Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases

Moloud Abdar¹, Sharareh R. Niakan Kalhori², Tole Sutikno³, Imam Much Ibnu Subroto⁴, Goli Arji⁵

¹Undergraduate of Computer Engineering, Department of Engineering, Damghan University, Iran
²Department of Health Information Management, Tehran University of Medical Sciences, Iran
³Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
⁴Department of Informatics Engineering, Universitas Islam Sultan Agung, Semarang, Indonesia
⁵Health Information Management, Tehran University of Medical Sciences, Iran

ABSTRACT

Heart diseases are among the nation’s leading cause of mortality and morbidity. Data mining techniques can predict the likelihood of patients getting a heart disease. The purpose of this study is comparison of different data mining algorithms on prediction of heart diseases. This work applied and compared data mining techniques to predict the risk of heart diseases. After feature analysis, models by five algorithms including C5.0, Neural Network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) and Logistic Regression, developed and validated. C5.0 Decision tree has been able to build a model with greatest accuracy 93.02%. KNN, SVM, Neural Network have been 88.37%, 86.05% and 80.23% respectively. Produced results of decision tree can be simply interpretable and applicable; their rules can be understood easily by different clinical practitioner.

Keyword:
C5.0 Algorithm
Data Mining
Heart Disease
Neural Network

1. INTRODUCTION

According to the latest statistics from the World Health Organization (WHO), heart diseases have a great deal of attention in medical research due to its impact on human health [1]. Cardiovascular disease is the number one cause of death in industrialized countries and not only have a major impact on individuals and their quality of life in general, but also on public health costs and the countries’ economies. Diagnosis of heart disease was more costly decision in diagnosis. Artificial Intelligence (AI) techniques were used vastly in medical diagnosis. With the advancement of science, the volume of accumulated data in various fields has been increased that it is well known the explosion of information [2]. When analyzing the accumulated data they could reveal their hidden useful information. By performing data mining, which is a new science, we able to extract the hidden knowledge of the data. Performing data mining reveals useful relationship existed among data, and this rule can apply for right decision making [3], [4]. Classification is one of the subdivisions of data mining, which acts in accordance with If-Then rule. Its purpose is to predict a variable based on other features that are known as predictors. Neural Network, support vector machine and decision Tree are different form of classification algorithms [5-9]. The purpose of this study is comparison of different machine learning algorithm on prediction of heart diseases.

This section summarises various technical articles on KDD process and data mining classification techniques applied on heart diseases datasets:
Ram Bilas Pachori and his colleagues [10] have been studying and diagnosing heart disease using tunable-Q wavelet obtained from heart rate signals. Since manual data entry occurs with errors and also it is time consuming, Tunable-Q Wavelet Transform (TQWT) method is recommended in the present study. Using the least squares support vector machine (LS-SVM), they have reported the accuracy of 96.8%, sensitivity equal to 100%, and specificity of 93.7%.

Another study conducted by Yongqiang Lyu et al. [11] has been based on an evaluation model of coronary artery disease by using data mining algorithm. In this research a new dynamic model, which makes it possible to assess lifetime, suggests linear time-invariant approach to assess CHD. The model result based on SYNTAX scores indicates a 5% possible error [12] in this study they have used J4.8 Decision tree method, and the reported precision was 84.1 percent.

In another study using genetic algorithm, SVM and SSVM conducted by Sumit Bhatia et al. [13] in classification of cardiac patients the features have been selected by genetic algorithm to help the SSVM in the best mode of input selection, the obtained precision is 72.55%, while the precision obtained by GA-SSVM has improved the result and its precision equals to 90.57%. Peter C. Austin and colleagues [14] discuss heart malfunctions in their paper. The associated physicians have divided the patients into two groups of "with" and "without" disease. They have found that the use of decision tree in data mining will have better results than regression model. Using MV5, Saba Bashir et al. [15] applied MV5 algorithm and its precision was 88.52%.

