International Research Journal of Multidisciplinary Science & Technology
www.irjmrs.com ISSN : 2455-930X

COMPARISON OF PERFORMANCE OF DATA-MINING ALGORITHMS ON HEART DISEASE PREDICTION

MAI Navid[1], NH Niloy[2]*

*[1&2]Ruhea College, Bangladesh *Corresponding author's*
E-mail: niloynh1997@gmail.com

Abstract - *There are many uses of Data mining in business industries and medical facilities. In the medical spectrum the data taken is used to predict certain diseases. Heart disease is one of the most widely known diseases out there, it can happen anytime to anyone of any age. Here we use algorithms such as C5.0, Neural Network, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and finally Logistic Regression to model the data and to compare performance of these algorithms in Data mining techniques. Where the practitioner can understand these techniques and use them for future prediction for patients saving time and money.*

Key words: Regression, Classification, Nearest Neighbor, Neural Networks, Support Vector Machine, Data-mining, Heart Disease.

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 (WHO, 2015). 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 (Bagheri and Shaltooki 2014). 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 (Vijayajothi *et al*., 2014, Tan et al., 2009). 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 redactors. Neural Network, support vector machine and decision Tree are different form of classification algorithms (Nikola and Elisa, 2015; Lotte *et al*., 2007; Anderson and Peterson, 2001; Anderson *et al*., 1998; Padmavathi and Ramakrishna 2015). The purpose of this study is comparison of different machine learning algorithm on prediction of heart diseases.

This section summarizes various technical articles on KNN process and data mining classification techniques applied on heart diseases datasets:

Ram BilasPachori and his colleagues (Shivnarayan *et al*., 2015) 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 YongqiangLyu *et al*. (2015) 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 [Mai *et al*., 2011) 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*. (2008) 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 (Peter *et al*., 2013) 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*. (2014) applied MV5 algorithm and its precision was 88.52%. Another research done by Jasmine Nahar *et al*. (2013) for finding relationship between heart disease risk factors in men and women. Itrefers 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*. (2014) 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 (Rajeswari *et al*., 2012), 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 (2013) 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.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 (Quinlan, 1986; Quinlan, 1994; Quinlan, 1996; Xindong *et al*., 1996). 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 (Shuonan *et al*., 2014). 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 = 1,2,3,4, ..., N]) is used in the following formula: (,). Also for calculate an instance belonging the Ci is used to the formula:

freq (Ci, S) / |S|

The training set can be calculated accordingly:

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

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:

2. $info\,x(S) = -\sum_{j=1}^{n}\frac{|Si|}{S}\times info(Si)$

3. $gain(X) = info(S) - info\,x\,(Si)$

4. $Split\ Info(X) = -\sum_{i=1}^{n}\left(\frac{|Si|}{|S|}\times log_2\frac{|Si|}{|S|}\right)$

5. $Gain\ ratio = \frac{\Delta info}{Split\ Info}$

6. $\text{Specificity} = \frac{TN}{FP+TN}$

7. $\text{Sensitivity} = \frac{TP}{TP+FN}$

8. $\text{Precision} = \frac{TP}{TP+FP}$

9. $\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$

Figure 1 formula used in calculating.

3.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 (Sumit 2008), (Vapnik, 1995).

4.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 (Yazdani *et al*., 2009).

5.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 (Daubechies, 2009; Demuth et al, 2009; Leng et al., 2006; Frank et al., 2003).

6. Accuracy Measurement

In order to evaluate the prediction rate, there are several indices such as specificity, sensitivity, precision, and accuracy to assess to assess the models' validity. These indices(equation 6-9) are calculated by the cofusion matrix. 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 (Alizadeh et al., 2011), (Han and Kamber, 2006). FN= The number of positively labeled data, which falsely have been classified as "Negative". TN= The number of negatively labeled data, which have been classified as "Correct". TP= The number of positively labeled data, which have been classified as "Correct". FP= The number of negatively labeled data, which falsely have been classified as "Positive".

7.Data Set

In this study 270 record with 13 features has been used (UCI Archive, 2015). Patients' attributions applied for modeling, their definitions and their range of values presented in Table 1.

Table 1. Patients' attributions applied for modeling, their definitions and their range of values.

