Prediction of Heart Disease Using Classification Based Data Mining Techniques

Sujata Joshi and Mydhili K. Nair

Abstract Data Mining is an interesting field of research whose major objective is to find interesting and useful patterns from huge data sets. These patterns can be further used to make important decisions based on the result of the analysis. Healthcare industry today generates huge amount of data on a day to day basis. This data has to be analysed and hidden and meaningful patterns can be discovered. Data mining plays a promising and significant role in this aspect. Data Mining techniques can be used for disease prediction. In this research, the classification based data mining techniques are applied to healthcare data. This research focuses on the prediction of heart disease using three classification techniques namely Decision Trees, Naïve Bayes and K Nearest Neighbour.

Keywords Data mining · Classification technique · Heart disease · Healthcare · Decision tree · Naïve bayes · K-Nearest neighbor · Dataset

1 Introduction

Heart Disease is a class of diseases that involve the heart, the blood vessels or both. The most common causes of heart disease are atherosclerosis and/or hypertension. Atherosclerosis is a condition that develops when a substance called plaque builds up in the walls of the arteries. This buildup narrows the arteries, making it harder for blood to flow through. If a blood clot forms, it can stop the blood flow. This can

S. Joshi (✉)
Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India
e-mail: sujata_msrp@yahoo.com

M.K. Nair
Department of Information Science and Engineering, M. S. Ramaiah Institute of Technology, Bangalore, Karnataka, India
e-mail: mydhili.nair@gmail.com

L.C. Jain et al. (eds.), *Computational Intelligence in Data Mining - Volume 2*,
Smart Innovation, Systems and Technologies 32, DOI 10.1007/978-81-322-2208-8_46

cause a heart attack or stroke. The major risk factors for heart diseases are age, gender, high blood pressure, diabetes mellitus, tobacco smoking, processed meat consumption, excessive alcohol consumption, sugar consumption, family history, obesity, lack of physical activity, psychosocial factors, and air pollution.

Heart disease is the leading cause of deaths worldwide, however since the 1970s, mortality rate due to heart related diseases have declined in many high-income countries. At the same time, heart related deaths and diseases have increased at a fast rate in low and middle-income countries. Although heart disease usually affects older adults, the symptoms may begin in early life, making primary prevention efforts necessary from childhood. Therefore risk factors may be modified by having healthy eating habits, exercising regularly, and avoiding of smoking tobacco.

In today's world, most of the hospitals maintain their patient data in electronic form through some hospital database management system. These systems generate huge amount of data on a daily basis. This data may be in the form of free text, structured as in databases or in the form of images. This data may be used to extract useful information which may be used for decision making. This requirement has led to the use of Knowledge Discovery in Databases (KDD) which is responsible for transforming data of low-level into high-level knowledge for decision making. Data mining which is one of the KDD process aims at finding useful patterns from large datasets. These patterns can be further analyzed and the result can be used for effective decision making and analysis. The various tasks of data mining are classification, clustering, association analysis and outlier detection. In this paper, various data mining classification techniques are applied to healthcare data related to heart diseases. It has helped to determine the best prediction technique in terms of its accuracy and error rate on the specific dataset.

2 Related Work

There has been an increase in the number of people suffering from heart diseases in the recent years [1]. With the advent of information technology and its applications data mining plays a very important and apt role in early detection of diseases. Data mining is extensively used in all fields and healthcare industry in particular [2–6]. In the healthcare industry, the data mining techniques are used for diagnosis of diseases [7], disease prediction [8], and analysis [9]. Data mining techniques can be applied for predicting the outcome of interest. Hence prediction is a very important task. The issues and guidelines of Predictive data mining in clinical medicine is discussed in [10]. Research work [7, 11, 12] related to heart disease diagnosis using data mining techniques is the motivation for this work. Classification based on Gini index is discussed in [13]. The data mining techniques Decision tree, Naïve Bayes and KNN are discussed in [8, 10, 14, 15]. A model based on Combination of Naïve Bayes Classifier and K-Nearest Neighbor is proposed in [16]. A clinical decision support system using association rule mining is discussed in [17]. A prediction system for lung cancer detection is proposed in [18]. A diagnostic tool is proposed

in [19] for skin diseases. In [6, 9], the researchers analyze healthcare data using different data mining techniques. After the extensive literature survey of the dataset, algorithms, methods employed by the authors, results and future work, it is found that there is a lot of scope in discovering efficient methods of medical diagnosis for various diseases and their analysis. This work is an attempt to predict the occurrence of heart diseases using classification data mining techniques namely Decision Tree, Naïve Bayes and K-Nearest Neighbor techniques.

