International Journal of Advanced Computer Science and Information Technology (IJACSIT)
Vol. 6, No. 1, 2017, Page: 1-9, ISSN: 2296-1739

www.elvedit.com

Proposing a Data Mining Approach for Cost Reduction of Heart Disease based on Decision Tree

Authors

Alireza Motaharzadeh alireza.motahar@yahoo.com
Department of Computer Engineering/ Damavand Branch, Islamic Azad University, Damavand, Iran Damavand, Iran

Javad Hosseinkhani jhkhani@gmail.com
Department of Computer Engineering/ Damavand Branch, Islamic Azad University, Damavand, Iran Damavand, Iran

Abstract

Nowadays heart disease is very common and is a major cause of mortality .proper and early diagnosis of this disease is very important. Diagnostic methods and treatments for this disease have many side effects and are so expensive. Therefore, researchers are looking for cheaper ways with high precision to diagnose this disease. Studies used characteristic collected from patients and various algorithms of data mining to increase the accuracy. In this thesis, a data set including important and effective characteristics for diagnosis of heart disease is collected. The data set in this thesis is collected from 118 cardiac patients that were referred to Heart Specialized Hospital in Jamaran. Data mining with extracting knowledge from the abundance of medical information can effectively help doctors in prediction and diagnosis proper treatment of diseases. To predict , diagnose and treatment of cardiac patients ,four models of decision tree data mining models and also artificial network was carried out on data set, which decision tree algorithms had the highest accuracy with 74/74 percent , that in my study is the highest achieved accuracy. Attention to high risks of performing invasive diagnostic methods in cardiac patients, including coronary angiography. On the other hand, successful experiences in data mining methods in medicine have been obtained; therefore this study has been presented the model based on data mining techniques that have capability to predict heart disease and diagnosis proper treatment for cardiovascular diseases. In this study, data mining with predicting that a person have heart disease or not, or diagnosis proper treatment method, can help doctors to reduce the cost of treatments of cardiac patients and quality of presenting better services.

Key Words

Heart Disease, Data mining, Diagnosis and treatment, Cost Reduction

I. INTRODUCTION

Modern data mining knowledge is one of the developing knowledge which has consolidated its position among all kinds of fields in recent years in such way that its growth is increasing in comparison with other top knowledges. Data mining is a tools to uncover the essential knowledge to solve problems in various fields. Nowadays, data mining has become a competitive advantage in many fields of industry and business. Data mining can be defined as the process of extracting knowledge, patterns and trends that were not previously known of databases and use of data to build predictive models [1].

Healthcare is one of the areas that in recent years have been the center of data mining experts' consideration. One of the important challenges of medical organizations, indicating the quality of their services against the affordable prices for consumers of these services. The quality of services depends on the correct diagnosis and prescribed the correct treatment and poor clinical diagnosis can lead to undesirable results that are unexpected. Quality of services includes correct identification of disease and effectiveness of their treatment. Clinical poor diagnosis can lead to dire consequences that are unacceptable. Also, hospitals are required to minimize the cost of clinical testing. Using computer-based information and decision-making systems can carry us away from such errors. Nowadays many hospitals use species of hospital information system (HIS) to manage patient data or healthcare usage [2].

These systems create a lot of data, including numbers, texts, graphs and photos. Unfortunately, these data are rarely used in clinical decision-making. Valuable asset of hidden information exist in these data. The main question that preoccupied the minds of most decision-makers in this area is how we can transform data into useful information that medical staff will be able to make smart clinic decisions. Hospitals must also reduce their cost of clinical trials. They can achieve this using proper computer data and decision support systems. Medical data are diversified including data related to patient, resource management data, costs and etc. Medical organizations must have the ability to properly analysis data in order to achieve success. These organizations can only achieve their objectives in this way and keep indicators of public health at an optimal level of standards and a kind of community can ensure the growth and flourish and prosperity of its own people that have the ability to supply health for its own people in first step. On the other hand medical environments in terms of information are technically rich environments. But in contrast, this wealth is faced with knowledge poverty. Massive amounts of data are exist in medical systems, and this an important advantage to be able to increase the quality of services offered in the field of healthcare to a desirable level by extracting hidden knowledge within this information using data mining techniques and apply this knowledge in procedures. In between the field of heart diseases due to sensitivity of heart health in continuing human life, has a particular importance and improving the diagnosis and treatment of these areas can save many lives.

