ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
December 2013,
Vol. 6, No. (4):
Pgs. 457-466

Heart Disease Prediction System Using Data Mining Techniques

ABHISHEK TANEJA

Department of Computer Science, S.A. Jain College, Ambala City, India.

(Received: November 15, 2013; Accepted: November 25, 2013)

ABSTRACT

In today's modern world cardiovascular disease is the most lethal one. This disease attacks a person so instantly that it hardly gets any time to get treated with. So diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. A wrong diagnosis by the hospital leads to earn a bad name and loosing reputation. At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India. The purpose of this paper is to develop a cost effective treatment using data mining technologies for facilitating data base decision support system. Almost all the hospitals use some hospital management system to manage healthcare in patients. Unfortunately most of the systems rarely use the huge clinical data where vital information is hidden. As these systems create huge amount of data in varied forms but this data is seldom visited and remain untapped. So, in this direction lots of efforts are required to make intelligent decisions. The diagnosis of this disease using different features or symptoms is a complex activity. In this paper using varied data mining technologies an attempt is made to assist in the diagnosis of the disease in question.

Key words: Data mining, Decision Tree, Neural Network, Naive Bayes, cardiovascular disease.

INTRODUCTION

Data mining has already established as a novel field for exploring hidden patterns in the huge datasets. Medical science is another field where large amount of data is generated using different clinical reports and other patient symptoms. Data mining can also be used heavily for the same purpose in medical datasets also. These explored hidden patterns in medical datasets can be used for clinical diagnosis. However, medical datasets are widely dispersed, heterogeneous, and huge in nature. These datasets need to be organized and integrated with the hospital management systems.

Cardiovascular diseases are one of the highest-flying diseases of the modern world[1]. According to world health organization about more than 12 million deaths occurs worldwide, every year due to heart problems. It is also one of the fatal diseases in India which causes maximum casualties. The diagnosis of this disease is intricate process. It should be diagnosed accurately and correctly. Due to limitation of the potential of the medical experts and their unavailability at certain places put their patients at high risk. Normally, it is diagnosed using intuition of the medical specialist. It would be highly advantageous if the techniques will be integrated with the medical information system. A

computer based information or decision support systems can facilitate accurate diagnosis that's too at reduced cost[1-2]. This integration of varied data mining techniques with the existing medical decision support system requires comparison of different data mining techniques for finding out their suitability for the said job. This paper tries to find out different data mining techniques suitable for the said job.

Literature Review

A model Intelligent Heart Disease Prediction System built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Neural Network was proposed by Palaniappan and Awang, they used a CRISP-DM methodology to build the mining models on a dataset obtained from the Cleveland Heart Disease database[3]. The results demonstrated the strange strength of each of the methodologies in realizing the objectives of the specified mining objectives. Intelligent Heart Disease Prediction System was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease.

Another study experimented on a sample database of patients' records. The Neural Network is tested and trained with 13 input variables such as Age, Blood Pressure, Angiography's report and the like. The supervised network has been recommended for diagnosis of heart diseases[4]. Training was carried out with the aid of back propagation algorithm. Whenever unknown data was fed by the doctor, the system identified the unknown data from comparisons with the trained data and generated a list of probable diseases that the patient is vulnerable to. The success rate for imprecise inputs to retrieve the desired output is closest to 100%.

In another study the problem of identifying constrained association rules for heart disease prediction was studied by[5]. The underlying dataset encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing. Three constraints were introduced to decrease the number of patterns. First one necessitates the attributes to appear on only one side of the rule. The second one segregates attributes into uninteresting groups. The ultimate constraint restricts the number of attributes in a rule[9]. Experiments illustrated that the constraints reduced the number of discovered rules remarkably besides decreasing the running time. Two groups of rules envisaged the presence or absence of heart disease in four specific heart arteries.

