v. *Prediction*: This discovers relationship between independent variables and relationship between dependent and independent variables as name implied is one of a data mining techniques.

III. RELATED WORK

3.1. Data mining techniques used heart disease prediction

In this section the demining systems used for the classification of heart disease is analyzed. N. Deepika et al. proposed Association Rule for classification of Heart-attack patients . The extraction of significant patterns from the heart disease

data warehouse was presented. The screening clinical data of heart patients is stored in the data warehouse which is pre-processed to make the mining process more efficient. Handling the missing values in the pre-processing is the first stage of Association Rule. Next equal interval binning with approximate values based on medical expert advice on Pima Indian heart attack data. For all frequent patterns with the aid of the proposed approach the significant items were calculated. The frequent patterns with confidence greater than a predefined threshold were chosen and it was used in the design and development of the heart attack prediction system. The Pima Indian Heart attack dataset used was obtained from the UCI machine learning repository. Characteristics of the patients like number of times of chest pain and age in years were recorded. The actions comprised in the pre-processing of a data set are normalizing the values used to represent information in the database the removal of duplicate records, removing unneeded data fields, and accounting for missing data points. It might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm [4].

In the real world, data is incomplete and in case of medical data it is very obvious. To remove the number of inconsistencies which are associated with data we use Data pre-processing Application of Data Mining Technique in Healthcare and Prediction of Heart Attacks presented by K. Srinivas et al. [10]. Classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network are used to the massive Volume of healthcare data. Tanagra data mining tool was used for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analyzed. The results of comparison are based on 10 tenfold cross-validations. According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing. The comparison made among these classification algorithms out of which the naive Bayes algorithm considered as the best performance algorithm. The performance of various algorithms is listed below [10].

The algorithm used	Accuracy	Time taken
Naïve Bayes	52.33%	609ms
Decision list	52%	719ms
K-NN	45.67%	1000ms

Table 3: Performance Study of Data mining Algorithms

The algorithm used Accuracy and Time taken in Naïve Bayes 52.33% 609ms, Decision list 52% 719ms and K-NN 45.67% 1000ms. Diagnosis of heart disease was used Naïve Bayes, K-NN, Decision List in this Naïve Bayes has taken a time to run the data for accurate result when compared to other algorithms. Sudha et al. [1] to propose the classification algorithm like Naïve Bayes, Decision tree and Neural Network for predicting the stroke diseases. The records with irrelevant data were removed from data warehouse before mining process occurs.

Data mining classification technology consists of classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. The testing data set was used for testing the classification efficiency. Then the classification algorithm like decision tree, naive Bayes and neural network was used for stroke disease prediction. The performance evaluation was carried out based on three algorithms and compared with various models used and accuracy was measured. While comparing these classification algorithms, the observation shows the neural network performance was more than the other two algorithms.

M A. Jabbar et al. proposed Association Rule mining based on the sequence number and clustering for heart attack prediction [12]. The entire database is divided into partitions of equal size. The dataset with 14 attributes was used in that work and also each cluster is considered one at a time for calculating frequent item sets. This approach reduces main memory requirement. To predict the heart attack in an efficient way the patterns are extracted from the database with significant weight calculation. The frequent patterns having a value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals were defined based on data exploration and all those models could answer complex queries in predicting heart attack.[18].

Mai Shouman, et al. [14] proposed k-means clustering with the decision tree method to predict the heart disease. In their work they suggested several centroid selection methods for k-means clustering to increase efficiency. The 13 input attributes were collected from Cleveland Clinic Foundation Heart disease data set. The sensitivity, specificity, and accuracy are calculated with different initial centroids selection methods and different numbers of clusters. For the random attribute and random row methods, ten runs were executed and the average and best for each method were calculated.

When comparing integrating k-means clustering and decision tree with traditional decision tree applied previously on the same data set, integrating k-means clustering with decision tree could enhance the accuracy of decision tree in diagnosing heart disease patients. In Addition, integrating k-means clustering and decision tree could achieve higher accuracy than the paging algorithm in the diagnosis of heart disease patients. The accuracy achieved was 83.9% by the enabler method with two clusters.