The decision support system in the field of cardiac diagnosis and proposed treatment for patients in this study is designed based on data mining. This is an experimental study designed to collect needed information and in order to perform data mining, going to hospital and obtaining the necessary information from the records of heart disease was required. The

information contained in the records of heart patients consisted of several parts and different tests of the patient's health status that have used medical specialist for mapping to different modes of person's health. After the information preparation, they will be imported in the form of database in Clementine 12 software for analysis and desired outcomes were extracted.

II. Review Related Work

Heart is the most important part of everyone's body and human life depends on the proper functioning it. If the heart would not function properly, other parts of the body such as brain and kidney will be affected [3]. Roa et.al developed a predict heart disease system prototype called smart heart using decision tree data mining techniques, Bayesian systems and neural networks. This system answers complex medical questions based on if that traditional systems cannot be held accountable. Using data mining on medical parameters such as age, gender, blood pressure, blood sugar and etc., possibility to get heart diseases can be predicted. This system is web-based, user-friendly, scalable, reliable and expandable [4].

Kumar and Research Associates surveyed on cardiovascular data set diseases in relation with data mining techniques such as classification, decision trees, artificial neural network (ANN), and support vector machine (SVM). The performance of these techniques was compared through sensitivity, specificity, precision, error rate, true positive rate and false positive rates. Error rate results for a decision tree, ANN and SVM were 2.756, 0.2755, 0.2248 and 0.1588, respectively. Precision and accuracy of reviews in decision tree model, ANN and SVM were 81.08, 79.05, 80.06 and 84.12%, respectively. The analysis indicates that SVM prediction model has the minimal error rate and highest precision and proper accuracy performance for coronary heart diseases among these four classification models [5].

Srinivas indicated the rate of heart disease that has been reported by patients in Andhra Pradesh region of India that Singareni coal mine also located in this region is higher compared to other regions. They predict heart disease by variables such as age, ethnicity, education level of the individual, income, body mass index, etc., and with the help of classification techniques in data mining and data mining algorithms such as neural networks, Bayesian network, decision tree and support vector machine that decision tree had higher precision than other algorithms [6].

Leung have used data mining in decision tree, neural networks and support vector machine algorithms for classification of data and the results were compared with each other [7].

Chaitraly et al. have used neural networks in their study titled prediction and diagnosis of heart disease using neural networks for diagnosis and prognosis of heart disease that 13 physician parameters were used in order to perform data mining. Eventually the accuracy of predicting heart disease is 100 percent [8].

Christine et al. have shown in their study that the decision trees are more capable to predict the possibility of getting heart disease than logistic regression model [9].

Ordonze et al. have used C5.0 decision tree algorithm, and 8 correlation rules algorithms using 25 risk factors for heart disease prediction and concluded that correlation rules generally creates easier predictable rules than decision trees [10].

Karaolis et al. have used decision tree algorithm in a study entitled predict coronary artery disease in order to assess the risk factors for coronary artery disease. They classified risk factors into two general category as prior risk factors and posterior risk factors. Its findings are similar to the rules extracted from decision tree algorithm in the present study [11].

Jyoti et al. have used the decision tree model that has the highest accuracy rate (83 percent) to predict the risk of getting heart disease. However, the decision tree presented in this study had a higher accuracy. The noted difference can be shown from the greater number of variables used in the study.

III. PROPOSED METHOD

Data mining is an iterative process and therefore can be repeated many times [1]. To ensure the successful implementation of a data mining project, always should have a clear methodology for implementation. CRISP-DM methodology according to the nature of the data has been selected among methodologies that have been defined as standard data mining procedures. Based on this methodology, data mining project consists of a six-stage life cycle that processes interact with each other. CRISP cycle is shown in Figure 1.

The first stage is CRISP-DM standard process that objectives and requirements will be clearly characterized. Objectives translation and its constraints will be defined in the formulation, data mining problem definition and providing a basic strategy for achieving objectives.