3 Classification

Classification is one of the important data mining tasks. The objective of classification is to assign a class to previously unseen data accurately. Classification consists of two stages:

Stage 1: Model construction
Stage 2: Model usage

Classification creates a model for the attributes of the dataset. A dataset is divided into training set and test set. In the first stage the training set is used to build the classification model using a learning algorithm. In the second stage, the learned model is put into operational use i.e. it is used to validate the test set. If the model performs well, then the model is now ready for prediction.

3.1 Classification Techniques

In this study, the classification techniques, Decision tree, Naïve Bayes and KNN are explored and applied to the dataset.

3.1.1 Decision Trees

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on an attribute, each branch denotes the outcome of test and each leaf node holds the class label. The first node in the tree is the root node. First, an attribute is selected and placed at the root node, and a branch is made for each possible value. This splits up the data set into subsets, one for every value of the attribute. Now repeat the process recursively for each branch, using only those instances that actually reach the branch. When all instances at a node have the same classification, the tree development can be stopped. To select the best split the measures used generally are Gini, Entropy or Classification error.

3.1.2 Naïve Bayes Classifier

Classification based on Bayes Theory is known as Bayesian Classification. Naive Bayes classifier is a statistical based classifier which is based on Bayes Theory. It assumes that attributes are statistically independent. This classifier is based on probabilities.

Given two events A and B, P(A) is prior probability and $P(A|B)$ is posterior probability, then according to Bayes theorem

$$\mathrm{P(A|B)} = \mathrm{P(B/A)P(A)/P(B)} \text{ and } \mathrm{P(B|A)} \text{ is computed as } \mathrm{P(A \cap B)/P(A)}$$

These Bayesian probabilities are used to determine the most likely next event for the given instance given all the training data. Conditional probabilities are determined from the training data.

This classifier yields optimal prediction (given the assumptions). It can also handle discrete or numeric attribute values.

3.1.3 K-Nearest Neighbor

Nearest neighbor method is a instance based classification technique that remembers all the instances. When the new instance is encountered, it uses previous instances as a model and compares it with the new instance. Prediction for the current instance is the one with the most similar previously observed instance. K-NN classifies the instances using the K nearest neighbors. This classifier has faster training rate but is slow when the dataset is large since it has to evaluate all instances.

4 Methodology

4.1 TOOL Used

WEKA [20] Tool (Waikato Environment for Knowledge Analysis), is a set of data mining algorithms and tools which can be used for analysis of data. WEKA is developed in JAVA. WEKA allows analyzing the data sets saved in .arff format using various algorithms. In this study, the Decision tree, Naïve Bayes and K-NN algorithms are applied to heart data set and the results of applying these techniques are shown.

4.2 Data Source

The heart diseases data set from the UCI [21] Learning Repository is used for this study. The heart data set consists of 303 records and 14 attributes. The attributes are listed in Table 1.

Table 1 Attributes of the heart.arff file

No	Attribute	Type
1	age	Real
2	sex	{female, male}
3	cp	{typ_angina, asympt, non_anginal, atyp_angina}
4	trestbps	Real
5	chol	Real
6	restecg	{left_vent_hyper, normal, st_t_wave_abnormality}
7	thalach	real
8	restecg	{left_vent_hyper, normal, st_t_wave_abnormality}
9	exang	{no, yes}
10	oldpeak	real
11	slope	{up, flat, down}
12	ca	real
13	thal	{fixed_defect, normal, reversable_defect}
14	num	{'<50', '>50_1','>50_2', '>50_3', '>50_4'}

4.3 Decision Tree

The decision tree is created by selecting the best split at every node. To select the best attribute for the split, the information gain is computed at each node and the attributes are ranked accordingly. Here the attribute evaluator used is Gain Ratio AttributeEval and the search method used is Ranker method from WEKA Tool. The ranked attributes are listed in Table 2.

Table 2 Attribute ranking based on information gain

Info gain	Rank	Attribute
0.17	12	thal
0.16	13	ca
0.15	9	exang
0.13	8	thalach
0.11	3	cp
0.10	10	oldpeak
0.09	11	slope
0.065	2	sex
0.060	1	age
0.022	7	restecg
0	6	fbs
0	5	chol
0	4	trestbps