Another research done by Jasmine Nahar et al. [16] for finding relationship between heart disease risk factors in men and women. It refers to the fact that coronary heart disease risk in women is less than men. Doing exercise men and women can easily overcome their chest pain. One of the extracted points in this paper introduces "Rest ECG" in both forms of normal and hyper, and "Slope being flat" is introduced as a risk factor. However, the research result indicate that Rest ECG for men is considered a risk factor only in its hyper form. The study concludes that Rest ECG should be considered as important factor to predict heart disease in women. The research techniques including Apriori, Predictive Apriori and Tertius have compared to each other and precision of predictive Apriori was 90%. Kyle Walker et.al [17] note that heart disease is the principal cause of death in America, Texas. Therefore, the performed a study on different areas of Texas using cluster analysis and result show that factors such as poor hygiene and economic deprivation and other conditions affect the outbreak of disease.

In the paper presented by K. Rajeswari and colleagues [18], they study the heart disease using Neural Network. They have studied the influence of feature selection for neural network algorithm in identifying patients with Ischemic heart disease. 12 features have been used in the paper. The result of their study shows that when all the features(attributes)are applied, the precision rate in training mode 89.4% and in test mode is 82.2%. An interesting point in the conclusion is that any reduction in features entry causes the precision decrease in both training and test modes. AV Senthil Kumar [19] applied fuzzy mechanism on cardiac patients The calculated precision in this paper was 94.11%. Some examples of research done on cardiac patients with different techniques have briefly mentioned below.

2. RESEARCH METHOD

The present study conducted by using data from the University of California, Irvine (UCI). This data includes 13 features classified into 2 classes of "with" and "without" heart disease. After feature analysis, models by five algorithms including decision tree, neural network, support vector machine and k-nearest neighborhood developed and validated.

2.1. C5.0 Algorithm

C5.0 algorithm developed from C4.5 algorithm is one of the most important and widely used algorithms in data mining. C4.5 itself is the extended form of ID3 algorithm. C5.0 has the ability to be applied for classifying as a decision tree or a set of rules. Because of the understandability of their rules set, they are preferred in many applications. The strength of the algorithm is in handling missing values or its large number of entries, as well as the fact that less time is necessary to learn it [20-23].

If S is training set and X contains n attributes so that the set S is divided into N sub categories: The algorithm to test the features makes use of element is called the gain ratio [24].

The number of samples in the S is displayed in (S1, S2, S3,...Sn). For calculating the number of samples that belong to Ci (the value Parameter i is [i = 1, 2, 3, ..., N]) is used in the following formula: freq(Ci,S). Also for calculate an instance belonging the Ci is used to the formula: freq(Ci,S)/|S|
Training set can be calculated according to the formula:

\[ info(S) = - \sum_{n=1}^{N} \left( \frac{freq(Ci,S)}{|S|} \log_2 \frac{freq(Ci,S)}{|S|} \right) \]

That \( info(S) \) includes information can be identified by all the samples in \( S \). After the division of \( S \) to all its subsets, Gain ratio is calculated as follows:

\[ info_x(S) = - \frac{|S|}{N} \sum \frac{|S_i|}{|S|} \times info(S_i) \]

\[ gain(X) = info(S) - info_x(S) \]

\[ Split Info(X) = - \sum_{i=1}^{N} \left( \frac{|S_i|}{|S|} \times log_2 \frac{|S_i|}{|S|} \right) \]

\[ Gain \ ratio = \frac{\Delta info}{Split \ Info} \]

6. Specificity = \( \frac{TN}{FP + TN} \)

7. Sensitivity = \( \frac{TP}{TP + FN} \)

8. Precision = \( \frac{TP}{TP + FP} \)

9. Accuracy = \( \frac{TP + TN}{TP + FN + FP + TN} \)

2.2. SVM Algorithm

Support Vector Machine (SVM) is a regulatory algorithm introduced by Vapnik in 1995. The base of the algorithm is using the precision to generalize the errors. The algorithm makes "hyperplane" and divides the data into classes so that all samples belonging to one class will be categorized on one side and the rest on the other side. Linear SVM Classifier is defined for the SVM classifying task, and dividing them occurs provided that the chosen line involves the most marginalized sure [13], [25].

