International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 01 | Jan 2020 www.irjet.net p-ISSN: 2395-0072

GENETIC ALGORITHM FOR FEATURE SELECTION TO IMPROVE HEART DISEASE PREDICTION BY SUPPORT VECTOR MACHINE

A. Sangeetha[1], Dr. B. Ananthi[2]

[1]Research Scholar, Dept. of Computer Science, Vellalar College for Women, Erode, Tamil Nadu, India
[2]Associate Professor, Dept. of Computer Science, Vellalar College for Women, Erode, Tamil Nadu, India

---

Abstract - *This research work has been developed to predict the heart disease based on feature selection and classification techniques. Heart disease prediction is the leading problem in medical data analysis. The prediction can be made from massive amount of healthcare data as information for data mining. There is only insufficient research that plays a major role in attention towards the critical features of predicting heart disease. The feature selection is essential for predicting the heart disease. The current work is implemented by Genetic algorithm (GA) for feature selection and also to focus on classifying the heart disease prediction. The best feature selection can be concluded by proper mutation results followed by best classification techniques. The classification technique performed for improves the accuracy value of the prediction. Hence, the heart disease prediction can be done in the primary stage.*

***Key Words*: Feature selection, Classification, Genetic Algorithm (GA), Support Vector Machine (SVM).**

1. INTRODUCTION

Data mining is process of extracting information from large amount of databases. Data mining is most useful in nontrivial information to the large volumes of data. Data mining is the process of extracting data for finding hidden patterns which can be modified into significant.

Heart disease is the premier cause of death in the world. Heart disease has garnered a highest deal of attention in medical research. The conclusion of heart disease is usually based on signs, symptoms and physical analysis of the patient. There are several causes that increase the risk of heart disease, such as smoking habit, body cholesterol level and family history of heart disease, obesity, high blood pressure, and insufficiency of physical exercise.

The aim of this paper is to select the correlated features or attributes of heart disease dataset by using Genetic Algorithm (GA). After the feature selection, classification techniques are applied for predict the heart disease. Then the prediction value is compared to other classification techniques.

2. LITERATURE REVIEW

HeonGyu Lee, Ki Yong Noh, and Keun Ho Ryu1 (2007), proposed multi-parametric features of HRV from ECG bio signal. The HRV indices proposed that the classification of the patients with CAD from normal people. For classification, they employed the proposed multi-parametric features, which allow them to choose a classifier from a large pool of well-studied classification methods. By considering several supervised methods including extended Naïve Bayesian classifiers (TAN, STAN), decision tree (C4.5), associative classifier (CMAR) and SVM. In experimental results, SVM showed the best performance among the tested methods [3].

MrudulaGudadhe, Kapil Wankhade, SnehlataDongre (2010), proposed a decision support system for heart disease classification based on support vector machine (SVM) and Artificial Neural Network (ANN). A multilayer perceptron neural network (MLPNN) with three layers is employed to develop a decision support system for the diagnosis of heart disease. The multilayer perceptron neural network is trained by back-propagation algorithm which is computationally efficient method. They have also proposed a decision support system for heart disease classification based on Support Vector Machine and MLP neural network architecture [6].

Rashmi G SabojiPrem Kumar Ramesh (2017), proposed a scalable solution in predicting the heart disease attributes and validated its accuracy. The implementation of random forest algorithm on Spark framework for predicting heart disease, with as small as 600 dataset records, they plan to explore other healthcare disease prediction, such as early prediction of certain types of cancer, etc. They are also planning investigate the impact of large supervised datasets at a colossal scale on performance and accuracy, running on high performance clusters [7].

Theresa Princy. J. Thomas (2016), explained that different classification techniques used for predicting the risk level of each person based on age, gender, Blood pressure, cholesterol, pulse rate. The patient risk level is classified using datamining classification techniques such as Naïve Bayes, KNN, Decision Tree Algorithm, and Neural Network. etc., Accuracy of the risk level is high when using more number of attributes. As per the analysis mode, it is seen that many authors use various technologies and different number of attributes for their study. Hence, different technologies give different precision depending on a number of attributes considered. Using KNN and ID3 algorithm the risk rate of heart disease was detected and accuracy level also provided for different number of attributes. In future, the numbers of attributes could be reduced and accuracy would be increased using some other algorithms [10].

Madhura Patil, Rima Jadhav, vishakha Patil, Aditi Bhawar, Mrs. Geeta Chillarge (2019), proposed that implementing of a system which will help to predict heart disease depending on the patients clinical data related to the factor associated with heart disease. By using medical dataset of the patients such as age, sex, blood pressure, overweight and blood sugar and by applying SVM classifier we can predict that the patients getting a heart disease or not. In addition classification accuracy, sensitivity, and specificity of the SVM have been found to be high thus making it a superior alternative for the diagnosis. They are also doing analysis on the data from which they are getting at which age it mostly occur or which region gets influenced by that disease. So precaution can be taken to avoid the death due to the heart disease [5].