Author	Technique used	Data mining tool	Accuracy	Objective
Abhishek et al (2013)	J48	Weka 3.6.4	95.56%	HDP System Using DM Techniques
	Naive Bayes		92.42%	
	J48		94.85%	
Chaitrali et al (2012)	Neural Network	Weka 3.6.6	100%	Prediction of HD
Monali Et al	C4.5	WEKA		Study and Analysis of Data mining Algorithms for Healthcare Decision Support System
	Multilayer Perceptron			
	Naïve Bayes			
Nidhi et al (2012)	Naive Bayes	Weka 3.6.6	90.74%, 99.62%, 100%	Analysis of HDP using Different DM Techniques
	Decision Trees	TANAGRA	52.33%, 52%, 45.67%	
		Weka 3.6.0	86.53%, 89%, 85.53%	
	Neural	.NET	96.5%,	
	networks	platform	99.2%, 88.3%	
Resul et al (2009)	Neural networks	SAS base software 9.1.3	97.4%	Diagnosis of valvular HD
Rashe-Dur Et al (2013)	Neural Network	WEKA	79.19%	Comparison of Various Classification Techniques
	Fuzzy Logic	TANAGRA	83.85%	
	Decision Tree	MATLAB		
Resul et al (2009)	Neural networks	SAS base software 9.1.3	89.01%	diagnosis of HD

Table 4: Shows different Data Mining tools used on heart disease detection with prediction accuracy

3.2. ALGORITHMS USED IN HEALTHCARE DATA

Healthcare data comprises of a large amount of data which includes EMR (Electronic Medical Records), administrative reports and other benchmarking techniques [12], detailed process of diagnosis, prevention and treatment of disease, any injuries, physical and mental impairments in humans. Data mining is able to search for new and valuable information from these large volumes of data mainly for predicting various diseases as well as in assisting for diagnosis for the doctors in making their clinical decision.

i. *Anomaly Detection*: Anomaly detection is used in discovering the most significant changes in the data set [2]. Bo Lie et al [9] had used three different anomaly detection method, standard support vector data description, density induced support vector data description and Gaussian mixture to evaluate the accuracy of the anomaly detection on uncertain dataset of liver disorder dataset which is obtained from UCI. The method is evaluated using the AUC accuracy. The results obtained for a balanced dataset by average was 93.59%. While the average standard deviation obtained from the same dataset is 2.63.

ii. *Clustering*: The clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data [2]. Rui Velosoa [17] had used the vector quantization method in clustering approach in predicting the readmissions in intensive medicine. The algorithms used in the vector quantization method are k-means, k-mediods and x-means. The datasets used in this study were collected from patient's clinical process and laboratory results. The evaluation for each of the algorithms is conducted using the Davies-Bouldin Index. The k-means obtained the best results while x-means obtained fair results while the k-mediods obtained the worst results. From the results the works by these researchers provide a useful result in helping to characterize the different types of patients having a higher probability to be readmitted.

iii. *Classification:* Classification is the discovery of a predictive learning function that classifies a data item into one of several predefined classes [2].

IV. DISCUSSION

The different data mining methods discussed in this paper, the accuracy of the results varies depending on the features of the data sets and the size of data set between the training and testing sets taken from various papers. Most common features among the healthcare data sets are highly imbalanced data sets, where by the minority and majority classifier are not balanced resulting prediction erroneous when run by the classifiers. Another characteristics of healthcare data sets are the missing values. The sample size of the data is often seen as other characteristics as the data available are usually in small scale. There is no one suitable data mining method to resolve all this issues.

The numerous heart attack predicting system techniques are presented in this paper. In this paper Heart attack prediction system methodology is categorized in two types. At first type data mining technique (mainly classification technique) are analyzed. The second type intelligent techniques used for heart disease prediction are analyzed. In data mining approach the heart disease data warehouse contains the screening clinical data of heart patients used for heart disease diagnosis. Data mining techniques combined with intelligent and evolutionary computation are discussed in this paper. In many papers author used the dataset of Heart disease from UCI Machine Learning Repository, University of California. Hence it might be a good choice for training.

