Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Prediction of Heart Disease Based on Risk Factors Using Genetic SVM Classifier

Rajwant Kaur
M.Tech Student
Department of Computer Science & Engineering
SGGSWU, Fatehgarh Sahib, Punjab, India

Sukhpreet Kaur
Assistant Professor
Department of Computer Science & Engineering
SGGSWU, Fatehgarh Sahib, Punjab, India

Abstract: Nowadays, Heart Disease is a major cause of dejection and fatality. All over the world, deaths due to Heart Disease is increasing rapidly than from any other disease. It is very difficult to predict the probable complications regarding Heart Disease well in advance. To identify the probable complications, many systems are made that uses clinical data sets for identifications. Some of the systems predict heart disease based on risk factors. A lot of visible risk factors that are common in Heart Disease patients can be used effectively for diagnosis. System based on risk factors helps not only medical experts for prediction but also warn the patients in advance about the probable presence of heart disease. These systems are also helpful to save money and time. Hence, taking an assumption forward a hybrid technique is proposed for prediction of Heart Disease on basis of risk factors. Data mining tools used for this system is: - Support Vector Machine (SVM) Classifier and Genetic Algorithm. The hybrid implemented technique uses the global optimization advantage of GA for initialization of Support Vector Machines (SVM) Weights. This technique makes system fast, more stable and accurate as compare to others. The system was implemented in MATLAB and on the basis of risk factors an accuracy of implemented system is 95%.

Keywords: Heart Disease, Clinical Dataset, Risk Factors, Support Vector Machine, Genetic Algorithm,

I. INTRODUCTION

Data mining is a knowledge discovery technique to analyze data and summarize it into useful information [5]. The analyzed data can be used for various applications. One of the most important applications is use in medical field. Nowadays, as the population of world increases, medical industries generates huge amount of data related to patients and diseases diagnoses etc. and facing major challenges. A major challenge of medical industries is to diagnose disease accurately and high cost of medical tests. Data mining techniques are used to solve these challenges, with high quality of service. Data mining provides techniques to find out hidden patterns and relations of medical data [1]. The mortality rate caused by heart disease has been increasing all over the world. So, there is need for development of heart disease prediction methods. According to one survey in 2008 approximately 17.3 million people died from Heart Diseases representing 30% of all global deaths. An estimated 7.3 million deaths were due to coronary heart disease and 6.2 million were due to stroke [1]. Recently, many researches in medical industries have been able to identify risk factors of Heart Diseases but more contribution is necessary to use this knowledge to reduce causes of deaths [1]. Various data mining techniques have been used to make clinical decision support systems, to get accurate results on the basis of information collected by researches from study. These systems allow patients to calculate the Heart Disease risks. IHDPS (intelligent heart disease prediction system) is capable to discover and extract hidden knowledge associated with heart disease on the basis of historical Heart Disease database [6].

II. DATA MINING TECHNIQUES

In order to discover unknown patterns, data mining techniques are used to explore, analyze and extract medical data with the help of different algorithms. Researchers are using data mining techniques for the diagnosis of many diseases such as heart disease, diabetes, stroke, and cancer [1]. Many researchers have been applying different data mining techniques such as Neural Network, Random Forest, Decision tree, K-Nearest Neighbor, Genetic Algorithm etc. One of the systems uses KNN. This approach combines KNN with Genetic Algorithm to improve the classification accuracy of heart disease data set. It used genetic search as a goodness measure to prune redundant and irrelevant attributes and to rank the attributes which contribute more towards classification. Least ranked attributes are removed and classification algorithm is built based on evaluated attributes. This classifier is trained to classify Heart Disease data set as either healthy or sick [7]. In another system, proposed Rule based model compare the accuracies of other applied rules to the individual results of Support Vector Machine (SVM), Decision Tree, and Logistic Regression on the Cleveland Heart Disease database in order to present an accurate model of predicting Heart Disease. Using the Cleveland Heart Disease database, this technique provides guidelines to train and test the system and thus attain the most efficient model of the multiple rule based combinations [10]. Abdullah, and Rajalaxmi(2012), have developed the data mining model by using Random Forest classifier to improve the prediction accuracy and to investigate various events related to CHD. That model could help the medical practitioners for predicting CHD with its various events and how it might be related with different

segments of the population [3]. Amin and Syed (2013)**,** have proposed a technique for prediction of heart disease using major risk factors. That technique was involved two most successful data mining tools, neural networks and genetic algorithms. The implemented technique involved to predict the risk of Heart Disease with an accuracy of 89% [1]. By studying the different techniques discussed above, this paper proposes a novel approach amalgamate Genetic Algorithm and SVM Classifier for prediction of Heart Disease based on risk factors.