- Data recognition stage

This stage includes data collection for heuristic analysis usage and also defining the basic information for initial assessment of high quality data and selection of useful required data.

Data preparation stage

Preparing raw data to final data, these data are used in all subsequent stages and therefore this stage requires more analysis and effort. Elements and analyzed identifiers selection will be assigned to data exploration and preparing them for modeling tools by cleaning raw data.

- Modeling stage

We optimize the modeling results by selecting and applying appropriate modeling techniques and given data mining methods which in case of requiring we can go back and optimize the modeling analysis.

- The assessment Stage

At this stage we find out if whether the selected model we bring us to our determined objective in first stage or not. Make decisions about using data mining results for validation is done at this stage as well.

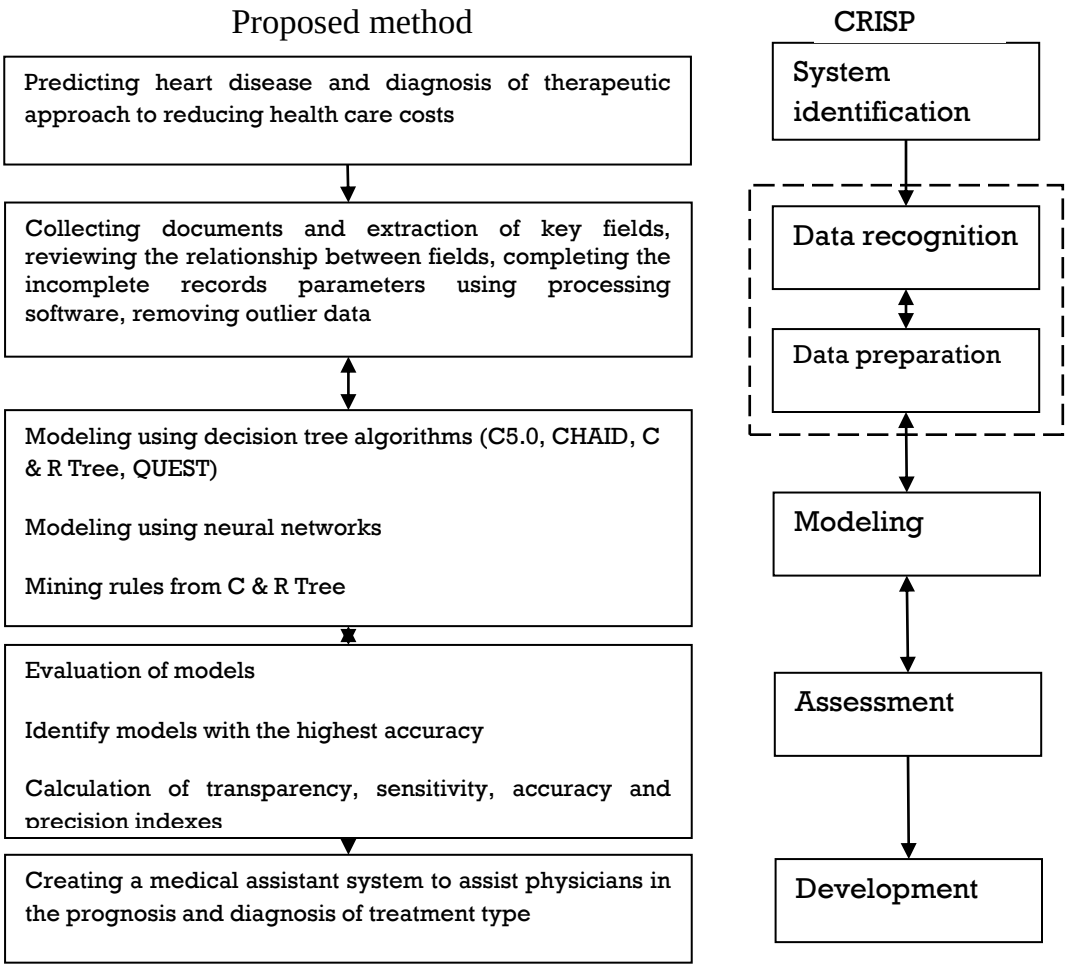

FIGURE 1. CRISP CYCLE

- Development stage

Using created model, for example, can produce a simple report of outputs and for a complex example, is the completion of processing of parallel data mining in other fields which these templates are used and converted into useful knowledge and after improving them, patterns that are effective will be used in a running system.