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

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 2. variables which are significantly correlated with target variable by using logestic Regression

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

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

8.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 3. Comparison on model fitness and model accuracy of four various applied machine learning algorithms

	Model Fitness (through using training set)				Model Accuracy (through using testing set)			
Algorithms	Specificity	Sensitivity	Precision	Training Accuracy	Specificity	Sensitivity	Precision	Testing Accuracy
C5.0	89.62 %	84.61 %	85.71 %	87.50 %	90.90 %	95.23 %	90.90 %	93.02%
SVM	84.90 %	79.48 %	79.48 %	82.61 %	90.90 %	80.95 %	89.47 %	86.05%
KNN	91.50 %	79.48 %	87.32 %	86.41 %	88.63 %	88.09 %	88.09 %	88.37%
Neural Network	91.50 %	78.20 %	87.14 %	85.87 %	86.36 %	73.80 %	83.78 %	80.23%

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 Tangent, CP and Slope are so important in prediction of heart diseases.

In a study conducted to comparing between data mining tools for heart diseases data set in [G.Subbalakshmi, K. Ramesh, M. Chinna Rao. 2011] and [Aditya M, Prince K, Himanshu A, Pankaj K.2014] 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 (Jesmin et al., 2013) 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 AlizadehSani et al [Roohallah *et al*., 2013] have used the C4.5 and Bagging algorithms to diagnosing coronary heart disease. For C4.5 algorithms have reported the best accuracy rate. Rajeswari et al., 2012 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 (John and Somasundaram, 2012] have been used hybrid attribute selection method for prediction of heart disease. The accuracy obtained by this model was 83.62 %. Kemal Polat and SalihGunes [Kemal and SalihGunes 2007) by use of C4.5 decision tree algorithm obtained 92.59 % accuracy.

9.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

[1] WHO Report, 2015. The Top 10 Causes of Death, last accessed 12/9/2013 from http:// who.int/mediacentre/factsheets/fs310/en/, (accessed 1 April 2015).

[2] Bagheri,H., and Shaltooki, A.A. 2014. Big Data: Challenges, Opportunities and Cloud Based Solutions. International Journal of Electrical and Computer Engineering (IJECE), 5(2): 340-343.

[3] Vijayajothi P, Tan SY, Sarinder KD and Amandeep SS. 2014. A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. Published by Elsevier, Biocybernetics and Biomedical Engineering, 34(3): 139-145.

[4] Tan, K.C., Teoh, E.J., Yu, Q., Goh, K.C. 2009. A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 36: 8616-8630.

[5] Nikola K, Elisa C. 2015. Spiking neural network methodology for modelling classification and understanding of EEG spatio-temporal data measuring cognitive processes. Information Sciences, 294: 565-575.

[6] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F. and Arnaldi, B.A. 2007. Review of classification algorithms for EEG-based brain-computer interfaces. J. Neural Eng. 4(2):1-25.

[7] Anderson, C. and Peterson, D. 2001. Recent advances in EEG signal analysis and classification, in: R. Dybowski, V. Gant (Eds.). Clinical Applications of Artificial Neural Networks, Cambridge University Press, UK. 175-191 (Chapter 8).

[8] Anderson, C., Stolz, E., Shamsunder, S. 1998. "Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. IEEE Trans. Biomed. Eng. 45(3): 277-286.

[9] Padmavathi, K. and Ramakrishna, K. S. 2015. Detection of Atrial Fibrillation using Autoregressive modeling. International Journal of Electrical and Computer Engineering (IJECE), 5(1): 64-70.

[10] Shivnarayan P, Ram BP, U. and Rajendra A. 2015. Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. Knowledge-Based Systems, 82: 1-10.

[11] Yongqiang, L., Jiaming, H., Yiran, W., Jijiang, Y., Yida, T., Wenyao, W., and Nazim, A. 2015. Dynamic evaluation model of coronary heart disease for ubiquitous healthcare. Computers in Industry, 69: 35-44.

[12] Mai, S., Tim, T., and Rob, S. 2011. Using Decision Tree for Diagnosing Heart Disease Patients. AusDM'11, Proceedings of the 9-th Australasian Data Mining Conference, Ballarat, Australia,.

[13] Sumit, B., Praveen, P.,and Pillai, G.N. 2008. SVM Based Decision Support System for Heart Disease Classification with Integer- Coded Genetic Algorithm to Select Critical Features. WCECS. Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, October 22 - 24,

[14] Peter, C., Austin, V., Jennifer, E. Ho., Daniel, L., and Douglas S. L. 2013. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology; 66(4): 398-407.