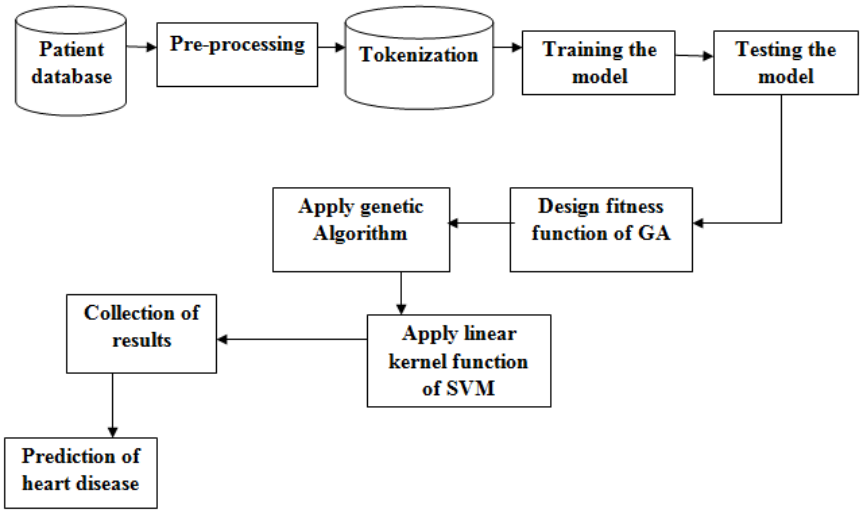
III. METHODOLOGY

A. The Data

Many risk factors that cause heart disease and it is very difficult to understand and categorized**.** Most of time Heart Disease is detected when a patient reaches at last stage of disease. The risk factors help to analyze the disease in advance. The dataset was composed of 12 important risk factors are sex, age, family history blood pressure, Smoking Habit, alcohol consumption, physical inactivity, diabetes, blood cholesterol, poor diet and obesity. The system is used to know whether the patient has risk of Heart Disease or not. The dataset contains 50 people data collected from different services done by American Heart Association.

B. SVM weight optimization by Genetic Algorithm

Support Vector Machine weight optimization by Genetic Algorithm. This system uses Linear Kernel function for learning and training the SVM Classifier. But the results produced by this were not much better. To get more accurate results the genetic algorithm is used with SVM as optimizer. For this, the fitness function of Genetic Algorithm is used as input for SVM Classifier. The Linear Kernel function is used to Train the dataset using the weights optimized by Genetic Algorithm.



Process of SVM + Genetic Algorithm Approach for Heart Disease Prediction

C. SVM Classifier.

A SVM Classifier is used having input space and feature space. The input space is based on the final set of risk factors for each patient. The feature space is obtained by maximizing the margin between the two classes for which training is fast and gives the best output results. The first step is to initialize the weights of SVM Classifier with the use of configure function available in MATLAB. According to the fitness function configured weights are passed to Genetic Algorithm for optimization. When the weights are optimized, the linear kernel function is used for training and learning. And a training function (Xtrain) is to update the weight values according to the linear kernel function values. Learning stops at misclassification after modify classifier weights and adjusting these to an optimum value at which results of classification is accurate. The obtained output would be presence or absence of Heart Disease.

IV. PERFORMANCE MEASURES

In this approach the accuracy rates of classification for the data set is measured. The system is developed using MATLAB 2012a. The data, basis on the risk factor related to Heart Disease collected from 50 people through case studies. The different performance parameters are used to obtain the accurate results. In this classification problem a prediction has four possibilities: True Positive (TP) Rate, True Negative (TN) Rate, False Positive (FP) Rate and False Negative (FN) Rate. Where TP and TN are correct classification and FP is the outcome incorrectly predicted as positive and FN is outcome incorrectly predicted as negative.

