International Journal of Economics & Management Sciences

Assari et al., Int J Econ Manag Sci 2017, 6:3
DOI: 10.4172/2162-6359.1000415

Research Article Open Access

Heart Disease Diagnosis Using Data Mining Techniques

Ramin Assari[1]*, Parham Azimi[2] and Mohammad Reza Taghva[1]

[1]*Department of IT Management, Allameh Tabataba'i University, Tehran, Iran*
[2]*Faculty of Mechanical Engineering and Industrial Engineering, Islamic Azad University, Qazvin, Iran*

Abstract

In recent decades, heart disease has been identified as the leading cause of death across the world. However, it is considered as the most preventable and controllable disease at the same time. According to World Health Organization (WHO), the early and timely diagnosis of heart disease plays a remarkable role in preventing its progress and reducing related treatment costs. Considering the ever-increasing growth of heart disease-induced fatalities, researchers have adopted different data mining techniques to diagnose it. According to results, application of the same data mining techniques leads to different results in different datasets. This study tries to assist healthcare specialists to early diagnose heart disease and assess related risk factors. To this end, the main heart disease diagnosis indices were identified using experts' opinions. Then, data mining techniques were applied on a heart-related dataset. Finally, the main heart disease diagnosis indices were identified and a model was developed based on extracted rules. Visual Studio was used to write the algorithm code.

Keywords: Bayesian network; Data mining; Decision tree; Heart disease; K-nearest neighbor; Support vector machines

Introduction

In the past decade, heart disease has been the leading cause of death in different continents and countries in the world, regardless of the income level of countries [1]. According to WHO report, heart disease is the leading cause of death across the world, accounting for 7.2 million deaths, i.e., 12.8% of all fatalities in the world [2]. Figure 1, illustrates deaths from heart disease across the world (scale: 1:100000). According to recent research predictions, cardiovascular diseases will become the leading cause of death up to 2030.

Although cardiovascular diseases have been identified as the leading cause of death in the world in the past decade, they have been introduced as the most preventable and controllable diseases [3]. The complete and correct treatment of a disease depends on the timely diagnosis of that disease [4]. An accurate and systematic tool for identifying high-risk patients and extracting data for timely diagnosis of heart disease seems a critical need.

Every day, modern computer-based systems collect large amounts of data using automatic data record systems in different fields. Data mining technology is the product of the evolution of database technology, IT and storage devices [5]. The current challenge is to make data mining and knowledge discovery systems applicable to a wider range of domains [6]. Researchers are adopting data mining techniques to diagnose different diseases including diabetes [7], stroke [8], cancer [9] and heart disease [10]. Considering the high rate of cardiovascular-induced fatalities, researchers have tried to adopt data mining systems to diagnose heart disease [11].

Literature Review

Every day, modern computer-based systems collect large amounts of data using automatic data record systems in the healthcare field where data mining can extract a valuable knowledge from them. The next section briefly explains heart disease and the application of data mining techniques in treating such diseases.

Heart disease

As the leading cause of death in the world, heart disease, according to WHO, accounts for 3.8 million and 3.4 million deaths in males and females, respectively.

The symptoms and incidence of heart disease differ from one person to another. However, they commonly include chest pain, jaw pain, neck pain, back pain, stomach disorders; arms and shoulders pains and shortness of breath [12]. Different heart problems induce different heart diseases including coronary artery disease, heart failure and stroke [13].

Although heart disease has been identified as the most chronic disease across the world, it is the most preventable one at the same time. A healthy life style (primary prevention) and timely diagnosis (secondary prevention) are two main elements of heart disease control. Conducting regular check-ups (secondary prevention) plays a remarkable role in the diagnosis and early prevention of heart disease complications [14]. Several tests including, chest X-rays, angiography, echocardiography and exercise tolerance test contribute to this important issue. However, these tests are costly and require accurate medical equipment.