2.3. KNN Algorithm

K-nearest neighbor algorithm is a method for classification based on similarity to other cases. Those close to others, are called a "neighbor". When a case is new, its distance from each of the cases in the model is calculated. Applying this classification, specifies the case as being the nearest neighbor, which is the most similar. Therefore, it puts the case into the group that contains the nearest neighbors. The algorithm is also able to calculate values continuously for a target. In this situation, the average or the median target value of the nearest neighbor is used to obtain the predicted value of new case [26].

2.4. Neural Network Algorithm

Artificial Neural Network is a data processing algorithm, originated from human brain. The system includes a large number of tiny processors to handle data processing. The processors act in the form of an interconnected network parallel to each other to solve a problem. Using programming knowledge, in this networks a data structure is designed that can act as neurons. This data structure is called the neuron [27-30].
2.5. Accuracy Measurement

In order to evaluate the prediction rate, there are several indices such as specificity, sensitivity, precision, and accuracy to assess the models’ validity. These indices (equation 6-9) are calculated by the confusion matrix (Figure 1). This matrix is a useful tool for analyzing the performance of classification method in data diagnosis or observations of various categories. The ideal state, most parts of the relevant data with the observations should be located on the main diagonal of the matrix, and the remaining values of the matrix are zero or near zero [31], [32].

\[ FN = \text{The number of positively labeled data, which falsely have been classified as } \text{"Negative"}. \]
\[ TN = \text{The number of negatively labeled data, which have been classified as } \text{"Correct"}. \]
\[ TP = \text{The number of positively labeled data, which have been classified as } \text{"Positive"}. \]
\[ FP = \text{The number of negatively labeled data, which falsely have been classified as } \text{"Positive"}. \]

By means of logestic regression variables which are significantly correlated with target variable are selected as predictor (P<=0.05). they are presented an defined in Table 2.

2.6. Data Set

In this study 270 record with 13 features has been used [33]. Patients’ attributions applied for modeling, their definitions and their range of values presented in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Definition</th>
<th>Categories of Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age of patient</td>
<td>[29-77]</td>
</tr>
<tr>
<td>Sex</td>
<td>Gender of patient</td>
<td>(1 = male; 0 = female)</td>
</tr>
<tr>
<td>CP</td>
<td>chest pain type</td>
<td>[1-4]</td>
</tr>
<tr>
<td>RBP</td>
<td>resting blood pressure</td>
<td>[94-200]</td>
</tr>
<tr>
<td>SC</td>
<td>serum cholesterol in mg/dl</td>
<td>[126-564]</td>
</tr>
<tr>
<td>FBS</td>
<td>fasting blood sugar &gt; 120 mg/dl</td>
<td>[0-1]</td>
</tr>
<tr>
<td>RER</td>
<td>resting electrocardiographic results</td>
<td>[0-2]</td>
</tr>
<tr>
<td>MHRA</td>
<td>maximum heart rate achieved</td>
<td>[71-202]</td>
</tr>
<tr>
<td>EIA</td>
<td>exercise induced angina</td>
<td>[0-1]</td>
</tr>
<tr>
<td>Oldpeak</td>
<td>ST depression induced by exercise relative to rest</td>
<td>[0-6.2]</td>
</tr>
<tr>
<td>Slope</td>
<td>the slope of the peak exercise ST segment</td>
<td>[1-3]</td>
</tr>
<tr>
<td>NUM</td>
<td>number of major vessels (0-3) colored by flourosopy</td>
<td>[0-3]</td>
</tr>
<tr>
<td>Thal</td>
<td>Normal, fixed defect, reversible defect</td>
<td>[3, 6, 7]</td>
</tr>
<tr>
<td>Variable to be predicted</td>
<td>Class of Heart Disease</td>
<td>Absence (1) or presence (2) of heart disease</td>
</tr>
</tbody>
</table>

