An Ensemble Classifier Characterized by Genetic Algorithm with Decision Tree for the Prophecy of Heart Disease

K. Chandra Shekar, Priti Chandra and K. Venugopala Rao

Abstract The prediction of heart disease is critically significant for diagnosis of diseases and treatment. The data mining techniques that can be applied in medicine, and in particular some machine learning techniques including the mechanisms that make them better suited for the analysis of medical databases. Extensive amounts of data gathered in medical databases require specialized tools for storing and accessing data, for data analysis, and for effective use of data. In particular, the increase in data volume causes great difficulties in extracting useful information and also consumes time for decision support. Intuitively, this large amount of stored data contains valuable hidden information, which could be used to improve the decision making process of an organization. For prediction of heart disease, many researchers have used some machine learning algorithms like Bayesian Classification, Neural Networks, Support Vector Machines, and K-nearest neighbor algorithms. We propose a hybrid technique of ensemble classifier to provide a better solution for the classification problem. The output from this hybrid scheme gives the optimized feature. This output is then given as the input to the decision tree classifier for predicting the occurrence and possibly obtaining the type of heart disease. Here, the features are initialized through the decision tree and fitness is evaluated via genetic algorithm.

Keywords Ensemble classifier · Genetic algorithm · Decision tree · Fitness function · Selection · Crossover · Mutation · Reproduction

K. Chandra Shekar (✉)
JNTUH, Hyderabad, Telangana, India
e-mail: chandhra2k7@gmail.com

P. Chandra
ASL, DRDO, Hyderabad, India
e-mail: priti_murali@yahoo.com

K. Venugopala Rao
GNITS, Hyderabad, India
e-mail: kvgrao1234@gmail.com

H. S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*, Lecture Notes in Networks and Systems 74,
https://doi.org/10.1007/978-981-13-7082-3_2

1 Introduction

In recent years, data mining has been extensively used in the areas of bioinformatics, science and engineering, genetics, and medicine [1]. Data mining is an interdisciplinary field of study in databases, machine learning, and visualization. Data mining is the research domain which deals with discovering the relationships and global patterns that exist hidden among large amounts of data [2, 3]. Most of the healthcare organizations are facing a major challenge of providing quality services to their patients like accurate automated diagnosis and administering treatment at affordable costs [4]. Data mining helps in identifying the patterns from successful medical case sheets for different illnesses and it also aims to find knowledge which is useful for the diagnostics [5]. It is a collection of various techniques and algorithms, through which we can extract informative patterns from raw data [6]. It plays a vital role in tackling the data overload in medical diagnostics. Data mining technology provides a deep insight providing a user oriented approach to discover novel and hidden patterns in the data. This helps in evaluating the effectiveness of medical treatments [7]. The data generated by healthcare transactions is enormous. This medical data containing patients' symptoms is analyzed to perform medical research [8].

With the development of information technology, extensive medical data is available. Medical data classification plays a significant role in various medical applications [9–11]. Medical classification can be widely used in hospitals for the statistical analysis of diseases and therapies [12, 13]. It addresses the problems of diagnosis, analysis and teaching purposes in medicine [14–16]. Medical data has made a great progress over the past decades in the development and use of classification algorithms [17–19]. In healthcare, these medical data can be transformed into aggregations to calculate average values per patient and compare with ranges/other values, to group data into clusters of similar data, etc. [20–22].

Ensemble Methods are the methods that use a combination of models to improve classifier and predictor accuracy. Bagging and Boosting are the two such general strategies. According to the Wolpert's no free lunch theorem, a classifier may perform well in few specific domains, but never in all application domains. Therefore, by combining the outputs of multiple classifiers, the ensemble of classifiers strategically extends the power of aggregated method to achieve better prediction accuracy.

2 Related Work

Akhil Jabbar [1] had proposed an algorithm which combines K-Nearest Neighbor with genetic algorithm for effective classification. Muthukaruppan et al. [23] had proposed particle swarm optimization (PSO), which is based on fuzzy expert system involving four stages. Lahsasna et al. [24] proposed a fuzzy rule-based system (FRBS) to serve as a decision support system for Coronary heart disease (CHD) diagnosis that not only considers the decision accuracy of the rules but also their transparency at

the same time. Yilmaz et al. [25] had presented a new data preparation method based on clustering algorithms for the diagnosis of heart and diabetes diseases. Kim et al. [26] had proposed a Fuzzy Rule-based Adaptive Coronary Heart Disease Prediction Support Model (FbACHD_PSM), which gives content recommendation to coronary heart disease patients.

3 Proposed Method

The proposed methodology integrates with supervised machine learning technique which is based on a hybrid approach for providing a better decision system using dual decision tree and genetic algorithm. Genetic algorithms are one of the best methods for search and optimization problems.

A decision tree is a tree structure classifier that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label.

Pros of Decision Trees (DTs):

- DTs do not require any domain knowledge.
- DTs are easy to comprehend.
- The learning and classification steps of a DT are simple and fast.