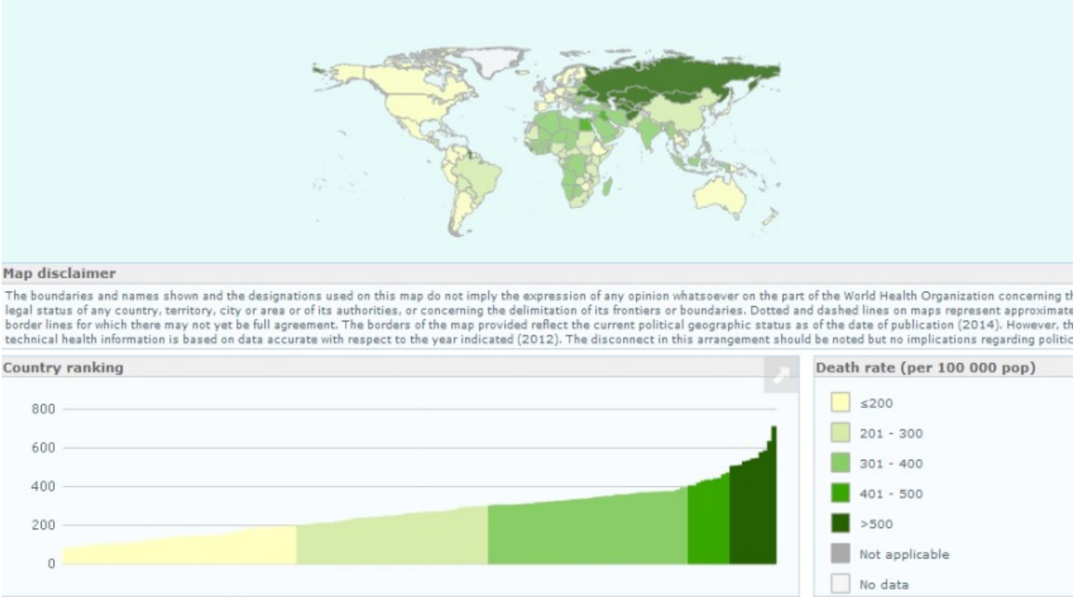
Applications of data mining to healthcare data

Data mining scholars have long studied the application of tools and equipment in improving the process of data analysis in large and complex datasets. Adopting data mining techniques in the medicine field is of high importance in diagnosing, predicting and deeply understanding of healthcare data. These applications include treatment centers analysis aimed at improving treatment policies and prevention of any mistake in hospitals, early diagnosis of diseases, prevention of diseases and hospital death reduction.

***Corresponding author:** Ramin Assari, IT Management, Management Department, Allameh Tabataba'i University, Iran, Tel: 98 2144737510; E-mail: assari20002000@gmail.com

Received December 13, 2016; **Accepted** March 27, 2017; **Published** March 29, 2017

Citation: Assari R, Azimi P, Taghva MR (2017) Heart Disease Diagnosis Using Data Mining Techniques. Int J Econ Manag Sci 6: 415. doi: 10.4172/2162-6359.1000415



Map disclaimer

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning th legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement. The borders of the map provided reflect the current political geographic status as of the date of publication (2014). However, th technical health information is based on data accurate with respect to the year indicated (2012). The disconnect in this arrangement should be noted but no implications regarding politic

Figure 1: Heart diseases-induced fatalities across the world in 2013.

Heart specialist's record and store large amounts of patients' data. This provides a great opportunity for extracting a valuable knowledge from such datasets. Researchers are adopting statistical approaches as well as data mining techniques to help treatment and healthcare specialists diagnose and determine heart disease risk factors in patients. Statistical analyses have identified a number of risk factors for heart diseases including age, blood pressure, smoking [15], total cholesterol [16], diabetes [17], and hypertension, heart disease background in family, obesity and lack of physical activity [18]. The awareness of heart disease risk factors assists treatment and healthcare specialists to identify patients who are subject to high risk factors.

Researchers have employed different data mining techniques to help specialists and physicians diagnose heart disease [19]. Some techniques are more common such as Naïve Bayes, decision tree and K-nearest neighbor. However, there are other classification-based data mining techniques such as kernel density, neural network, bagging algorithm, sequential minimal optimization, direct Kernel self-organizing map and support vector machine. The next section briefly explains those techniques which were used in this study.

Decision tree

There are different types of decision trees. They only differ in the mathematical model they use to select the class of attribute during rule extraction. Gain ratio decision tree is the most common, successful type [20]. It is a relationship between entropy (information gain) and classified information.