![Figure 1. Confusion matrix](image)

By means of logestic regression variables which are significantly correlated with target variable are selected as predictor (P<=0.05). they are presented an defined in Table 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Definition</th>
<th>Categories of Values</th>
<th>B</th>
<th>Wald</th>
<th>Sig</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Gender</td>
<td>1 = male; 0 = female</td>
<td>1.104</td>
<td>6.337</td>
<td>0.012</td>
<td>3.018</td>
</tr>
<tr>
<td>CP</td>
<td>chest pain type</td>
<td>[1-4]</td>
<td>0.731</td>
<td>13.648</td>
<td>0.000</td>
<td>2.077</td>
</tr>
<tr>
<td>RBP</td>
<td>resting blood pressure</td>
<td>[94-200]</td>
<td>0.023</td>
<td>5.238</td>
<td>0.022</td>
<td>1.023</td>
</tr>
<tr>
<td>EIA</td>
<td>exercise induced angina</td>
<td>[0-1]</td>
<td>1.236</td>
<td>10.182</td>
<td>0.001</td>
<td>3.442</td>
</tr>
<tr>
<td>NUM</td>
<td>number of major vessels (0-3) colored by flourosopy</td>
<td>[0-3]</td>
<td>1.133</td>
<td>25.224</td>
<td>0.000</td>
<td>3.106</td>
</tr>
<tr>
<td>Thal</td>
<td>Normal, fixed defect, reversible defect</td>
<td>[3, 6, 7]</td>
<td>0.397</td>
<td>16.848</td>
<td>0.000</td>
<td>1.488</td>
</tr>
</tbody>
</table>
3. RESULTS AND ANALYSIS

This section presents the experimental results and analysis done for this study. In this work, four classifiers including C5.0, SVM, KNN and Neural Network are conducted. Data divided into trainset and testset (70% and 30% respectively). The training set is used to build the classifier and test set used to validate it. Model development is conducted in two main steps including model fitness and model accuracy. To calculate the model fitness criteria we used the data of training set; however, to compute the model accuracy measurements, data of testing set is applied which is merely much more valuable to judge about our models accuracy. Related results of these experiments are demonstrated in Table 3.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>Training Accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5.0</td>
<td>89.62 %</td>
<td>84.61 %</td>
<td>85.71 %</td>
<td>87.50 %</td>
<td>90.90 %</td>
<td>95.23 %</td>
<td>90.90 %</td>
<td>93.02%</td>
</tr>
<tr>
<td>SVM</td>
<td>84.90 %</td>
<td>79.48 %</td>
<td>79.48 %</td>
<td>82.61 %</td>
<td>90.90 %</td>
<td>80.95 %</td>
<td>89.47 %</td>
<td>86.05%</td>
</tr>
<tr>
<td>KNN</td>
<td>91.50 %</td>
<td>79.48 %</td>
<td>87.32 %</td>
<td>86.41 %</td>
<td>88.63 %</td>
<td>88.09 %</td>
<td>88.09 %</td>
<td>88.37%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>91.50 %</td>
<td>78.20 %</td>
<td>87.14 %</td>
<td>85.87 %</td>
<td>86.36 %</td>
<td>73.80 %</td>
<td>83.78 %</td>
<td>80.23%</td>
</tr>
</tbody>
</table>

C5.0 Decision tree has been able to build a model with greatest accuracy since the model prediction accuracy is 93.02%. Model accuracies obtained from other classifiers are different as this value for KNN, SVM, Neural network have been 88.37%, 86.05% and 80.23% respectively. By analyzing the variables importance in C5.0 model we find that attention to features such as Thal, CP and Slope are so important in prediction of heart diseases (Figure 2).