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex. Tree Pruning can be done through two approaches:

- Pre-pruning—The tree is pruned by halting its construction early.
- Post-pruning—This approach removes a sub-tree from a fully grown tree.

The cost complexity of a decision tree is measured by two parameters, the number of leaves in the tree and the error rate of the tree.

Genetic algorithms (GA) were invented by John Holland in 1975. Genetic algorithms can be applied for search and optimization problems. GA uses genetics approach as its model for problem solving. Each solution in genetic algorithm is represented through chromosomes. Chromosomes are made up of genes, which are individual elements that represent the problem. The collection of all chromosomes is called the population [1, 27].

In general, there are three operators that can be applied in GA.

(1) **Selection**:
This operator is used in selecting individuals for reproduction with the help of fitness function. Fitness function in GA is the value of an objective function for its phenotype. The chromosome has to be first decoded, for calculating the fitness function.

(2) **Crossover**:
This is the process of taking two parent chromosomes and producing a child from them. This operator is applied to create better string.

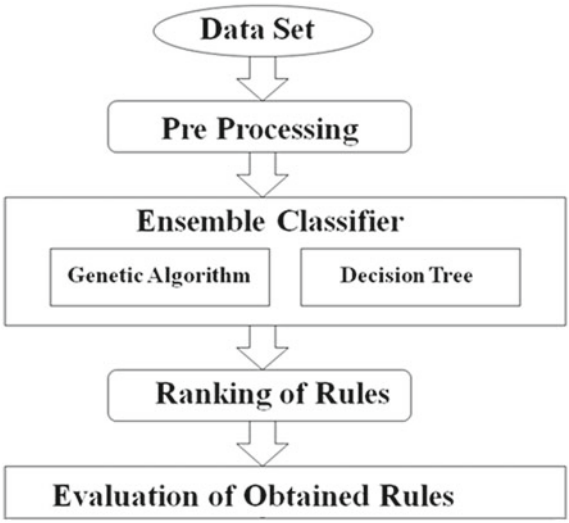

Fig. 1 Proposed system architecture for an ensemble classifier characterized by genetic algorithm with dual decision tree

(3) **Mutation**:
This operator is used to alter the new solutions in the search for better solution. Mutation prevents the GA to be trapped in a local minimum.

The proposed system architecture (Fig. 1) consists of an ensemble classifier characterized by genetic algorithm with dual decision tree facilitates as follows, in the first stage multiple risk factors such as age, hypercholesterolemia, hypertension, diabetes, obesity, stress level, alcohol taken, etc., are taken as input. This input is preprocessed to fill up the missing values, remove noise and inconsistencies if any in the data and then is given to the hybrid scheme which consists of genetic algorithm and decision tree. Here, the features are initialized through decision tree and fitness is evaluated via genetic algorithm. The output from this hybrid scheme gives the optimized feature. This output is then given as the input to the decision tree classifier for obtaining the type of heart disease.

In decision tree, both training and testing phases are carried out. In the training phase, a classifier known as iterative dichotomizer or random forest classifier can be utilized. This classifier makes use of number of decision trees at training stage in order to enhance classification rate. This random classifier contains two steps namely oob (out-of-bag) and permutation to avoid classification error and to measure the importance of variable. This classifier has a combined group of techniques to process such as randomized node optimization, bagging and CART model. In random optimization algorithm, the best tree model is given as output and in bagging it repeatedly selects the random sample with the replacement of the training set and fit trees. After that CART (classification and regression) is done for recognizing the type of attack and finally the output is displayed using a tree structure. The output displays the type of heart attack for the patient to occur. This can be determined in the classification step by comparing the information stored in the database. If

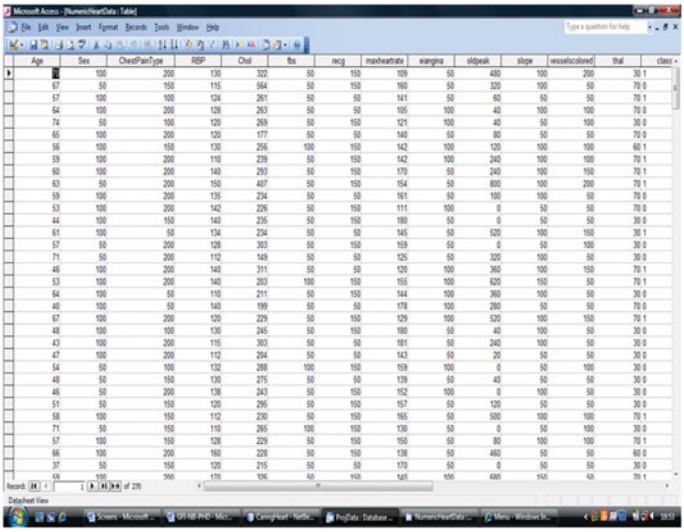

Fig. 2 Heart disease dataset

there is any type of attack possibility is predicted then it will show the prediction by percentage value by the utilization of the regression method. Decision tree has four major advantages for predictive analytics namely it implicitly performs feature selection, it needs relatively very less effort from users for data preparation, the nonlinear relationships between parameters do not affect tree performance, and it is very simple to explain.