In entropy technique, the attribute which minimizes entropy and maximizes information gain is selected as the tree root. To select tree root, it is first necessary to calculate the information gain of each attribute. Then, the attribute maximizing information gain should be selected. Information gain, or entropy, is derived from relation 1.3 [21].

$$E = -\sum_{i=1}^{k} p_i \, log_2^{p_i} \qquad (2\text{-}1)$$

Where k is the number of response variable classes, p_i is the ratio of the number of the ith class events to total number of samples (occurrence probability of i)

Bayesian network

Bayesian network is a statistical technique predicting the membership class of the studied sample using the probability theory. Bayesian network practices classification process in accordance with Bayes' theorem. It assumes that the influence of the value of a theorem on a class is independent from the influence of other attributes. This assumption is called "class conditional independence". This assumption was made to simplify engaged calculation and this is why it was named "Naïve", i.e., simple.

This technique calculates the prior probability of the response variable and the conditional probability of other variables. The prior and conditional probabilities of the initial training are calculated. Then, for every test dataset sample, the probability of the occurrence (presence) of each case of response variable is calculated. Afterwards, the response variable with the highest occurrence probability is selected. The probability of test sample for the response variable value is derived from relation 4.3 [22].

$$P\left(v = c_i\right) = P\left(c_i\right) \times \sum_{j=1}^{n} P(a_j = v_j | class = c_i) \qquad (2\text{-}2)$$

Where V, ci, aj and vj are test sample, response variable value, data attribute and the test sample value, respectively.

K-nearest neighbor

This classification technique is called a memory-based technique, since the training samples should be stored in memory during run-time [23-27].

If a is the first sample denoted by (a1, a2,…, an), and b is the second sample denoted by (b1, b2,…, bn), the distance between them is calculated by relation 2-3.

$$\sqrt{\left(a_1 - b_1\right)^2 + \left(a_2 - b_2\right)^2 + \left(a_n - b_n\right)^2} \qquad (2\text{-}3)$$

Support vector machine

Given availability of support vectors, Support Vector Machine (SVM) is the boundary determining the best data classification and separation. In SVM, only those data lying inside support vectors are used as the base data for machine and building a model. This means that this algorithm is not sensitive to other data. It aims to find the best data boundary with the farthest possible distance from all classes (their support vectors). SVM transfers data to a new space with respect to their predetermined classes so that data can be classified and separated linearly (using hyperplanes). Then, it searches for support lines (or support planes in multi-dimensional space) and tries to determine the equation of a straight line that maximizes the distance between each two classes. Each support vector is characterized with an equation describing the boundary line of each class.

Dataset

This study used Cleveland Clinic Foundation dataset known as "Cleveland Clinic Foundation Heart Disease Dataset". This dataset included 13 attributes (Table 1) and 303 samples, 3 of which were incomplete and hence excluded from this study. The continuous values of this dataset were discretized using equal frequency method. This technique classifies continuous values into 5 classes.

Selected data mining techniques were applied on the studied dataset after dataset discretization. 10-fold cross-validation method was used to validate the results. This technique classifies dataset into 10 portions. 9 portions were used for training the algorithm and 1 portion was used for evaluation in each run-time. The process was repeated 10 times. This procedure is helpful specifically in datasets with small number of samples by prevention of over-fitting. Finally, the sensitivity, specificity and accuracy of each method were calculated.

Sensitivity is the ratio of true positives. Specificity is the ratio of true negatives. Accuracy is the ratio of true positives and true negatives combined (relations 2.4 to 2.6).

Sensitivity=True Positive/Positive (2-4)

Specificity=True Negative/Negative (2-5)

Accuracy=(True Positive + True Negative)/(Positive + Negative) (2-6)

Results

This section compares the accuracy, sensitivity and specificity of the employed techniques in terms of their confusion matrices.

Decision tree, Naïve Bayes, K-Nearest Neighbor and Support Vector Machine were applied to the studied dataset. Table 2 shows the sensitivity, specificity and accuracy of these data mining techniques.