In year 2010, a study was conducted for predictive model for the Ischemic Heart Disease (IHD); they applied Back-propagation neural network (BPNN), the Bayesian neural network (BNN), the probabilistic neural network (PNN) and the support vector machine (SVM) to develop classification models for identifying IHD patients on a data obtained from measurements of cardiac magnetic field at 36 locations (6 × 6 matrices) above the torso[6]. The result shows that BPNN and BNN gave the highest classification accuracy of 78.43 %, while RBF kernel SVM gave the lowest classification accuracy of 60.78 %. BNN presented the best sensitivity of 96.55 % and RBF kernel SVM displayed the lowest sensitivity of 41.38 %. Both polynomial kernel SVM and RBF kernel SVM presented the minimum and maximum specificity of 45.45 % and 86.36 %, respectively.

After reviewing the above literatures the researcher was motivated to work on a classification model that is sought to predict heart disease cases based on patterns generated from International Cardiovascular Hospital database[7]. This study is different in two ways from the studies that are presented above, the first one is the data that is used for this study is collected from transthoracic echocardiography report of patients and the second one is the classification models are developed on a much larger dataset. As far as the knowledge of the researcher is concerned this study will be the first of its kind in Ethiopia that applied data mining to predict heart disease[8].

Research Methodology

In this study, to develop a prediction model that can predict heart disease cases based on measurements taken from transthoracic echocardiography examination, and we have used

the Knowledge Discovery in Database (KDD) methodology as described by[2].

To define the problem and determine medical goals, I have thoroughly discussed with medical fraternity particularly cardiologists at PGI, Chandigarh. Since the knowledge gained from the different experts are a high-level description of the problem from the medical point of view, a literature review was carried out and relevant works related to data mining and heart disease have been reviewed to have more knowledge about the domain.

Furthermore, a real time observation of the system was performed to understand the business process of the hospital. A key sub goal in this step is determination of data mining goals and their success criteria. The goals are obtained by translating medical goals into data mining goals.

I have used data collected from PGI, Chandigarh which contains transthoracic echocardiography report of 7,008 patients from the year 2008 to the first quarter of year 2010. Data was collected from various measurements that were taken during the echocardiography examination that also included information of 20 variables. In an effort to reduce the number of variables, then I turned to a domain expert for assistance. The expert selected 15 of the most important variables for inclusion in the dataset.

As the hospital keeps the record of each patient in a separate hard file, therefore that file is converted into a separate Microsoft Word file, in order to integrate the data it was needed to create a database with variables of interest and record the values of each variable into the new database. After recording, the new database now contains 7,339 instances each instance resembling a single file.

The selected data was checked for noise, inconsistency and missing values using distribution frequency while outlier detection was done using box plots. Noises and inconsistencies identified in the dataset were corrected manually, while missing values were replaced with the most probable value determined with regression and outliers were replaced with the mean value of the attributes. All the data cleaning was performed after addressing issues and requirements of the tools selected for the preprocessing phase.

At this stage after consulting with the domain expert a few transformations were implemented on the dataset to make the data more suitable for the data mining algorithms.

The other data transformation like attribute selection was necessary to reduce the number of features a classification algorithm has to examine and reduce errors from irrelevant features. I have used best first search method to select the best attributes from 15 attributes that were available.

In the next step I have selected appropriate data mining technique for developing a predictive model. After thoroughly checking the available algorithms in Weka machine learning software the algorithms Decision Tree, Neural Network and Bayesian Classifier were selected for this study.

To employ the selected classification algorithms four experiments were designed and the experiments were conducted on a full training dataset containing 7,339 instances. In all of the experiments two scenarios were considered, one containing all 15 attributes and the other only 8 selected attributes. 10-Fold Cross Validation was adopted for randomly sampling the training and test data sets. The Weka 3.6.4 machine learning software was used for these purposes.

All the models built were evaluated to see how they fulfill data mining goals. Algorithms were evaluated on the basis of classification accuracy, area under the ROC curve and confusion matrix table.

Experimentation

Keeping in view the goal of this study to predict heart disease using classification techniques, I have used three different supervised machine learning algorithms i.e., Decision Tree Classification, Bayesian Classifier and Neural Network.

Four experiments were conducted for this study and for all experiments two situations were considered, one containing all the 15 attributes and

the other containing 8 selected attributes. With four experiments and eight different situations a total of eight models were developed.