Table 1. A Confusion Matrix for Prediction Outcomes:

Prediction	Disease		
		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

The following set of Evaluation measures are being used to find out the results [3].

Sensitivity: A high sensitivity is clearly important where the test is used to identify a serious but treatable disease.

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

Specificity: The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease.

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

Accuracy: Accuracy measures correctly figured out the diagnostic test by eliminating a given condition.

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

Precision: is the fraction of retrieved instances that are relevant. Precision is calculated by:

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

Recall: - is the fraction of relevant instances that are retrieved. Recall is calculated by:

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure: - A measure that compile precision and Recall is the harmonic mean of Precision and Recall, the traditional F-measure and balanced F-score. It totally depends upon value Precision and Recall. F-measure is calculated by:

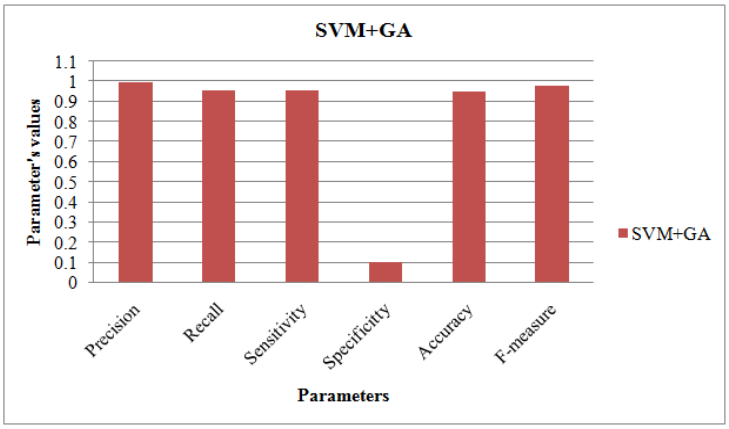
$$\text{F-measure} = \frac{(2*Precision\ *Recall\ )}{(Precision\ +Recall\ )}$$

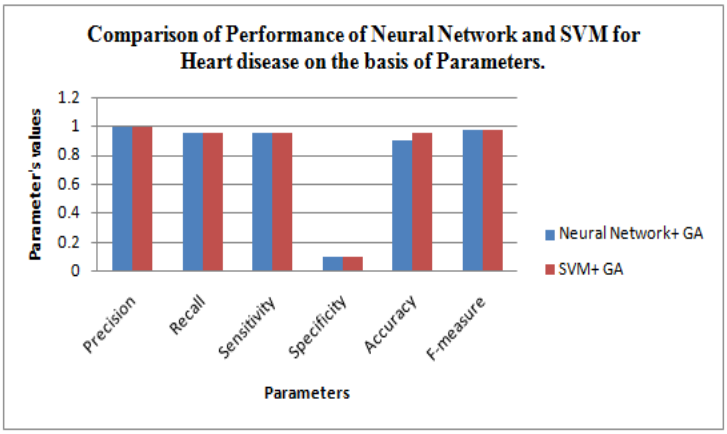
The results on the Test Set will be displayed in the form of two-dimensional confusion matrix having a row and column for each class. The number of test cases for which the actual class is the row and predicted class is the column, are shown by the matrix elements.

V. RESULTS AND DISCUSSION

The data of 50 people collected based on the risk factors through different case studies provided at the web site of American Heart Association. 70% data is used for training and 30% for testing. The accuracy of Heart Disease prediction on the data is calculated as 95%. The obtained results are shown in figure 1. Results show that SVM and Genetic Algorithm give better accuracy than the Neural Network

Algorithms	Precision	Recall	Sensitivity	Specificity	Accuracy	F-measure
Neural Network+ GA	0.991361	0.954545	0.954545	0.103571	0.904444	0.972605
SVM+ GA	0.996805	0.957895	0.957895	0.103361	0.95152	0.976963





The sensitivity of classification system using Neural Network + GA was 0.954545 and with use of SVM + GA, the sensitivity of classification system results 0.957895. The precision of Neural Network + GA was 0.991361 and the

precision of SVM + GA is 0.996805. The accuracy of Neural Network + GA was computed 0.991361 and accuracy of SVM + GA is 0.996805. Results computed on the basis of different parameters shows SVM + GA perform better than Neural Network + GA.