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1=male
		0=female
Cp	Discrete	Chest pain type:
		1-typical angina
		2-atypical angina
		3-non-angina pain
		4-asymptomatic
Trestbps	Continuous	Resting blood pressure in (mm Hg)
Chol	Continuous	Serum cholesterol in (mg/dl)
Fbs	Discrete	FBS>120 (mg/dl)
		1=yes
		0=no
Restecg	Discrete	Resting electrocardiographic results:
		0=nrmal
		1-having ST-T wave abnormality
		2=showing probable or defined left ventricular hypertrophy
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise-induced angenia:
		1=yes
		0=no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment:
		1=up sloping
		2=flat
		3=down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that range between 0 and 3
Thal	Discrete	3=normal
		6=fixed defect
		7=reversible defect
Diagnosis	Discrete	Diagnosis classes:
		1=healthy
		2=patient who is subject to possible heart disease

Table 1: Cleveland clinic foundation heart disease dataset indices.

Their accuracy ranges from 79% to 84.33%. According to Table 2, SVM achieved the highest accuracy (84.33%).

SVM and Naïve Bayes achieved the highest accuracy, followed by KNN (k=7 resulted in the best accuracy as compared to other values) and decision tree, respectively. Weka and IBM SPSS Modeler were used to implement data mining techniques.

Identification of influential factors

After implementing the above mentioned techniques and performing related analyses, the following influential indices were identified. Table 3 compares all 13 input attributes in terms of their significance based on the results obtained from each technique.

Accuracy	Specificity	Sensitivity	Method
0.79	0.8148	0.7608	Decision Tree
0.8366	0.8641	0.8043	Naïve Bayes
0.81	0.8395	0.7753	KNN K=7
0.8433	0.895	0.7826	SVM

Table 2: Results of employed data mining techniques.

Attribute	Importance degree				Avg.
	Naïve Bayes	KNN K=7	SVM	Decision Tree	
Age	0.7	0.08	-	0.01	0.04
Sex	-	0.08	0.06	0.05	0.047
Cp	0.11	0.08	0.18	0.24	0.152
Trestbps	0.04	0.08	-	-	0.03
Chol	0.05	0.08	-	-	0.032
Fbs	0.06	0.08	0	-	0.035
Restecg	0.07	0.08	0	-	0.037
Thalach	0.05	0.07	0.09	-	0.052
Exang	-	0.08	0.09	0.08	0.062
Oldpeak	0.09	0.07	0	0.04	0.05
Slope	-	0.07	0.13	0.04	0.06
Ca	0.21	0.08	0.25	0.2	0.185
Thal	0.15	0.08	0.19	0.33	0.187

Table 3: Significance comparison of input indices in each method.

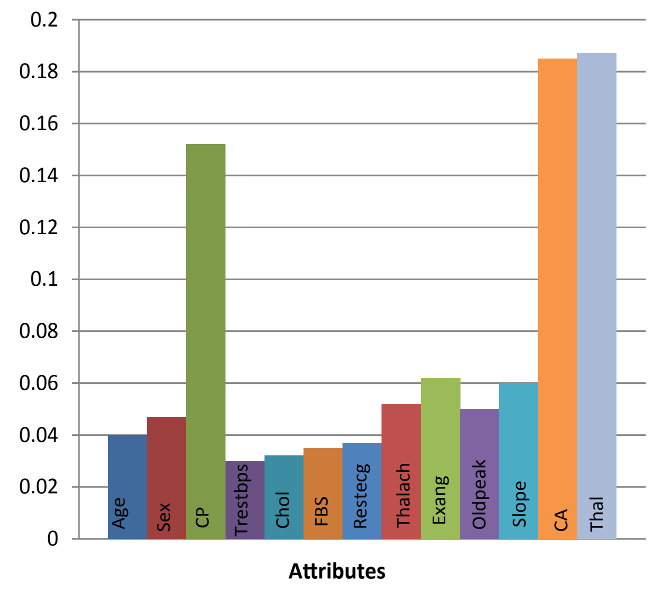


Figure 2: Comparing significance of input indices in each method.

The variables thal (0.187), ca (0.185) and cp (0.152) had the highest significance in all 4 employed techniques respectively, followed by trestbps (0.03) and chol (0.032) as the lowest. Figure 2 compares the attributes in terms of significance.