![Figure 2. variable importance for heart diseases prediction based on C5.0 model](image)

Figures 3 and 4 are comparative ROC curves based on risk of heart diseases. This figures show two ROC curve for logistic regression and C5.0 decision tree C5.0 has outperformed than logistic regression with area under curve (AUC) 0.869. AUC for logistic regression was 0.835. Overall, these results of area under curve reveals better performance of C4.5 decision tree classification algorithm.
In a study conducted to comparing between data mining tools for heart diseases data set in [34] and [35] variable like blood pressure, blood sugar, age and sex showed a significant association with heart diseases. The study conducted by Jasmine Nahar and her colleagues [16] also pointed out that sex was highly important in predicting heart disease, whereas in this study features such as resting blood pressure, sex, chest pain type, exercise induced angina and number of major vessels played a major role. In a paper Zahra Alizadeh Sani et al [36] have used the C4.5 and Bagging algorithms to diagnosing coronary heart disease. For C4.5 algorithms have reported the best accuracy rate. K. Rajeswari et al. [18] applied neural network on ischemic heart disease that the accuracy obtained for training and testing was 89.4 % and 82.2 % respectively. T. John Peter and K. Somasundaram [37] have been used hybrid attribute selection method for prediction of heart disease. The accuracy obtained by this model was 83.62 %. Kemal Polat and Salih Gunes [38] by use of C4.5 decision tree algorithm obtained 92.59 % accuracy.

4. CONCLUSION

In this study, KNN, SVM, C5.0, Logistic Regression and Neural Network were implemented on UCI dataset. Based on investigated methods, decision tree has achieved the best performance. There are different issues that influence the performance of applied models including type of problem and type of input data (discrete or continuous), due to the fact that dataset mainly was discrete, decision tree able to handle numerical data. Because output variable labeled with two classes; ‘with’ and ‘without’ heart diseases, decision tree yielded better performance than other algorithms. Decision trees are able to generate understandable rules and can perform classification without requiring much computation and clearly indicate that which fields are most important for prediction or classification.

REFERENCE


IJECE Vol. 5, No. 6, December 2015 : 1569 – 1576


BIOGRAPHIES OF AUTHORS

Moloud Abdar received his Undergraduate (Bachelor) degree in Computer Engineering (Software Engineering) from the University of Damghan, Iran in 2015. He has more than 7 conference and journal papers about the Data Mining. Currently, his research interests include data mining, web and text mining, artificial intelligence and image processing.

Sharareh R. Niakan Kalahari holds Ph.D. of medical informatics. She was graduated from the university of Manchester in 2011. Currently, she is Associated Professor in Department of Health Information Management at Tehran University of Medical Sciences (TUMS), Iran. Her main interest is machine learning and optimization in medical domain.

Tole Sutikno is an Associated Professor in Department of Electrical Engineering at Universitas Ahmad Dahlan (UAD), Indonesia. He is Editor-in-Chief, Principle Contact and Editor for some Scopus indexed journals. He is also one of founder of the Indonesian Publication Index (IPI). He has completed his studies for B.Eng., M.Eng., Ph.D. in Electrical Engineering from Diponegoro University, Gadjah Mada University and Universiti Teknologi Malaysia, respectively. His research interests include the field of industrial electronics and informatics, power electronics, FPGA applications, embedded systems, data mining, information technology and digital library.

Imam Much Ibnu Subroto was born in Semarang Indonesia in 1973. Currently he is a Head Department of Electrical Engineering Postgraduate Study of Universitas Islam Sultan Agung, Semarang, Indonesia. He is also one of founder of the Indonesian Publication Index (IPI). He received B.E.E from Gadjah Mada University, Yogyakarta, Indonesia in 1998. He received M.Sc. from Universiti Teknologi Malaysia, Johor in 2007. He received Ph.D. degree from Universiti Teknologi Malaysia in 2015. His research interest in data mining, intelligence system and education technology.

Goli Arji is Ph.D. student in Health Information Management, Tehran University of Medical Sciences (TUMS), Iran. She is interested in data mining, fuzzy logic, clinical decision support system, telemedicine and consumer health informatics.