Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 8, August 2013, pg.102 – 107

RESEARCH ARTICLE

Intelligent Heart Disease Prediction Model Using Classification Algorithms

Pramod Kumar Yadav[1], K.L.Jaiswal[2], Shamsher Bahadur Patel[3], D. P.Shukla[4]

[1]Research Scholar, Department of Computer Application & Physics,
Govt. P. G. Science College Rewa (M.P.), India

[2]Assistant Professor and In charge of BCA, DCA & PGDCA, Department of Physics,
Govt. P.G. Science College, Rewa (M.P.), India

[3]Research Scholar, Department of Computer Science & Mathematics,
Govt. P. G. Science College Rewa (M.P.), India

[4] Professor and Head, Department of Computer Science & Mathematics,
Govt. P. G. Science College Rewa (M.P.), India

[1] yadav.pramod181@gmail.com, [2] drkanhaiyalaljaiswal@gmail.com, [3] sspatel12@gmail.com, [4] shukladpmp@gmail.com

Abstract— Data mining technique have led over various methods to gain knowledge from vast amount of data. So, different research tools and techniques like association rule, Classification algorithms, and decision tree etc. This paper analyses the performance of various classification function techniques in data mining for prediction heart disease from the heart disease data set. The classification algorithms used and tested in work are Logistics, Multi-layer Perception and Sequential Minimal Optimization algorithms. The performance factor used for analyzing the efficiency of algorithm are clustering accuracy and error rate. The result show logistics classification function efficiency is better than multi-layer perception and sequential minimal optimization.

Key Terms: - Data mining; sequential minimal optimization; multilayer perception; logistics; Disease prediction

I. INTRODUCTION

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. It is "the science of extracting useful information from large databases". It is one of the tasks in the process of knowledge discovery from the database. [1]. Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans. It is a process to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. There are two primary goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand Description focuses on finding patterns describing the data that can be interpreted by humans.

The medical data mining has the high potential in medical domain for extracting the hidden patterns in the datasets [3]. These patterns are used for clinical diagnosis and prognosis. The medical data are widely

distributed, heterogeneous, voluminous in nature. The data should be integrated and collected to provide a user oriented approach to novel and hidden patterns of the data. A major problem in medical science or bioinformatics analysis is in attaining the correct diagnosis of certain important information. For an ultimate diagnosis, normally, many tests generally involve the classification or clustering of large scale data.

The test procedures are said to be necessary in order to reach the ultimate diagnosis. However, on the other hand, too many tests could complicate the main diagnosis process and lead to the difficulty in obtaining the end results, particularly in the case of finding disease many tests are should be performed. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers. Classification is one of the most important techniques in data mining. If a categorization process is to be done, the data is to be classified, and/or codified, and then it can be placed into chunks that are manageable by a human [12]. This paper describes classification function algorithms and it also analyzes the performance of these algorithms. The performance factors used for analysis are accuracy and error measures. The accuracy measures are True Positive (TP) rate, F Measure, ROC area and Kappa Statistics. The error measures are Mean Absolute Error (M.A.E), Root Mean Squared Error (R.M.S.E), Relative Absolute Error (R.A.E) and Relative Root Squared Error (R.R.S.E).

II. HEART DISEASE PREDICTION

Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis for widely distributed in raw medical data which is heterogeneous in nature and voluminous. These data should be collected in an organized form. This collected data can be integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. From the analysis of World Health Organization, they estimated 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths occur in United States and other developed countries due to cardio vascular diseases. On the above discussion, it is regarded as the primary reason behind deaths in adults. Heart disease kills one person every 34 seconds in the United States. The following paper reviewed about predicting of heart disease using data mining technique.

III. METHODS

A. Data source

In this paper, we use the heart disease data from machine learning repository of UCI [11]. We have total 303 instances of which 164 instances belonged to the healthy and 139 instances belonged to the heart disease. 15 clinical features have been recorded for each instance.

S.N.	Clinical feature	description
01	Age	Age in year
02	Sex	Value 1:Male,value 0:Female
03	Chest Pain Type	value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic
04	Fasting Blood Sugar	value 1: >120 mg/dl; value 0: <120 mg/dl
05	Restecg	resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy
06	Exang	exercise induced angina (value 1: yes; value 0: no
07	Slope	the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
08	CA	number of major vessels colored by floursopy (value 0-3)
09	Thal	Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10	Trest Blood Pressure	mm Hg on admission to the hospital
11	Serum Cholestrol	mg/dl
12	Thalach	maximum heart rate achieved
13	Oldpeak	ST depression induced by exercise
14	Smoking	value 1: past; value 2: current; value 3: never
15	Obesity	value 1: yes; value 0: no

Table 1- Clinical features and their description

B. Classification Algorithms

Classification algorithm plays an important role in heart disease prediction. In this paper we have analyzed three Classification Algorithms. The algorithms are namely logistic, Multilayer perception and Sequential Minimal Optimization.