Conclusion

Heart disease is the leading cause of death across the world. It accounts for 7.2 million deaths, i.e., 12.8% of fatalities in the world. Although cardiovascular diseases have been identified as the leading cause of death in the past decade, they are the most preventable and controllable diseases at the same time. Deaths from cardiovascular diseases show an ever-increasing trend. On the other hand, their early diagnosis plays an important role in improving patients' health status and decreasing fatalities. Therefore, this study aimed to aid physicians to early diagnose such diseases and assess heart disease risk factors in studied individuals. After studying different papers, selecting different data mining techniques and implementing them on the selected dataset, SVM technique achieved the highest accuracy (84.33%). In SVM as well as the other employed techniques, thal, ca and cp were introduced as the most influential indices on average. This technique is expected to be implemented in future on a localized dataset with non-aggressive indices in general. This, in turn, imposes lower costs and complications on patients.

References

1. World Health Organization (2011) The top ten causes of death.
2. World Health Organization (2013) Deaths from coronary heart disease.
3. Center for Disease Control and Prevention (2014) Heart Disease and Family History.
4. Paladugu S (2010) Temporal mining framework for risk reduction and early detection of chronic diseases. University of Missouri-Columbia.
5. Obenshain MK (2004) Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology 25: 690-695.
6. Shillabeer A, Roddick JF (2006) Towards role based hypothesis evaluation for health data mining. Electronic. Journal of Health Informatics 1: 1-9.
7. Porter T, Green B (2009) Identifying Diabetic Patients: A Data Mining Approach.
8. Panzarasa S, Quaglini S, Sacchi L, Cavallini A, Micieli G, et al. (2010) Data mining techniques for analyzing stroke care processes. In the Proc. of the 13th World Congress on Medical Informatics.
9. Li L, Tang H, Wu Z, Gong J, Gruidl M, et al. (2004) Data mining techniques for cancer detection using serum proteomic profiling. Artificial intelligence in medicine 32: 71-83.
10. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications 36: 7675-7680.
11. Lakshmi K, Krishna MV, Kumar SP (2013) Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability. International Journal of Scientific and Research Publications 3: 1-10.
12. Centers for Disease Control and Prevention (2013) Chronic Disease Prevention and Health Promotion. Accessed 27 September 2013, from http://www.cdc.gov/nccdphp/
13. U.S department of health and human services (2005) High Blood Cholesterol What you need to know.
14. Department of Health & Aging AG (2012) Seniors and Aged Care Australia websites have been replaced.
15. Heller RF, Chinn S, Tunstall Pedoe HD, Rose G (1984) How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. British Medical Journal 288: 1409-1411.
16. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, et al. (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97: 1837-1847.

17. Simons LA, Simons J, Friedlander Y, McCallum J, Palaniappan L (2003) Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia 178: 113-116.

18. Shahwan-Akl L (2010) Cardiovascular disease risk factors among adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing 1: 1-7.

19. Helma C, Gottmann E, Kramer S (2000) Knowledge discovery and data mining in toxicology. Statistical Methods in Medical Research 9: 329-358.

20. Quinlan JR (1986) Decision trees and multi-valued attributes. In: Hayes, Michie D (eds.) Machine intelligence. Oxford University Press.

21. Han J, Kamber M (2006) Data Mining Concepts and Techniques: Morgan Kaufmann Publishers.

22. Bramer M (2007) Principles of data mining: Springer.

23. Alpaydin E (1997) Voting over multiple condensed nearest neighbors. Artificial Intelligence Review 11: 115-132.

24. National Center for Chronic Disease Prevention and Health Promotion (2013) Know the facts about heart disease.

25. Rajkumar M, Reena GS (2010) Diagonsis of Heaer Disease using Datamining Algorithm. Global Journal of Computer Science and Technology 10: 38-43.

26. Shouman M (2014) Prototype Development of a Novel Heart Disease Risk Evaluation Tool Using Data Mining Analysis. Zagazig University, Egypt.

27. Tu MC, Shin D, Shin D (2009) Effective Diagnosis of Heart Disease through Bagging Approach. Paper presented at the 2nd International Conference on Biomedical Engineering and Informatics.