The performances of the models in this study were evaluated using the standard metrics of accuracy, precision, recall and F-measure which were calculated using the predictive classification table, known as Confusion Matrix. ROC area was also used to compare the performances of the classifiers.

In this regard I have conducted four experiments. For all the experiments two settings was done, one containing all the 15 variables and the other containing 8 chosen variables. All the experiments were done on a full training dataset containing all the instances and cross validation was used for randomly sampling the training and test sets.

The first experiment was designed to evaluate the performance of a J48 classifier Unpruned tree in predicting heart disease and to investigate the effect of attribute selection on the performance of the model. In this experiment two situations were considered, one containing all 15 attributes and the other containing the selected 8 attributes.

On the first scenario the algorithm was run on a full training set containing 7,339 instances with 15 attributes. It took 0.89 second to build the model and the model generated a tree with a size of 473 and 323 leaves.

On the second scenario the algorithm was run on a full training set containing 7,339 instances with selected 8 attributes. It took 0.36 second to build the model and the model generated smaller and less complex tree with a size of 126 and 71 leaves making it less complex and faster than the experiment conducted on all attributes.

In the first experiment I evaluated the performance of J48 classifier unpruned tree in predicting heart disease. The result of which is given in table 1.1 below and their detail performance measures used is depicted in table 1.2.

In the first case the algorithm was run on a full training set containing 7,339 instances with 15 attributes. It took 0.89 second to build the model and the model generated a tree with a size of 473 and 323 leaves.

In the second case situation the algorithm was run on a full training set containing 7,339 instances with selected 8 attributes. It took 0.36 second to build the model and the model generated smaller and less complex tree with a size of 126 and 71 leaves making it less complex and faster than the experiment conducted on all attributes.

Table 1: Confusion Matrix for Experiment I

Model	**Confusion Matrix**		
	Yes (Predicted)	No (Predicted)	Actual
J48 unpruned with all attributes	2,841	206	Yes
	213	4,079	No
J48 unpruned with selected attributes	Yes (Predicted)	No (Predicted)	Actual
	2,874	172	Yes
	157	4,135	No

The results of this experiment showed that a J48 unpruned decision tree algorithm is highly capable in predicting heart disease cases. Furthermore, the results showed the impact of attribute selection on classification accuracy, Decision tree size and model complexity.

Table 2: Detailed Performance Measures for Experiment 1

Model	Accuracy	TP Rate	TN Rate	Precision	F-Measure	ROC Area
J48 unpruned with all attributes	94.29%	0.932	0.95	0.943	0.943	0.942
J48 unpruned with selected attributes	95.52%	0.944	0.963	0.955	0.955	0.965

Table 3: Confusion Matrixes for Experiment 2

Model	Confusion Matrix		
	Yes (Predicted)	No (Predicted)	Actual
J48 unpruned with all attributes	2,872	174	Yes
	162	4,130	No
J48 unpruned with selected attributes	Yes (Predicted)	No (Predicted)	Actual
	2,892	155	Yes
	171	4,121	No

Second experiment was designed to find out the performance of a J48 classifier pruned tree in predicting heart diseases.

The result of this experiment is given below in the table1.3 and detailed performance measures used is given in table 1.4.

The model correctly identified 2,892 patients out of 3047 patients who had heart disease and the remaining 155 were identified incorrectly to be free from the disease while they actually had the disease. Regarding to Precision score of the model, 94.4% of patients labeled as belonging to class Yes does indeed belong to class Yes while 96.4% of patients labeled as belonging to class No does indeed belong to class No.

The third experiment was designed to evaluate the performance of Naïve Bayes Classifier in predicting heart disease. In this experiment two scenarios were considered, one containing all 15 attributes and the other containing the selected 8 attributes. The intention here is to investigate the effect of attribute selection on the performance of the models.