V. CONCLUSION

Heart disease is one of the leading causes of death worldwide and the early prediction of heart disease is important. The computer aided heart disease prediction system helps the physician as a tool for heart disease diagnosis. Some Heart Disease classification system is reviewed in this paper. From the analysis it is concluded that, data mining plays a major role in heart disease classification. Neural Network with offline

training is a good for disease prediction in early stage and the good performance of the system can be obtained by pre-processed and normalized dataset. The classification accuracy can be improved by reduction in features.

The objective of our work is to provide a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. The analysis shows that different technologies are used in all the papers with taking different number of attributes. So, different technologies used shown the different accuracy to each other. In some papers it is shown that neural networks given the accuracy of 100% in prediction of heart disease. On the other hand, this is also given that Decision Tree has also performed well with 99.62% accuracy by using 15 attributes [15].

So, different technologies used shown the different accuracy depends upon number of attributes taken and tool used for implementation. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of heart disease. Although applying data mining techniques to help health care professionals in the diagnosis of heart disease is having some success, the use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention.

REFERENCES

[1] A. Sudha, P. Gayathiri and N. Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", International Journal of Computer Applications, Vol. 43, No. 14, pp. 0975 – 8887, 2012.
[2] Carlos Ordonez, Edward Omiecinski, Mining Constrained Association Rules to Predict Heart Disease, IEEE. Published in International Conference on Data Mining (ICDM), p. 433440, 2001.
[3] Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology, 2012.
[4] D. Shanthi, G. Sahoo and D. K. Bhattacharyya and S. M. Hazarika, Networks, Data Mining and Artificial Intelligence: Trends and Future Directions, 1st ed. Narosa Pub House, 2006.
[5] Dr. K. Usha Rani, analysis of heart diseases dataset using neural network approach, International Journal of Data Mining & Knowledge Management Process, 2011.
[6] G. E. Vlahos, T. W. Ferratt, and G. Knoepfle, "The use of computer-based information systems by German managers to support decision making," Inf. Manag., vol. 41, no. 6, pp. 763– 779, 2004.
[7] H. Thomas and L. Paul, Statistics: Methods and Applications, 1st ed. StatSoft, Inc, 2005.
[8] I. H. Witten, E. Frank, and M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). 2011.
[9] K. Srinivas, G. Raghavendra Rao and A. Govardhan, "Survey on prediction of heart morbidity using data mining techniques", International Journal of Data Mining & Knowledge Management Process, Vol. 1, No. 3, pp. 14 - 34, 2011.
[10] K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, "Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering, Vol. 02, No. 02, pp. 250 - 255, 2011.
[11] Latha Parthiban and R.Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Biological and Medical Sciences, 2008.
[12] M A. Jabbar, Priti Chandra and B. L. Deekshatulu, "Cluster based association rule mining for heart attack prediction", Journal of Theoretical and Applied Information Technology, Vol. 32, No.2, pp. 197 - 201, 2011
[13] Mai Shouman, Tim Turner and Rob Stocker, "Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients", Proceedings of the International Conference on Data Mining, 2012.
[14] M. Karegar, A. Isazadeh, F. Fartash, T. Saderi, and A. H. Navin, "Data-Mining by Probability-Based Patterns," pp. 353– 360, 2008.
[15] Nan-Chen Hsieh & Lun-Ping Hung & Chun-Che Shih, Intelligent Postoperative Morbidity Prediction of Heart Disease Using Artificial Intelligence Techniques, J Med Syst, 2012.
[16] Oleg Yu. Atkov, Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters, Elsevier, 2012.
[17] Shantakumar B.Patil, Dr.Y.S. Kumaraswamy, Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction, (IJCSNS) International Journal of Computer Science and Network 228 Security ,2009.
[18] Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naïve bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012.
[19] Tom Dent, "Predicting the risk of coronary heart disease", PHG foundation publisher, 2010.
[20] World Health Organization, "Global status report on no communicable diseases", 2010.G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)
[21] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
[22] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
[23] K. Elissa, "Title of paper if known," unpublished.
[24] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[25] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[26] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.