Sequential Minimal Optimization

The SMO class implements the sequential minimal optimization algorithm, which analyzed this type of classifier [4]. It is one of the highest methods for learning support vector machines. Sequential minimal optimization is often slow to compute the solution, particularly when the data items are not linearly separable in the space span by the nonlinear mapping. This should be happen, because of noise data. Both accuracy and run time depend critically on the values that are given to two parameters: the degree of polynomials in the non-linear mapping (–E) and the upper bound on the coefficients values in the equation for the hyper plane (–C). By default both are set to be 1. The best settings for a heart disease dataset can be found only by experimentation [4].

Algorithm 1: SMO
1. Input: C, kernel, kernel parameters, epsilon
2. Initialize b and all □'s to 0
3. Repeat until KKT(Karush-Kuhn-Tucker) satisfied (to within epsilon):
– Find an example *e1* that violates KKT (prefer unbound examples here, choose randomly among those)
– Choose a second example *e2*. Prefer one to maximize step size (in practice, faster to just maximize |E1 – E2|). If that fails to result in change, randomly choose unbound example. If that fails, randomly choose example. If that fails, re-choose e1.
– Update α1 and α2 in one step
– Compute new threshold b

Multi-Layer Perception

Multilayer Perception classifier is based on back propagation algorithm to classify instances of data. The network is created by an MLP algorithm. The network can also be modified and monitored during training phase. The nodes in this neural network are all sigmoid. The back propagation neural network is referred as the network of simple processing elements working together to produce an output. The multilayer feed-forward neural network should be learned by performing the back propagation algorithm. It should be learned by a set of weights for predicting the class label of tuples. The neural network consists of three layers namely input layer, one or more hidden layers, and an output layer [15]. Each layer should be made up of units. The input layer of the network correspond to the attributes should be measured for each training values. To make input layer, the inputs are fed simultaneously into the units. These inputs are passed through the input layer and are then weighted and fed simultaneously to a second layer of neuron like units, which is known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used [6]. At the core, back propagation is simply an efficient and exact method for calculating all the derivatives of a single target quantity (such as pattern classification error) with respect to a large set of input quantities (such as the parameters or weights in a classification rule) [15]. To improve the classification accuracy we should reduce the training time of neural network and reduce the number of input units of the network [13].

Algorithm 2: MLP
1. Apply an input vector and calculate all activations, a and u
2. Evaluate Dk for all output units via:
$\Delta i(t) = (di(t) – yi(t))g'(ai(t))$
(Note similarity to perceptron learning algorithm)
3. Backpropagate Dks to get error terms d for hidden layers using:
$\delta i(t) = g'(ui(t))\Sigma k\ \Delta k(t)wki$
4. Evaluate changes using:
□ij(t+1) = □ij (t) + $\eta\delta i(t)$ xj(t)
$wij(t+1) = wij (t) + \eta\Delta i(t)$ zj(t)

Logistic Algorithm:

The term regression can be defined as the measuring and analyzing the relation between one or more independent variable and dependent variable [18]. Regression can be defined by two categories; they are linear regression and logistic regression. Logistic regression is a generalized by linear regression [8]. It is mainly used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modeled directly by linear regression i.e. discrete variable changed into continuous value. Logistic regression basically is used to classify the low dimensional data having nonlinear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. So, the main motto of Logistic regression is to determine the result of each variable correctly Logistic regression is also known as logistic model/ logit model that provide categorical variable for target variable with two categories such as light or dark, slim/ healthy.

Algorithm 3: Logistic

1. Suppose we represent the hypothesis itself as a logistic function of a linear combination of inputs:
$h(x)=1 / 1 + \exp(wTx)$
This is also known as a sigmoid neuron.
2. Suppose we interpret
$h(x)$ as $P(y=1|x)$
3. Then the log-odds ratio,
$\ln (P(y=1|x)/P(y=0|x))=wT\ x$ which is linear
4. The optimum weights will maximize the conditional likelihood of the outputs, given the inputs.

IV. EXPERIMENTAL RESULTS

A. Accuracy Measure

The following table shows the accuracy measure of classification techniques. They are the True Positive rate, F Measure, Receiver Operating Characteristics (ROC) Area and Kappa Statistics. The TP Rate is the ratio of play cases predicted correctly cases to the total of positive cases. . It is a probability corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. F Measure is a way of combining recall and precision scores into a single measure of performance. Recall is the ratio of relevant documents found in the search result to the total of all relevant documents [2]. Precision is the proportion of relevant documents in the results returned. ROC Area is a traditional to plot this same information in a normalized form with 1-false negative rate plotted against the false positive rate.

TABLE 2: Accuracy measure for function algorithm

Algorithm	F Measures	TP Rate	ROC Area	Kappa Statistics
SMO	**69.3**	**70.52**	**86.8**	**53.81**
MLP	**69.7**	**69.53**	**91.2**	**52.79**
Logistic	**70.4**	**70.86**	**92.2**	**54.6**

From the graph, this work analyzed that, TP rate accuracy of logistic function performs better when compared to other algorithms. When compared to F Measure accuracy logistic function produced better results than MLP and SMO. The ROC Area of the point attains the highest accuracy in logistic function algorithm. At last the accuracy measure of Kappa statistics performs better in logistic function than other algorithm. As a result the logistic function performs better accuracy than multilayer perception and sequential minimal optimization.