Table 4: Detailed Performance Measures for Experiment 2

Model	Accuracy	TP Rate	TN Rate	Precision	F-Measure	ROC Area
J48 unpruned with all attributes	95.41%	0.943	0.962	0.954	0.954	0.964
J48 unpruned with selected attributes	95.56%	0.949	0.96	0.956	0.955	0.965

In the third experiment I tried to evaluate the performance of Naïve Bayes Classifier in predicting heart disease. Again as in previous two experiments two situations were considered, one with all the variables i.e., 15 variables and the other containing the selected variables i.e., 8 variables. The results of this experiments is given below in table 1.5 and detailed performance measures used is given in table 1.6.

Table 5: Confusion Matrixes for Experiment 3

Model	Confusion Matrix		
	Yes (Predicted)	No (Predicted)	Actual
Naive Bayes with all attributes	2,635	412	Yes
	178	4,114	No
Naive Bayes with selected attributes	Yes (Predicted)	No (Predicted)	Actual
	2,654	393	Yes
	163	4,129	No

The performance of Naïve Bayes model was better on the selected attributes. The classification accuracy increased to92.42% from 91.96%. And also, the execution time decreased by half compared to the model built on all 15 attributes.

Table 6: Detailed Performance Measures for Experiment 3

Model	Accuracy	TP Rate	TN Rate	Precision	F-Measure	ROC Area
Naive Bayes with all attributes	91.96%	0.865	0.959	0.92	0.919	0.97
Naive Bayes with selected attributes	92.42%	0.871	0.962	0.925	0.924	0.972

In the fourth experiment I explored the ability of Neural Network in predicting heart disease. From Neural Network Algorithms Multilayer Perception was selected to conduct the experiment. As in the previous cases two scenarios were considered, one containing all 15 attributes and the other containing the selected 8 attributes. The results of this experiments is given below in table 1.7 and detailed performance measures used is given in table 1.8.

Results showed that Neural Network model performed better on the selected attributes compared to the whole set of attributes. Classification accuracy increased to 94.85% from 93.83% also, the execution time decreased significantly to 34.14 seconds from 158.94 seconds.

For comparing the models and selecting the best model I have compared using different

Table 7: Confusion Matrixes for Experiment 4

Model	Confusion Matrix		
	Yes (Predicted)	No (Predicted)	Actual
Neural Network with all attributes	2,789	258	Yes
	195	4,097	No
Neural Network with selected attributes	Yes (Predicted)	No (Predicted)	Actual
	2,841	206	Yes
	172	4,120	No

performance measures like accuracy, TN Rate, TP Rate, F-Measure, ROC Area and execution time (time taken to build the model).

As presented on Table 1.9 all classification algorithms performed nearly equally well with a remarkable accuracy of up to 95.56% while the lowest accuracy score is 91.96%. A Pruned J48 tree classifier which was implemented on selected attributes achieved the highest accuracy (95.55%) while an unpruned J48 tree classifier which was implemented on selected attributes came out to be a close second with classification accuracy of 95.51%. On the other hand, Naïve Bayes classifier implemented on both selected attributes and the whole set of attributes scored the lowest classification accuracy which are 92.42% and 91.96% respectively.

Table 8: Detailed Performance Measures for Experiment 4

Model	Accuracy	TP Rate	TN Rate	Precision	F-Measure	ROC Area
Neural Network with all attributes	93.83%	0.915	0.955	0.938	0.938	0.969
Neural Network with selected attributes	94.85%	0.932	0.96	0.948	0.948	0.974

As presented on Table 1.9 all classification algorithms performed nearly equally well with a remarkable accuracy of up to 95.56% while the lowest accuracy score is 91.96%. A Pruned J48 tree classifier which was implemented on selected attributes achieved the highest accuracy (95.55%) while an unpruned J48 tree classifier which was implemented on selected attributes came out to be a close second with classification accuracy of 95.51%. On the other hand, Naïve Bayes classifier implemented on both selected attributes and the whole set of attributes scored the lowest classification accuracy which are 92.42% and 91.96% respectively.