3. METHODOLOGY

The current research work has series of steps to attain the objectives. The figure 1 shows the complete architecture of the heart disease prediction. The steps involved in this process are as follows,

- Dataset Collection
- Preprocessing
  - Genetic Algorithm
- Classification
  - Naïve Bayes
  - Random Forest
  - SVM

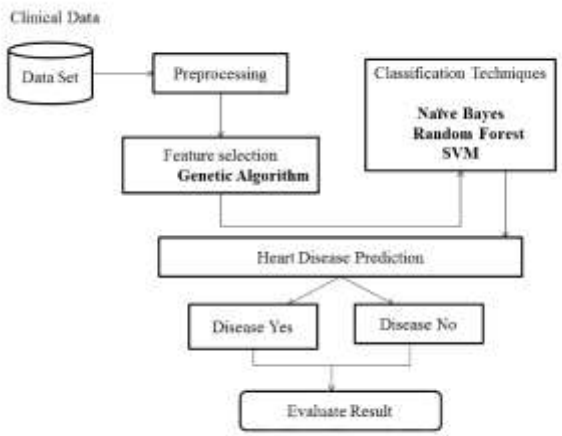


Figure 1 Architecture of heart disease prediction

A. Dataset Collection

The database for this research work has been taken from the StatLog dataset in UCI repository. It includes 14 attributes. The heart disease dataset included in this research work consists of total 270 instances with no missing values. This research work is aimed at predicting the heart disease irrelevant of the disease types.

B. Preprocessing

The collected datasets are preprocessing here by using feature selection method. The Genetic algorithm is used to feature selection and select the specific attributes.

Genetic Algorithm (GA)

The genetic algorithm parameters are crossover, elitism, mutation, and population size used in my research work. The class labels are represented in the form of binary type that is used to predict the disease as mention as 1 otherwise it will be 0. The above structure is followed by a genetic algorithm parameter specified in type category.

Here mutation rate 0.03 is default value of inversion mutation method. It is a subset of attributes to invert merely the entire string into subsets. The population size is used for retrieving the useful information and it little diversity based on random solutions. These random solutions are derived the population. The proportion of the existing population is used to create a new generation by the end of this phase, 6 features are selected from a set of 14 features in the original dataset.

C. Classification

Classification is used to predict the class labels. Here, the three classification algorithms (Naïve Bayes, Random Forest and SVM) are used to classify the data.

Naïve Bayes

To predict the class label of a tuple, the same training data is used to classify the disease. The class label attribute, disease, has two distinct values (namely, [yes, no]). In a real dataset hypothesis was tested, which gives multiple evidence (feature). Then the probabilities of disease for the dataset were calculated as shown in the equation (1).

$$P(A|B) = P(B|A)P(A)/P(B) \quad (1)$$

Where,

P(B|A) – Probability of the evidence gives that hypothesis is true (Likelihood)

P(A|B) – Probability of the hypothesis given that the evidence is there (Posterior Probability)

P(A) – Number of probabilities of disease in the data set (Class Prior probability)

P(B) – Total number of available features (Predictor Prior Probability).

Random Forest

The Random Forest classification is a collection of the decision tree classifiers. It is one of the ensemble method classifier. The Decision Tree is generated using a random selection of attributes at each node to determine the split. The popular class is returned by each tree votes during classification process [4].

Step 1: Randomly select m features from entire n features m<<n.

Step 2: Surrounded by that m features, calculate the node d using the best split point.

Step 3: Split the node into daughter nodes from the best split.

Step 4: Repeat step 1 to 3 until one number of node has been reached.

The accuracy of a random forest depends on the strength of the individual node.

Support Vector Machine

It finds a hyper plane for data separation using essential tuples. Here, the two class problem was performed by linear separable. Let the disease data set D. Corresponding to the classes disease=yes and disease-no respectively. There are infinite numbers of possible separating hyper planes. The hyper plane make the straight line that line can be separate all the tuples of class 0 from class 1. To find the best separating plane based on data points that are to be find the best hyper plane. The class labels split the two possible separating hyper planes that associated with their margins. Both hyper plane can correctly classified the given data tuples with the larger margin. The associated margins give the largest separation between the classes. A splitting hyper plane can be written as,

$$W. X + b = 0, \quad (2)$$

W is a vector, (no. of attributes)

b is a scalar,

X is a value of attributes.

The scalar b weight can be adjusted the hyper plane denoting the sides of the margin can be written as,

$$H_{1:} w_0 + w_1x_{1+} w_2x_{2>=0} \quad (3)$$

$$H_{2:} w_0 + w_1x_{1+} w_2x_{2<=1} \quad (4)$$

Training tuples that fall on hyperplane H_1 or H_2 satisfy equation (3) and (4) that is called support vectors. They are equally closed to the Maximum Margin Hyper plane [MMH].

4. RESULTS AND DISCUSSION

A. PERFORMANCE METRICS

Precision

The peoples predicted as heart diseases are TP and FP. The peoples actually having a heart disease are TP.

Precision = TP/ (TP+FP)

Recall

The Peoples having heart diseases are TP and FN. The people diagnosed by the model having a heart disease are TP.

Recall = TP / (FP+FN)

F-Measure

F-Measure is the balanced mean value of precision and recall.