In our proposed hybrid technique, we can predict the accurate type of heart attack and optimum feature selection for reducing dimensionality, training time, and overfitting. The proposed methodology can be implemented using MATLAB platform and the experimental results can be analyzed and compared with the conventional methods.

4 Results and Discussions

The experimental results attained from the proposed method are compared with the existing methods in terms of classification accuracy and time complexity with respect to the heart disease dataset (Fig. 2).

The proposed approach generates the optimized features through genetic algorithm. The classification accuracy is higher when compared with the existing methods (Table 1).

The reduction of time complexity is expected due to the optimization performed on the features (Table 2).

Table 1 Accuracy analysis

Classification method	Accuracy (%)
Random subspace	78.91
Decision tree	66.67
Multilayer perceptron	79.25
Proposed approach	85.37

Table 2 Accuracy analysis

Classification method	Time complexity (s)
Random subspace	0.18
Decision tree	0.17
Multilayer perceptron	27.88
Proposed approach	0.12

Thus, it is natural to realize the efficient of the proposed approach as the accuracy has increased and the time complexity has reduced significantly.

5 Conclusion and Future Directions

Majority of the health care organizations are facing a severe challenge in the provision of quality services like diagnosing patients correctly and administering treatment at reasonable costs. Data mining helps to identify the patterns of successful medical therapies for different illnesses and also it aims to find useful information from large collections of data. With the development of information technology, extensive medical data is available. Medical data classification plays an essential role in most of the medical applications. Ensemble Methods are the methods that use a combination of models, to improve classifier and predictor accuracy. The purpose of this research is to enhance performance of the heart disease prediction system by avoiding mis-prediction rate. We further plan to compare the performance of our ensemble classifier with the existing and traditional classifiers. We would also like to move towards hybrid generic intelligent systems to further improve the predictive accuracy.

References

1. Akhil Jabbar M, Deekshatulu BL, Chandra P (2013) Classification of heart disease using K-nearest neighbor and genetic algorithm. Procedia Technol 10:85–94
2. Karaolis MA et al (2010) Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans Inf Technol Biomed 559–566
3. Oztekin A, Delen D, Kong ZJ (2009) Predicting the graft survival for heart–lung transplantation patients: an integrated data mining methodology. Int J Med Inform e84–e96

4. Kurt I, Ture M, Turhan Kurum A (2008) Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl 366–374
5. Tsipouras MG et al (2008) Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Trans Inf Technol Biomed 447–458
6. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 7675–7680
7. Kahramanli H, Allahverdi N (2008) Design of a hybrid system for the diabetes and heart diseases. Expert Syst Appl 82–89
8. Huang Y et al (2007) Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med 251–262
9. Ramon J et al (2007) Mining data from intensive care patients. Adv Eng Inform 243–256
10. Cho BH et al (2008) Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artif Intell Med 37–53
11. Sarkar BK, Sana SS, Chaudhuri K (2012) A genetic algorithm-based rule extraction system. Appl Soft Comput 238–254
12. Karaboga D, Ozturk C (2011) A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Appl Soft Comput 652–657
13. Wright A, Chen ES, Maloney FL (2010) An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 891–901
14. Chandra Shekar K, Sree Kanth K, Ravi Kanth K (2012) Improved algorithm for prediction of heart disease using case based reasoning technique on non-binary datasets. Int J Res Comput Commun Technol 1(7)
15. Deepika N, Chandra Shekar K, Sujatha D (2011) Association rule for classification of heart-attack patients. Int J Adv Eng Sci Technol 11(2):253–257
16. Polat K, Güneş S (2007) An expert system approach based on principal component analysis and adaptive neuro fuzzy inference system to diagnosis of diabetes disease. Digit Signal Process 702–710
17. Yap BW, Ong SH, Mohamed Husain NH (2011) Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst Appl 13274–13283
18. Özçift A (2011) Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Comput Biol Med 265–271
19. Pal D et al (2012) Fuzzy expert system approach for coronary artery disease screening using clinical parameters. Knowl-Based Syst 162–174
20. Seera M, Lim CP (2014) A hybrid intelligent system for medical data classification. Expert Syst Appl 2239–2249
21. Acharya UR et al (2013) Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. Comput Methods Programs Biomed 624–632
22. Exarchos TP et al (2007) A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree. Artif Intell Med 187–200
23. Muthukaruppan S, Er MJ (2012) A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Syst Appl 11657–11665
24. Lahsasna A et al (2012) Design of a fuzzy-based decision support system for coronary heart disease diagnosis. J Med Syst 3293–3306
25. Yilmaz N, Inan O, Uzer MS (2014) A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases. J Med Syst
26. Kim J-K et al (2014) Adaptive mining prediction model for content recommendation to coronary heart disease patients. Clust Comput 881–891
27. Vijay Bhasker G, Chandra Shekar K, Lakshmi Chaitanya V (2011) Mining frequent itemsets for non binary data set using genetic algorithm. Int J Adv Eng Sci Technol 11(1):143–152