ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

Comparative Analysis of Classification Function Techniques for Heart Disease Prediction

Dr. S.Vijayarani[1], S.Sudha[2]

Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, India[1]

M.Phil Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India[2]

ABSTRACT: The data mining can be referred as discovery of relationships in large databases automatically and in some cases it is used for predicting relationships based on the results discovered. Data mining plays an important role in various applications such as business organizations, e-commerce, health care industry, scientific and engineering. In the health care industry, the data mining is mainly used for Disease Prediction. Various data mining techniques are available for predicting diseases namely clustering, classification, association rules, regression and etc. This paper analyses the performance of various classification function techniques in data mining for predicting the heart disease from the heart disease data set. The classification function algorithms used and tested in this work are Logistics, Multi Layer Perception and Sequential Minimal Optimization algorithms. Comparative analysis is done by using Waikato Environment for Knowledge Analysis or in short, WEKA. It is open source software which consists of a collection of machine learning algorithms for data mining tasks. The performance factors used for analysing the efficiency of algorithms are clustering accuracy and error rate. The result shows that logistics classification function efficiency is better than multi layer perception and sequential minimal optimization.

Keywords: Data mining, Disease prediction, logistics, multi layer perception, sequential minimal optimization.

I. INTRODUCTION

Data mining can be defined as the extraction of useful knowledge from large data repositories. Compared with other data mining application fields, medical data mining plays a vital role and it has some unique characteristics. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature: Massive data collection, Powerful multiprocessor computers and Data mining algorithms

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).

The rest of this paper is organized as follows. Section 2 describes the review of literature. Section 3 discusses the classification function algorithms used for predicting the heart disease. Experimental results are analyzed in Section 4 and Conclusions are given in Section 5.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

II. LITERATURE REVIEW

In [7], the cardiovascular heart disease is predicted by the classification techniques namely artificial neural networks, RIPPER, decision tree and support vector machine. The author concluded that the support vector machine performs well when compared to other algorithm because it attains least error rate and highest accuracy.

In [1] the classification data mining techniques is used for analyzing the performance of an algorithm. The algorithms used are WAC, naïve bayes, and Apriori. As a result the performance efficiency is evaluated by using classification matrix.

In [14], the heart disease is analyzed by neural network approach which includes variable length rate with momentum and the back propagation algorithm. The author finds that the efficiency is high in classification process by applying parallel approach which is included in the training phase.

In [9] the heart attack is predicted by applying association rule mining technique. The proposed algorithm CBARBSN is based on sequence numbers and clustering of the transactional database. The proposed algorithm CBARBSN performed well than the existing ARNBSN algorithm. Based on the execution time the performance is compared.

In [11] based on the probability of decision support the heart disease is predicted. As a author analyzed that decision tree performs well and sometimes the accuracy is similar in Bayesian classification.

III.HEART DISEASE PREDICTION

Heart disease prediction plays a vital role in data mining because in worldwide most of the death occur in heart diseases. Medical diagnosis plays an important role and yet complicated task that needs to be executed efficiently and accurately. To reduce cost for achieving clinical tests an appropriate computer based information and decision support should be needed. Comparative studies of various techniques are available for an accurate implementation and good efficient for automated systems**.** [12]

The World Health Organization (WHO) analyzed that twelve million deaths occurs worldwide due to Heart diseases. Many of the deaths occurs in United States and other developed countries due to cardio vascular diseases. Heart disease is the major causes of different countries include India. In every 34 seconds the heart disease kills one person. There are various categories in Heart disease but it mainly focuses on three types namely Cardiovascular Disease, Cardiomyopathy and Coronary heart disease.

Now a day's most of the people are affected by heart diseases. There are different types of heart diseases. They are discussed as follows.

Coronary Artery Disease: When the combination of fatty material, calcium and scar tissue (plague) builds up in the arteries that supply the heart with blood through this, the disease should develops. Through these arteries called the coronary arteries, the heart muscle (myocardium) gets the oxygen and other nutrients it needs to pump blood. Coronary artery disease is America's No.1 killer, affecting more than 13 million Americans. *Enlarged Heart (Cardiomegaly):* A heart condition that causes the heart to become larger than normal as a result of heart disease. Cardiomegaly is most often linked to high blood pressure, but it can also occur as a result of other heart conditions, such as congestive heart failure, and other non-cardiac causes such as long-term anemia [17]. *Heart Attack:* A heart attack is the death of, or damage to, part of the heart muscle because the supply of blood to the heart muscle is severely reduced or stopped.