B. Error Rate

The table 3 shows the Error rate of classification techniques. They are the Mean Absolute Error (M.A.E), Root Mean Square Error (R.M.S.E), Relative Absolute Error (R.A.E) and Root Relative Squared Error (R.R.S.R) [10]. The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes. The root mean square error (RMSE) is defined as frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. It is a good measure of accuracy, to compare the forecasting errors within a dataset as it is scale-dependent. Relative error is a measure of the uncertainty of measurement compared to the size of the measurement. The root relative squared error is defined as a relative to what it would have been if a simple predictor had been used. More specifically, this predictor is just the average of the actual values. Thus, the relative squared error

manipulates by taking the total squared error and normalizes it by dividing by the total squared error of the simple predictor. One reduces the error to the same dimensions as the quantity by taking the square root of the relative squared error is being predicted.

TABLE 3: Error rate for function algorithm

Algorithm	R.A.E	R.M.S.E.	R.R.S.R.	M.A.E.
SMO	**95**	**34.83**	**99**	**26.15**
MLP	**47.79**	**30.7**	**85.53**	**12.36**
Logistic	**46.40**	**27.43**	**76.44**	**12**

From the graph, it is observed that SMO and MLP attains highest error rate. Therefore the logistic function algorithm performs well because it contains least error rate when compared to multilayer perception (MLP) and sequential minimal optimization (SMO) algorithm.

V. CONCLUSION

There are different data mining techniques that can be used for the identification and prevention of heart disease among patients. In this paper, three classification algorithms techniques in data mining are intelligent for predicting heart disease. They are function based Logistic, Multilayer perception and Sequential Minimal Optimization algorithm. By analyzing the experimental results, it is observed that the logistic classification algorithms technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least error rate. In future we tend to improve performance efficiency by applying other data mining techniques and optimization techniques. It is also enhanced by reducing the attributes for the heart disease dataset.

REFERENCES

[1] Mai Shouman, Tim Turner, Rob Stocker,(2012),"Using Data Mining Techniques In Heart Disease Diagnosis And Treatment ",Proceedings in Japan-Egypt Conference on Electronics, Communications and Computers,IEEE,Vol.2 pp.174-177.

[2] Anchana Khemphila and Veera Boonjing (2011), "Heart Disease Classification Using Neural Network And Feature Selection", in Proc. 21st International Conference on Systems Engineering,IEEE,vol.3 pp. 406-409.

[3] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi,and Constantinos S. Pattichis(2010),"Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE Transactions On Information Technology In Biomedicine, Vol. 14, No. 3.pp.559-566.

[4] K.Srinivas , B.Kavita Rani, Dr. A.Govardhan (2010), Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks ,IJCSE Vol. 02, No. 02, pp 250-255.

[5] M. Karaolis, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis(2009), "Association rule analysis for the assessment of the risk of coronary heart events," in Proc. 31st Annu. Int. IEEE Eng. Med. Biol. Soc. Conf., Minneapolis, MN, Sep. 2–6, pp. 6238–6241.

[6] Sellappan Palaniappan, Rafiah Awang(2008), "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8. pp 343-350.

[7] M. Karaolis, J. A.Moutiris, and C. S. Pattichis(2008), "Assessment of the risk of coronary heart event based on data mining," in Proc. 8th IEEE Int. Conf. Bioinformatics Bioeng., pp. 1–5.

[8] K. Polat, S. Sahan, H. Kodaz, and S. Guenes(2007), "A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS," Comput.Methods Programs Biomed., Vol. 88, no. 2, pp. 164–174.

[9] C.Ordonez(2006), "Comparing association rules and decision trees for disease prediction," in Proc. Int. Conf. Inf. Knowl. Manage.,Workshop Healthcare Inf. Knowl. Manage. ,IEEE,Arlington, VA, pp. 17–24.

[10] R. B. Rao, S. Krishan, and R. S. Niculescu(2006), "Data mining for improved cardiac care," ACM SIGKDD Explorations Newsletter., vol. 8, no. 1, pp. 3–10.

[11] S. A. Pavlopoulos, A. Ch. Stasis, and E. N. Loukis(2004), "A decision tree based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds," Biomedical Engineering OnLine, vol. 3, p.

[12] C. Ordonez, E. Omiecinski, L. de Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczvnska, and E. V. Garcia(2001), "Mining constrained association rules to predict heart disease," in Proc. IEEE Int. Conf. Data Mining (ICDM), pp. 431–440.

[13] C. L. Tsien, H. S. F. Fraser, W. J. Long, and R. L. Kennedy(1998), "Using classification trees and logistic regression methods to diagnose myocardial infarction," in Proc. 9th World Congress. ed. Inf., Vol. 52, pp. 493–497.

[14] J. Han and M. Kamber, Data Mining, Concepts and Techniques, 3rd edition, San Francisco, CA: Morgan Kaufmann,2011.

[15] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques over Heart Disease Data base" [IJESAT] international journal of engineering science & advanced technology ISSN: 2250–3676, Volume-2, Issue-3, 470 – 478

[16] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2005.