[15] Saba, B., Usman, Q., Farhan, HK., and Younus, J. 2014. MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble. Arab J SciEng; 39(11): 7771-7783.

[16] Jesmin, N., Tasadduq, I., Kevin, ST., and Yi-Ping, Ph. Ch. 2013. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Application; 40(4): 1086-1093.

[17] Kyle E., and Sean, M. 2014. Classifying high-prevalence neighborhoods for cardiovascular disease in Texas. Applied Geography; 57: 22-31, 2014.

[18] Rajeswari, K., Vaithiyanathan, V., and Neelakantan, T.R. 2012. Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks. International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), Procedia Engineering; 41: 1818-1823.

[19] Senthil, K. 2013. Generating Rules for Advanced Fuzzy Resolution Mechanism to Diagnosis Heart Disease. International Journal of Computer Applications; 77(11): 6-12.

[20] Quinlan, J R. 1986. Induction of decision trees. Machine Learning; 4: 81-106.

[21] Quinlan, J R. 1994. C4.5: Programs for machine learning. Machine Learning; 3:235-240.

[22] Quinlan, J R. and Bagging. 1996. Boosting and C4.5. Proceedings of 14th National Conference on Artificial Intelligence: 725-730.

[23] Xindong, W., Vipin, K., Ross Q., Qiang, Y., Hiroshi, M., Geoffrey, J. M., Angus, Ng., Bing, L., Philip, S., Zhi-Hua, Z., Michael, S., David, JH., and Dan, S. 2008. Top 10 algorithms in data mining. 14(1): 1-37.

[24] Shuonan, H., Rongtao, H., Xinming, S., Jun, W., and Chengshang, Y. 2014. Research on C5.0 Algorithm Improvement and the Test in Lightning Disaster Statistics", International Journal of Control and Automation, vol. 7, no1, pp. 181-190.

[25] Vapnik, V. N. 1995. The nature of statistical learning theory. New York: Springer.

[26] Yazdani, A., Ebrahimi, T., and Hoffmann, U. 2009. Classification of EEG signals using Dempster Shafer theory and a K-nearest neighbor classifier. IEEE. In: Proc of the 4th int EMBS conf on neural engineering: 327-30.

[27] Daubechies, I.1990. The wavelet transform, time-frequency localization and signal analysis. IEEE. Trans Inform Theor; 36: 961-1005.

[28] Demuth, H., Beale, M., and Hagan, M. 2009. Neural network Toolbox™ user's guide. The MathWorks, Inc.;.

[29] Leng, G., McGinnity, T.M., and Prasad, G. 2006. Design for self-organizing fuzzy neural networks based on genetic algorithms. IEEE. Trans. Fuzzy Syst; 14 (6): 755-766.

[30] Frank, H. F., Leung, H. K., Lam, S., Ling, H., and Peter, K. S. 2003 . Tuning of the structure and parameters of a neural network using an improved genetic algorithm. IEEE. Trans. Neural Networks; 14 (1): 79-88.

[31] Alizadeh, S., Ghazanfari, M., and Teimorpour, B. 2011 .Data Mining and Knowledge Discovery", Publication of Iran University of Science and Technology. 2nd ed. [Persian].

[32] Han, J. .2006. KamberM.chapter 1: introduction: Data Mining: Concepts and Techniques. Morgan Kaufman Publisher. 2nd ed.

[33] UCI Archive, Machine Learning Repository," https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/ (accessed 2 May 2015).

[34] Subbalakshmi, G., Ramesh, K., and Chinna, M. 2011. Decision Support in Heart Disease Prediction System using Naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE); 2(2): 170-176.

[35] Aditya, M., Prince, K., Himanshu, A., and Pankaj, K. 2014. Early Heart Disease Prediction Using Data Mining Techniques. Computer Science & Information Technology (CS & IT): 53-59.

[36] Roohallah, A., Jafar, H., Zahra, A., Hoda, M., Reihane, B., FahimeKh, A., and Fariba A. 2013. Diagnosing Coronary Artery Disease via Data Mining Algorithms by Considering Laboratory and Echocardiography Features. Official Journal of Rajaie Cardiovascular Medical and Research Center; 2(3): 133-139.

[37] John, P., and Somasundaram, K. 2012. Study and Development of Nevel Feature Selection Frmework for Heart Disease Preciction. International Journal of Scientific and Research Publications; 2(10): 1-7.

[38] Kemal P.,and Salih, G. 2007.A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Computer methods and programs in biomedicine; 88: 164-174.