The other performance measures used to compare the results were TP Rate (Sensitivity) and TN Rate (Specificity). Here again all the models scored astonishingly well with a tight difference in performance. The TP Rate and TN Rate were (TP Rate, TN Rate) = (0.932, 0.95), (0.943, 0.962), (0.865, 0.959), (0.915, 0.955), (0.944, 0.963), (0.949, 0.96), (0.891, 0.962), (0.932, 0.96) for J48 unpruned with all attributes, J48 pruned with all attributes, Naïve Bayes with all attributes, Neural Network with all attributes, J48 unpruned with selected attributes, J48 pruned with selected attributes, Naïve Bayes with selected attributes, Neural Network with selected attributes, respectively.

Table 9: Summarizing performance of various models

Model	Accuracy	TP Rate	TN Rate	Precision	F-Measure	ROC Area
J48 unpruned with all attributes	94.29 %	0.932	0.95	0.943	0.942	0.89
J48 pruned with all attributes	95.41 %	0.943	0.962	0.954	0.964	1.05
J48 unpruned with selected attributes	95.52 %	0.944	0.963	0.955	0.965	0.36
J48 pruned with selected attributes	95.56 %	0.949	0.96	0.955	0.965	0.41
Naive Bayes with all attributes	91.96 %	0.865	0.959	0.919	0.97	0.11
Naive Bayes with selected attributes	92.42 %	0.871	0.962	0.924	0.972	0.05
Neural Network with all attributes	93.83 %	0.915	0.955	0.938	0.969	158.94
Neural Network with selected attributes	94.85 %	0.932	0.96	0.948	0.974	34.14

A model built from J48 pruned with selected attributes scored the highest TP Rate while the model from Naïve Bayes with all attributes scored the lowest. It was easier for the J48 pruned with selected attributes model to identify negative cases correctly compared to the other models in contrast a model built from Neural Network with all attributes straggled a little bit to identify negative case correctly compared to the others.

One important thing observed here was that all the models were better in predicting negative cases compared to the positive ones.

In regard of the ROC Area, looking the area under the curve (AUC) as an indicator for the quality of separation, Table 1.9 confirms neural networks and Naïve Bayes classifiers were the most accurate classifiers. A neural network classifier implemented on selected attributes achieving ROC Area closer to the 'perfect classification' point than the result set from the other experiments.

Based on the time taken to build the models the two experiments implemented with the Naïve Bayes classifier took the shortest time span to build the models whereas, the experiments conducted with Neural Networks took the longest time and the J48 classifiers fall in between those two algorithms.

J48 classifier outperformed the other algorithms by achieving the highest accuracy, TP Rate, TN Rate, and F-Measure values, whereas, the Neural Network classifier achieved the highest ROC Area value and the Naïve Bayes classifier achieved the fastest execution time.

After the comparison of the models was performed the next step was selecting the best model based on those comparisons and to do so it is essential to see things from the clinician view.

Since heart disease is a fatal disease a clinician may prefer to keep the number of false positives low keeping true positives high, but it is still undesirable to tell a healthy patient that he or she is sick. Early diagnosis of a disease is a key factor for a successful treatment, therefore the classification models are expected to perform well at discovering positive instances, and when selecting the best model the emphasis is more on TP Rate. Overall accuracy, then, is not the spirit of classification for this study the spirit is to identify patients with heart disease accurately as much as possible. Ideally the TP Rate is expected to be as close to 1 as is reasonably possible. In other words one should be willing to sacrifice accuracy of negative classifications in exchange for improving the accuracy of positive classifications.

Based on this assumption from the Decision Tree algorithm the J48 classifier implemented on selected attributes is selected as the best predictive model for this study.

The experimental results have shown that, in general, J48 Decision Tree algorithm outperformed Naïve Bayes classifier and Neural Networks in the domain of predicting heart disease cases. One possible explanation for superiority of J48 classifier over Neural Network and Naïve Bayes classifier is the nature of the dataset used in this study. Decision Tree Algorithms tend to perform better on simple datasets and this leads to a conclusion that the classification problem presented by the dataset is a simpler one.

Conclusion and Future Work

In this study, the aim was to design a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography.