Heart Valve Disease: Valvular heart disease refers to several disorders and diseases of the heart valves, which are the tissue flaps that regulate the flow of blood through the chambers of the heart. *Congenital Heart Disease:* Congenital heart disease refers to a problem with the heart's structure and function due to abnormal heart development before birth. Congenital means present at birth [17].

A. Data Source

In order to compare the data mining classification techniques Cleveland cardiovascular disease dataset from UCI repository was used [5]. This dataset has 13 attributes and 303 instances. These data are analyzed in Weka tool [16].The attributes are

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

Attributes	Description
Age	Age in years
Sex	Male=1,Female=0
Cp	Chest pain type
Blood pressure	Resting Blood pressure upon hospital admission
Cholesterol	Serum Cholesterol in mg
blood sugar	Fasting blood sugar>120 mg/dl true=1 and false=0
Resting ECG	Resting electrocardiographic results
Thalach	Maximum Heart Rate
Induced Angina	Does the patient experiment angina as a result of exercise
Old peak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
CA	Number of major vessels colored by fluoroscopy
Thal	Normal ,fixed defect, reversible defect

B. Classification Function Algorithms

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

1. SMO

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 α'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 $|E_1 - E_2|$). If that fails to result in change, randomly choose unbound example. If that fails, randomly choose example. If that fails, re-choose e1.
   - Update α_1 and α_2 in one step
   - Compute new threshold b

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

2. Multi Layer Perceptron

Multilayer Perceptron 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 D_k for all output units via:

$$\Delta_i(t) = (d_i(t) - y_i(t))g'(a_i(t))$$

(Note similarity to perceptron learning algorithm)

3. Backpropagate D_ks to get error terms d for hidden layers using:

$$\delta_i(t) = g'(u_i(t))\sum_k \Delta_k(t)w_{ki}$$

4. Evaluate changes using:

$$\nu_{ij}(t+1) = \nu_{ij}(t) + \eta\delta_i(t)\, x_j(t)$$

$$w_{ij}(t+1) = w_{ij}(t) + \eta\Delta_i(t)\, z_j(t)$$

3. Logistic Function:

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 non linear 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:
   $h(x)=1 / 1 + \exp(w^T x)$
   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,
   $\ln (P(y=1|x)/P(y=0|x))=w^T x$ which is linear
4. The optimum weights will maximize the conditional likelihood of the outputs, given the inputs.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

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 1: Accuracy measure for function algorithm

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

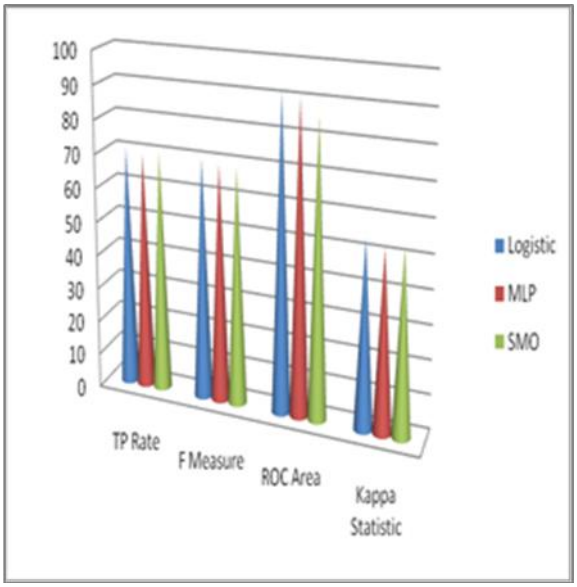

Fig1: Accuracy measure for function algorithm

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 perceptron and sequential minimal optimization.