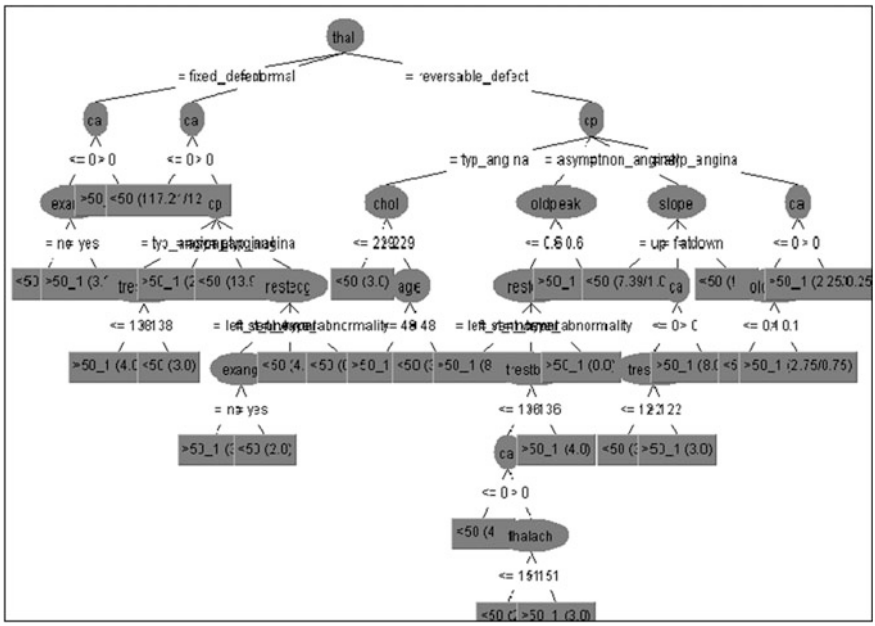


Fig. 1 Decision tree generated using J48 algorithm

Table 3 Results of decision tree algorithm

	No of instances	Percentage (%)
Correctly classified instances	279	92.0792
Incorrectly classified instances	24	7.9208
Total instances	303	

The attributes selected in the order are: **12**, **13**, **9**, **8**, **3**, **10**, **11**, **2**, **1**, **7**, **6**, **5, 4**.

The Decision Tree algorithm J48 is then applied to the heart data set and the decision tree in Fig. 1 is generated. This decision tree can be used for prediction. The results are shown in Table 3.

4.4 Naïve Bayes

The attribute evaluator used is Gain Ratio AttributeEval and the search method used is Ranker method. The ranked attributes are same as in Decision tree. The Naïve Bayes algorithm is applied to the heart data set and the results of few attributes are shown in Table 4.

The results are shown in Table 5.

Table 4 Results of few attributes using Naïve Bayes technique

Attribute	<50 (0.54)	>50_1 (0.45)	>50_2 (0)	>50_3 (0)	>50_4 (0)
cp					
typ_angina	17.0	8.0	1.0	1.0	1.0
asymp	40.0	105.0	1.0	1.0	1.0
non_anginal	70.0	19.0	1.0	1.0	1.0
atyp_angina	42.0	10.0	1.0	1.0	1.0
[total]	169.0	142.0	4.0	4.0	4.0
restecg					
left_vent_hyper	69.0	80.0	1.0	1.0	1.0
Normal	97.0	57.0	1.0	1.0	1.0
st_t_wave_abnormality	2.0	4.0	1.0	1.0	1.0
[total]	168.0	141.0	3.0	3.0	3.0

Table 5 Results of Naïve Bayes technique

	No of instances	Percentage (%)
Correctly classified instances	255	84.1584
Incorrectly classified instances	48	15.8416
Total instances	303	

Table 6 Results of K-nearest neighbor technique

	No of instances	Percentage (%)
Correctly classified instances	303	100
Incorrectly classified instances	0	0
Total instances	303	

4.5 K-Nearest Neighbor

The KNN algorithm is applied to the heart data set and the results are shown in Table 6.

5 Results and Conclusion

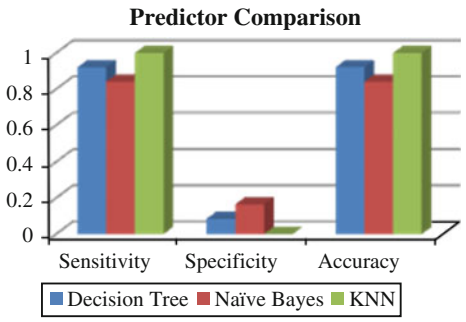
The evaluation measures used are Sensitivity, Specificity and Accuracy

(i) **Sensitivity = TP/P**
(ii) **Specificity = TN/N**
(iii) **Accuracy = (TP + TN)/(P + N)**

Table 7 Summarization of prediction techniques with performance

Prediction technique	Sensitivity	Specificity	Accuracy
Decision tree	0.921	0.085	0.922
Naïve bayes	0.842	0.165	0.842
KNN	1	0	1

Fig. 2 Comparison of prediction techniques



where TP is true positives, TN is true negatives, P and T are actual positives and actual negatives respectively. A good predictor must have high sensitivity, low specificity and high accuracy. The comparisons of these measures with respect to the three prediction techniques are summarized in Table 7.