Data collected from PGI, Chandigarh from the year 2008 to 2011 containing 7,339 instances was selected and preprocessed for this study. The models were built on the preprocessed Transthoracic Echocardiography dataset with three different supervised machine learning algorithms i.e. J48 Classifier, Naïve Bayes and Multilayer Perception using Weka 3.6.4 machine learning software.

The performances of the models were evaluated using the standard metrics of accuracy, precision, recall and F-measure. 10-Fold Cross Validation was adopted for randomly sampling

the training and test data samples. All eight models performed well in predicting heart disease cases. The most effective model to predict patients with heart disease appears to be a J48 classifier implemented on selected attributes with a classification accuracy of 95.56%.

Three data mining goals were defined based on the medical problems. The goals were evaluated against the selected model and the selected model built with J48 Decision Tree Algorithm successfully met all the three data mining goals.

Significant rules that are useful for predicting the presence of heart disease were extracted from the dataset. The domain expert confirmed that most of the rules generated are important in interpretation of echocardiography examinations.

From a total of 15 attributes that were available, 8 attributes that are highly relevant in predicting heart disease from Transthoracic Echocardiography dataset were selected.

Heart disease is a fatal disease by its nature and misdiagnosis of this disease can cause serious, even life threatening complications such as cardiac arrest and death. The best model selected for predicting heart disease could not exceed a classification accuracy of 95.56% and still much remains to fill the gap of 4.44% misclassified cases.

This study showed that data mining techniques can be used efficiently to model and predict heart disease cases. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior cardiologists to screen out patients who have a high probability of having the disease and transfer those patients to senior cardiologists for further analysis.

Most of the experiments conducted in this study were implemented with default parameters of the algorithms, further investigations should be performed with different parameter settings to enhance and expand the capabilities of the prediction models. In addition, the Neural Network and Naïve Bayes classification algorithms should be tested thoroughly. Missing values, noisy data, inconsistencies, and outliers presented a challenge in the data mining process. Therefore, statistical and machine learning approaches should be applied to control the quality of the data. Furthermore, keeping each patient's echocardiography examination result as a single file has made it difficult to apply any kind of data mining technique. Creating a database for echocardiography examination results and other examination results would be helpful for searching, retrieving and minimizing memory space. Currently, patient history and knowledge used for interpreting the echocardiography results are not stated on the final report. The researcher believes that stating this hidden knowledge can have a positive impact on researches that will be conducted in the future. The echocardiography offers two-dimensional images during examination. Unfortunately, the images generated from each examination are not stored by the hospital instead; they are discarded as soon as the examination is over. The hospital should find a way to store the image so that they can be used to extract relevant information related to the disease using intelligent image recognition systems.

As a future work, the researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

REFERENCES

1. David L. Olson and Dursun, D., *Advanced Data Mining Techniques*. Springer-Verlag Berlin Heidelberg (2008).
2. Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy., *Advances in Knowledge Discovery and Data Mining*, (AKDDM), AAAI/ MIT Press, Massachusetts (1996).
3. Palaniappan, S., Awang, R., *Intelligent Heart*

Disease Prediction System Using Data Mining Techniques, IJCSNS *International Journal of Computer Science and Network Security*. **8**(8): 343-350 (2008).

4. Guru, N., Anil, D., Navin, R., *Decision Support System For Heart Disease Diagnosis Using Neural Network*. Delhi Business Review. **8**(1): (2007).
5. Carlos, O., *Improving Heart Disease Prediction Using Constrained Association Rules*, Seminar Presentation at University of Tokyo (2004).
6. Kangwanariyakul, Y., Chanin, N., Tanawut, T., Thanakorn, N., *Data Mining of Magnetocardiograms for Prediction of Ischemic Heart Disease*. *EXCLI Journal.* **33**(9): 82-95 (2010).
7. Giudici, P., *Applied Data Mining Statistical Methods for Business and Industry*. John Wiley & Sons Ltd, Chichester , England (2003).
8. WHO., *Fact Sheet: The Top Ten Causes of Death*. World Health Organization. Geneva (2006).
9. Han, J. and Kamber, M., *Data Mining: Concepts and Techniques.* Second Edition, Morgan Kaufmann Publishers, San Francisco (2006).