F-Measure = 2 * Precision * Recall /

Precision + Recall

Analyze the performance evaluation for without feature selection (Genetic algorithm) method.

Table 1: Existing parameter values for without feature selection (Genetic Algorithm)

ALGORITHM PARAMETER	PRECISION	RECALL	F-MEASURE
NB	0.7368	0.8235	0.7775
RF	0.7894	0.7317	0.7593
SVM	0.8723	0.8723	0.8721

Analyze the performance evaluation for with feature selection (Genetic algorithm) method.

Table 2: Proposed parameter values for without feature selection (Genetic Algorithm)

ALGORITM PARAMETER	PRECISION	RECALL	F-MEASURE
NB	0.8055	0.9062	0.8528
RF	0.8421	0.8205	0.8311
SVM	0.9545	0.9130	0.9332

B. Comparison Results

Analyze the Heart Disease Prediction accuracy value is compared with NB (Naïve Bayes), RF (Random Forest), and SVM (Support Vector Machine).

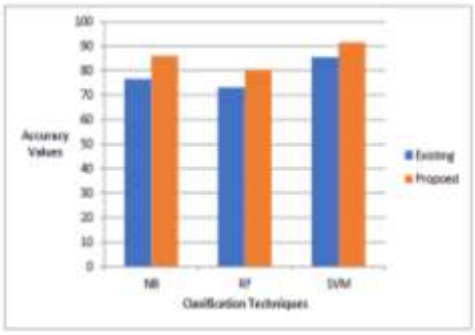
Accuracy=TP+TN/ (TP+TN+FP+FN).

Table 3: Comparison for Heart disease prediction accuracy

CLASIFICATION TECHNIQUES	ACCURACY WITHOUT GA	ACCURACY WITH GA
NB	76.47	85.29
RF	73.06	80.88
SVM	85.19	92.59

The table 3 shows the comparison between without Genetic algorithm feature selection and with Genetic algorithm feature selection.

Figure 1: Comparison graph for heart disease prediction accuracy



5. CONCLUISION AND FUTURE WORK

Heart Disease is an uncontrollable disease by its nature. This disease makes a heart attack and death. The clinical domain is concluded and steps are taken to apply apt techniques in the heart disease prediction. The heart disease prediction is implemented by classification techniques. It is help to predict heart disease depending on the patients clinical data related to the causes associated with heart disease. In existing system, heart disease prediction is analyzed the two classification techniques such as Naïve Bayes, and Random Forest. Random Forest provides better results as compare to Naïve Bayes.

In this proposed system, the Genetic Algorithm feature selection is used to reduce the attributes from the heart disease data set. The attributes will be reduced and the prediction accuracy values are increased. After the attribute selection Naïve Bayes and Random algorithms are applied for prediction and the SVM classifier algorithm are additionally used here to predict the heart disease. Compare to the Naïve Bayes, Random Forest and SVM algorithm, SVM classifier provides perfect results and high metrics values for heart disease prediction.

The future work will apply LSTM method based on deep learning techniques to classify heart disease diagnoses prediction. It formulates the problem as multi label classification and multivariate time series.

REFERENCES

1. Dhanashree S. Medhekar, Mayur P. Bote, Shruti D. Deshmukh, Heart Disease Prediction System using Naive Bayes, International Journal Of Enhanced Research In Science Technology & Engineering Vol. 2 Issue 3, March.-2013.

2. Garima Singh, Kiran Bagwe, Shivani Shanbhag, Shraddha Singh, Sulochana Devi, Heart disease prediction using Naïve Bayes, International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 03 Mar -2017.

3. HeonGyu Lee, Ki Yong Noh, and Keun Ho Ryu1, Predicting Coronary Artery Disease from Heart Rate Variability using Classification and Statistical Analysis, Seventh International Conference on Computer and Information Technology IEEE DOI 10.1109/CIT.2007.163.

4. Latha Parthiban and R.Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, World Academy of Science, Engineering and Technology International Journal of Medical and Health Sciences Vol:1, No:5, 2007.

5. Madhura Patil, Rima Jadhav, vishakha Patil, Aditi Bhawar, Mrs. Geeta Chillarge, Prediction and Analysis of Heart Disease using SVM Algorithm, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887 Volume 7 Issue I, Jan 2019.

6. Mrudula Gudadhe, Kapil Wankhade, SnehlataDongre, Decision Support System for Heart Disease based on Support Vector Machine and Artificial Neural Network, IEEE 978-1-4244-9034-/10 2010.

7. Rashmi G SabojiPrem Kumar Ramesh, A Scalable Solution for Heart Disease Prediction using Classification Mining Technique, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017) 978-1-5386-1887-5/17/2017 IEE.

8. Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr. D. P.Shukla, Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques, IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64.

9. Sheik Abdullah, R.R.Rajalaxmi, A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier, International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012) Proceedings published in International Journal of Computer Applications (IJCA).

10. Theresa Princy. J. Thomas, Human Heart Disease Prediction System using Data Mining Techniques, 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT] 978-1-5090-1277-0/16/2016 IEEE.