VI. CONCLUSION

Data mining techniques applied in Medical field provides innovative results and decision support system used to improve the health of patients and for other medical services. But still it needs to improve the system to predict probable complications in advance. In this study, a new approach that combines SVM Classifier with Genetic Algorithm to improve the performance of SVM classifier. The system predicts Heart Disease on the basis of risk factors and save the money and time to undergo costly medical tests. Risk factors data of 50 patients used to test the system and achieved results showed accuracy of 95%. For future, The system can be enhance and improve the performance of diagnosis of the Heart Disease by using the better optimization algorithm to create the model which will give the efficient result and also use another classification methods and algorithms.

REFERENCES

[1] Amin, Syed Umar, Kavita Agarwal, and Rizwan Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors", *Information & Communication Technologies (ICT), 2013 IEEE Conference on*, pp. 1227-1231. IEEE, 2013.

[2] Mythili, T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu. "A heart disease prediction model using svm-decision trees-logistic regression (sdl)."*International Journal of Computer Applications,* Issue 16, Vol.68, pp. 11-15, April 2013.

[3] Abdullah, A. S., and R. Rajalaxmi. "A data mining model for predicting the coronary heart disease using random forest classifier." In *International Conference in Recent Trends in Computational Methods, Communication and Controls*, pp.22-25, 2012.

[4] Deekshatulu, B. L., and Priti Chandra, "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm", *Procedia Technology* 10,pp. 85-94, 2013

[5] Masethe, Hlaudi Daniel, and Mosima Anna Masethe. "Prediction of Heart Disease using Classification Algorithms." In *Proceedings of the World Congress on Engineering and Computer Science*, Vol. 2, pp.22-25, 2014.

[6] Waghulde, Nilakshi P., and Nilima P. Patil. "Genetic Neural Approach for Heart Disease Prediction.*" International Journal of Advanced Computer Research",* Issue-16, Vol.4, 2014.

[7] Deekshatulu, B. L., and Priti Chandra, "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm", *Procedia Technology,Elsebier* ,Vol.10,pp. 85-94, 2013

[8] Cinetha, K., and P. Uma Maheswari. "Decision Support System for Precluding Coronary Heart Disease (CHD)." *International Journal of Computer Science and Mobile Computing,* Issue.2, Vol.3, pp.34-38, Feb 2014.

[9] Jabbar, M. Akhil, B. L. Deekshatulu, and Priti Chandra. "Classification of Heart Disease Using Artificial Neural Network and Feature Subset Selection."*GJCST,* Issue 3, Vol.13, 2013.

[10] Taneja, Abhishek. "Heart disease Prediction System Using data Mining Techniques." *Oriental Journal of Computer Science & Technology,* Issue 4, Vol.6, pp.457-466, December 2013.

[11] Nanotkar, Ashish P., and Mrs Rahila Sheikh. "Intelligent Prediction of Heart Disease Using Risk Factors Based on Data Mining Techniques."

[12] Dangare, Chaitrali S., and Sulabha S. Apte. "A data mining approach for prediction of heart disease using neural networks." *International Journal of Computer Engineering and Technology (IJCET)* Issue.3, Vol.3, 2012.

[13] Waghulde, Nilakshi P., and Nilima P. Patil. "Genetic Neural Approach for Heart Disease Prediction.*" International Journal of Advanced Computer Research",* Issue-16, Vol.4, 2014.

[14] Ade, Ms RR, Dhanashree S. Medhekar, and Mayur P. Bote. "Heart disease prediction system using svm and naive bayes." *IJESRT,* Issue 2, Vol.5, pp.277-9655, May 2013.

[15] Chitra, R., and V. Seenivasagam. "Heart disease prediction system using supervised learning classifier." *Bonfring International Journal of Software Engineering and Soft Computing* Issue 3, Vol.1, pp.01-07, 2013.