The experiments are conducted with WEKA tool and the algorithms applied on the heart dataset. The graph in Fig. 2 reveals that sensitivity and accuracy are high and specificity is low. Hence the predictors perform well on operational use. With respect to model creation the results show that KNN has highest accuracy as expected since KNN remembers all the instances. But when used for prediction the Decision Tree performs well when compared to other two methods for the given heart dataset.

References

1. Heart Disease—General Info and Peer reviewed studies: [Online] Available http://www.aristoloft.com
2. Patka, S., et al.: Recent trends and rapid development of applications in data mining. IOSR J. Comput. Sci. (IOSR-JCE) 73–78. e-ISSN: 2278-0661, p-ISSN: 2278-8727 (2014)
3. Tomar, D., Agarwal, S.: A survey of data mining approaches for healthcare. Int. J. Bio-Science and Bio-Technology **5**(5), 241–256 (2013)
4. El-Sappagh, S.H., et al.: Data mining and knowledge discovery: applications, techniques, challenges and process models in healthcare. Int. J. Eng. Res. Appl. (IJERA) **3**(3), 900–906. ISSN: 2248-9622 www.ijera.com (2013)
5. Koh, H.C., Tan, G.: Data mining applications in healthcare. J. Healthc. Inf. Manag. **19**(2), 65 (2011)

6. Obenshain, M.K.: Application of data mining techniques to healthcare data. Infect. Control Hosp. Epidemiol. **25**(8), 690–695 (2004)
7. Shouman, M., Turner, T., Stocker, R.: Using data mining techniques in heart disease diagnosis and treatment. In: Proceedings in Japan–Egypt Conference on Electronics, Communications and Computers, vol. 2, pp. 174–177. IEEE (2012)
8. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inf. **77**(2), 81–97 (2006)
9. Gosain, A.: Analysis of healthcare data using different data mining techniques, IEEE, ISBN: 978-1-4244-4711-4 (2009)
10. Milovic, B., Milovic, M.: Prediction and decision making in health care using data mining. Int. J. Public Health Sci. (IJPHS) **1**(2), 69–78 (2012). ISSN: 2252-8806
11. Melillo, P., De Luca, N., Bracale, M., Pecchia, L.: Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. IEEE J. Biomed Health Inf. **17**(3), 727–733 (2013)
12. Rao, R.B., Krishan, S., Niculescu, R.S.: Data mining for improved cardiac care. ACM SIGKDD Explor. Newsl. **8**(1), 3–10 (2006)
13. Suneetha, N., Hari, V.M.K., Kumar, V.S.: Modified gini index classification: a case study of heart disease dataset. Int. J. Comput. Sci. Eng. **2**(6), 1959–1965 (2010)
14. Han, J., Kamber, M.: Data Mining: Concepts And Techniques. Morgan Kaufmann, San Francisco (2001)
15. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, 4th edn. Pearson Publications, Boston
16. Ferdousy, E.Z., Islam, M.M, Matin, M.A.: Combination of Naïve Bayes classifier and K-nearest neighbor in the classification based predictive models. J. Comput. Inf. Sci. **6**(3), 48–56. ISSN: 1913-8989 (2013)
17. Cheng, C., Chanani, N., Vengopalan, J., Maher, K., Wang, D.: icuARM—an ICU clinical decision support system using association rule mining. IEEE J. Transl. Eng. Health Med. **1** (2013)
18. Krishnaiah, V., Narasimha, G., Chandra, N.S.: Diagnosis of lung cancer prediction system using data mining classification techniques. Int. J. Comput. Sci. Inf. Technol. **4**, 39–45 (2013)
19. Cataloluk, H., Kesler, M.: A diagnostic software tool for skin diseases with basic and weighted K-NN, IEEE. ISBN: 978-1-4673-1448-0/12 (2012)
20. WEKA: Data Mining Machine Learning Software. [Online] Available http://www.cs.waikato.ac.nz/ml/weka/
21. UCI Machine Learning Repository. [Online] Available http://archive.ics.uci.edu/ml/datasets.html
22. Cios, K.J., William Moore, G.: Uniqueness of medical data mining. J. Artif. Intell. Med. **26**(1), 1–24 (2002)