- Data preparation stage

Preparing raw data to final data, these data are used in all subsequent stages and therefore this stage requires more analysis and effort. Elements and analyzed identifiers selection will be assigned to data exploration and preparing them for modeling tools by cleaning raw data. In the data preparation stage data collection was done by specialists in order to cleanse data sets. First, empty fields of some heart disease cases were initialized by referring to the processor software and checking their medical history, including the history of medications and experiments. Records that still had missing values at this stage were omitted. Also some records had outlier data that were optimized based on averaging and deducing same data type and records were

omitted that had no similarity with other data. Therefore in data preparation stage, 125 initial records of patients were eventually remained that has reduced to 118 final records after refining and remove some records. Specification heart patient records are given in Table I.

TABLE I: SPECIFICATION OF HEART PATIENT RECORDS

Type	Description	Characteristic
Nominal	-	Sex
Numerical	-	Age
Rating	-	Blood Pressure
Rating	-	Blood Sugar
Rating	-	Cholesterol
Rating	-	Cigarette
Numerical	-	Weight
Rating	-	Job
Nominal	-	Cure

- Modeling

There are many data mining methods for modeling. In this phase various data mining techniques are used to find the optimized model and pattern. Modeling is done by using SPSS Clementine 12.0. Method in this study is predictive data mining. The decision tree algorithm is used in order to obtain the best ratio between different fields. A decision tree is used for classification. Classification is known as one of the best data mining techniques and include of the two stages. In the first stage that is deduction stage, the objective is to discover a model to define pre-identified data categories. Model is created based on training samples provided by the system. Deduction algorithm creates definitions for that particular category using characteristic values of samples that belong to each category. In the second stage that is prediction; by using deduction model we can predict the belonging of those samples that their belonging to any particular category is not specified.

Implementation of decision tree algorithms consist of C5.0, CHAID, C & R Tree, QUEST and neural networks on existing data was to predict methods of curing heart disease patients that accomplishing this was done by measuring blood pressure, blood sugar and smoking as the most important and affecting factors on heart disease. In order to train a decision tree, a categorical variable should be the output field and one or more input field should be existed. Input fields were obtained values fro, patient's experiments and output was considered as the type of cure. Set of rules usually hold the most related information to the decision tree. Tree will break into sub-branches based on a divided based on a standard branch division and this process continues until finally the data of each nod would be put on one category. To express the extracted rules, root direction to leaf will be scrolled and rules will be expressed conditionally. Tags of model category are described in Table II.

TABLE II: TAGS OF MODEL CATEGORY

Description	Category tag
Angiography: In this method the number of blocked coronary artery, the blockage and it is level will be cleared and the most direct way to detect coronary artery problems is the heart.	1
Angioplasty: In this method, the coronary arteries that became narrowed or blocked by fatty deposits and blood clots will be opened without surgery.	2
Medical treatment	3
Pace Maker: Is part of the heart or the device that artificially imitates beating created the heart beat and regulate its harmony.	4
Open Heart: If the number of narrow vessels were high, or narrowing is in a dangerous place such as main arteries in a dangerous place such as left main trunk artery or at the beginning of the coronary artery or at the coronary artery bifurcation site, there is no possibility of angioplasty and requires coronary artery bypass graft surgery.	5

IV. EVALUATION OF MODELS

In this phase of the modeling should be paid to evaluating the results of modeling. The evaluation results will improve the model and make it usable. To evaluate the model, the under study data was divided into two parts as training and testing. Data from the training sector (85 percent) produce the tree and test data (15%) test the produced tree and determine the tags of these records. There are various indicators for evaluating categorizing methods such as transparency, sensitivity, precision and accuracy that are calculated according to equations 1 to 4.