B. Error Rate

The table 2 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

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

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 2: Error rate for function algorithm

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

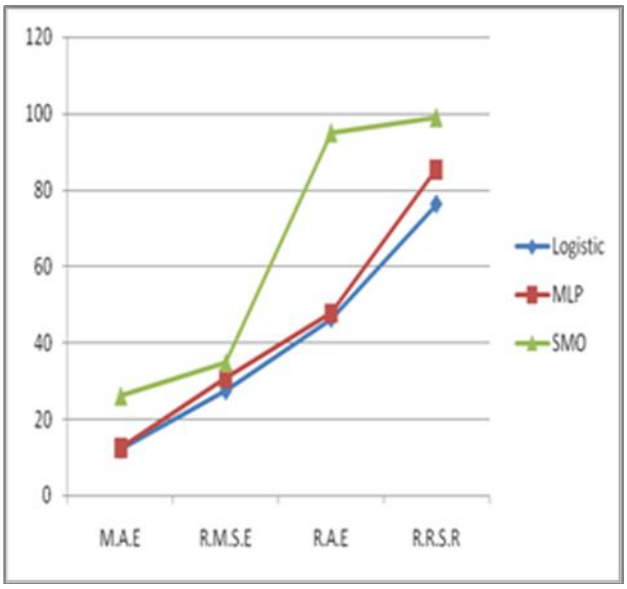

Fig 2: Error rate for function algorithm

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 perceptron (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 function techniques in data mining are compared for predicting heart disease. They are function based Logistic, Mutilayer perceptron and Sequential Minimal Optimization algorithm. By analyzing the experimental results, it is observed that the logistic classification function 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. 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
2. "Binary classification performances measure cheat sheet" www.damienfrancois.be/blog/files/modelperfcheatsheet.pdf
3. 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.
4. Ian H. Witten Eibe Frank, "WEKA Machine Learning Algorithms in Java", 2000 Morgan Kaufmann Publishers. All rights reserved
5. "Cleveland heart disease dataset" sci2s.ugr.es/keel/dataset.php?cod=57
6. Margaret H Dunham., "Data mining: Introductory and Advanced Topics", Published by Pearson Education., Inc., Copyright@2003.
7. Esra Mahsereci Karabulut & Turgay İbrikçi "Effective Diagnosis of Coronary Artery Disease Using The Rotation Forest Ensemble Method" June 2011 / Accepted: 30 August 2011 / Published online: 13 September 2011 # Springer Science+Business Media, LLC 2011
8. De Mantaras & Armengol E. (1998),"Machine learning from example: Inductive and Lazy methods", Data & Knowledge Engineering 25: 99-123

9. MA.JABBAR, Dr. PRITI CHANDRA, B.L.DEEKSHATULU "Cluster Based Association Rule Mining For Heart Attack Prediction" JTAIT Vol. 32 No.2 October 2011.
10. Nithyasri.B, Nandhini.K, Dr. E.Chandra "CLASSIFICATION TECHNIQES IN EDUCATION DOMAIN" (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1679-1684
11. Dr. D. Raghu. T. Srikanth, Ch. Raja Jacub, "Probability based Heart Disease Prediction using Data Mining Techniques" IJCST Vol. 2, Issue 4, Oct - Dec. 2011, ISSN: 0976-8491 (Online) | ISSN: 2229-4333(Print)
12. Ruben D. Canlas Jr.,"DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES", August 2009
13. H. Lu, R. Setiono, and H. Liu, "Effective Data Mining Using Neural Networks", IEEE, 1996
14. Dr. Usha Rani.K "Analysis of Heart Diseases Dataset Using Neural Network Approach" (IJDKP) Vol.1, No.5, September 2011
15. Rohit Arora, Suman, " Comparative Analysis of Classification Algorithms on Different Datasets using WEKA", *International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012*
16. Weka: http://www.cs.waikato.ac.nz/~ml/weka/
17. www.webmd.com/heart-disease/guide/heart-diseasesymptoms-Types
18. Yugal kumar, G. Sahoo, "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA", *I.J. Information Technology and Computer Science*, 2012, 7, 43-49 Published Online July 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.2012.07.06

BIOGRAPHY

Mrs. Dr. S.Vijayarani has completed MCA and M.Phil in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security issues and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. S.Sudha has completed M.Sc in Software Systems. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are privacy in data mining.