The confusion matrix is used to calculate the amount of indexes. This matrix is a useful tool to analyze the performance of categorizing method in data diagnosis or observations of other categories. Best case scenario is that the most of relevant data to observations locate on the diagonal of the matrix and the remaining values of matrix are zero or close to zero. The accuracy of the generated models to the data is shown in Table 3. The best results were obtained using the C & R Tree node. So the C & R Tree is used to determine treatment methods.

TABLE III: THE VALUES OF ACCURACY OF DATA MINING MODELS

Accuracy(Percent)	Model	Algorithm	Row
74.74	Algorithm C&R Tree	Decision tree	1
73.68	Algorithm Quest		2
67.37	Algorithm Chaid		3
65.26	Algorithm C 5.0		4
58.95	Neural network algorithm	Neural network	5

V. Conclusion & Future Works

In this study by implementing four models of decision tree models and also neural network, we compared them based on their accuracy which decision tree algorithms had the highest accuracy rate in prognosis and diagnosis of treatment method for heart patients and can help doctors to predict and diagnosis of appropriate individual treatment with characteristics such as age, sex, blood pressure, blood sugar, fat, overweight, smokers. The knowledge derived by C & R Tree models can be recommended as a model for predicting appropriate among all decision tree models treatment that indicates the affecting rules of heart disease patient's characteristics in prognosis and diagnosis treatment method. According to the cardiologist we can say that the risk characteristics of older age, smoking, high blood pressure, high levels of fat have the most effect in the treatment method and this is while according to comparisons based on prioritizing variables by reviewing algorithms, these variables have been accounted as the prime factors, which indicates the importance of these variables in predicting and treatment method. In this study, data mining by predicting if a person have heart disease or not or diagnosis and treatment method help to reduce the cost of curing heart disease and helps physicians provide better quality of their services.

VI. References

[1] Hosseinkhani, Javad, Suhaimi Ibrahim, Suriayati Chuprat, and Javid Hosseinkhani Naniz. "Web Crime Mining by Means of Data Mining Techniques." Research Journal of Applied Sciences, Engineering and Technology 7, no. 10 (2014): 2027-2032.

[2] Oztekin, Asil, Dursun Delen, and Zhenyu James Kong. "Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology." international journal of medical informatics 78, no. 12 (2009): e84-e96.

[3] Dangare, Chaitrali S., and Sulabha S. Apte. "A data mining approach for prediction of heart disease using neural networks." International Journal of Computer Engineering & Technology (IJCET) 3, no. 3 (2012): 30-40p.

[4] Roa, Diego, Javier Bautista, Nicolás Rodríguez, María Del Pilar Villamil, Andres Jiménez, and Oscar Bernal. "Data mining: A new opportunity to support the solution of public health issues in Colombia." In Computing Congress (CCC), 2011 6th Colombian, pp. 1-6. IEEE, 2011.

[5] Pandey, Atul Kumar, Prabhat Pandey, and K. L. Jaiswal. "A heart disease prediction model using Decision Tree." IUP Journal of Computer Sciences 7, no. 3 (2013): 43.

[6] Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) 2, no. 02 (2010): 250-255.

[7] Leung, KwongSak, KinHong Lee, JinFeng Wang, Eddie YT Ng, Henry LY Chan, Stephen KW Tsui, Tony SK Mok, Pete Chi-Hang Tse, and Joseph JY Sung. "Data mining on dna sequences of hepatitis b virus." IEEE/ACM transactions on computational biology and bioinformatics 8, no. 2 (2011): 428-440.

[8] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47, no. 10 (2012): 44-48.

[9] Colombet, Isabelle, Alan Ruelland, Gilles Chatellier, François Gueyffier, Patrice Degoulet, and Marie-Christine Jaulent. "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron

and logistic regression." In Proceedings of the AMIA Symposium, p. 156. American Medical Informatics Association, 2000.

[10] Ordonez, Carlos. "Comparing association rules and decision trees for disease prediction." In Proceedings of the international workshop on Healthcare information and knowledge management, pp. 17-24. ACM, 2006.

[11] Karaolis, Moutiris, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis. "Association rule analysis for the assessment of the risk of coronary heart events." In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6238-6241. IEEE, 2009.

[12] Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17, no. 8 (2011): 43-48.