Enhancing Heart Disorders Prediction with Attribute Optimization

Sushruta Mishra, Anuttam Dash, Piyush Ranjan, and Ajay Kumar Jena

Abstract The wide application of machine learning in dominant domains such as marketing, telecommunication, agriculture, and other industries has made an impact on its use in several other time critical applications. Health care is one of the vital sector where machine learning is finding acceptance in disease diagnosis. Though the medical zone is rich in raw information, but somehow not all information are successfully extracted that is needed to disclose uncertain trends & efficient decision making. Extraction of these uncertain patterns and associations usually turns unexploited. Modern optimization methodologies may be helpful in dealing with this scenario. In this research work, it is intended to use classification-based modelling algorithms which include Naïve Bayes, decision trees, artificial neural network (ANN), and support vector machine (SVM) with the use of health-related attributes like age, gender, level of blood pressure, and blood sugar, it can be used in predicting the probability of patients inheriting various disorders related to heart. Eventually, genetic algorithm is used as a feature optimizer which extracts the relevant attributes for classification. It is observed that with the use of genetic algorithm, the classification performance is enhanced with the implementation of the above classifiers.

Keywords Naïve bayes · Artificial neural network · Genetic algorithm · Data mining · Classification accuracy

S. Mishra (✉) · A. Dash · P. Ranjan · A. K. Jena
School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
e-mail: mishra.sushruta@gmail.com

A. Dash
e-mail: anuttam.dash@gmail.com

P. Ranjan
e-mail: piyushupadhaya22@gmail.com

A. K. Jena
e-mail: ajay.bbs.in@gmail.com

P. K. Mallick et al. (eds.), *Advances in Electronics, Communication and Computing*,
Lecture Notes in Electrical Engineering 709,
https://doi.org/10.1007/978-981-15-8752-8_14

1 Introduction

The process of extracting out the patterns, trends, and structures within the data available in pre-existing databases and thereby using them to build statistical models for generating information from them is known as data mining. In order to extract hidden patterns and structures within the data, it uses several disruptive technologies like statistical learning, machine learning, database technologies, etc. According to the report of World Health Statistics 2012 report, one in every three adults across the world has high blood pressure condition which is responsible for triggering almost half of the total number of deaths caused due to heart disease. Cardiovascular disease (CVD) is the major reason which generally influences not only the heart attack but also heart. Heart, one of the important muscular organs facilitates pumping of blood in the entire human body. Stroke, coronary heart disease, and cardiovascular diseases are the widely known examples of heart diseases. A condition of human heart caused due to contraction and hardening of the blood vessels connected to the brain or also sometimes by high blood pressure [1, 2] is referred to as a stroke. A lot of casualties are reported only due to heart diseases not only in India but also in other countries. In USA, an individual at every 34 s dies due to a heart disease. Some types of heart diseases are cardiomyopathy, coronary heart disease, and cardiovascular disease. Cardiovascular disease affects the heart by disturbing the routine of blood circulation in the human body. Thus, there is a need for efficient and accurate diagnosis. The doctor's experience and knowledge are determinant of the accuracy of the diagnosis of the disease, it might lead to unwanted treatment costs in some cases. Thus, there is a need for an automatic medical diagnosis system. This paper is an effort to elaborate various information extraction methods that may be used in these intelligent model-based systems.

2 Literature Survey

A Web-based automated system which uses different kinds of health-related attributes like age, gender, blood pressure level, blood sugar level, etc., which help in predicting occurrence of a heart disorders in an individual [3]. The complex what-if queries are addressed by the Naïve Bayes algorithm. The platform in which the system has been made is PHP which further makes it flexible and expandable. Hybrid approaches are found to be more efficient and accurate than a single model in terms of predicting heart diseases which is again confirmed by survey [4]. The data mining techniques that have been evaluated by the author using accuracy and sensitivity [5] as measures are namely artificial neural networks (MLP), Naïve Bayes, decision tress (C4.5). We observed that a greater number of attributes result in a better performance of Naïve Bayes, artificial neural networks than decision trees. The efficiency of different algorithms of decision trees like C5.0, ID3, C4.5, and J48 in predicting the different kinds of heart disorders has been identified [6]. ID3 mainly aims to construct decision trees

by using a definite number of instances of training. J48 decision tree has been built on top of ID3 algorithm. The latest version of ID3 algorithm is C4.5 to which C5.0 is an extension. Attribute selection measure results in the split criteria: Information gain can be complimented by the algorithms performance: K-NN, neural network, Naïve Bayes, decision tree which is used in the heart disease samples [7].Some of data mining techniques utilized for diagnosing heart diseases are Simple Cart, Bayes Net, J48, Naive Bayes and REPTREE [8]. We have used classifiers like decision tree, Naïve Bayes, classification by clustering [9] to identify heart diseases. To reduce the features from 13 to 6, we have used genetic model. It was observed that among the three classifiers: Naive Bayes, Simple Cart and REPTREE, Naive Bayes is found to be the best. By using clustering and integration of feature subset selection and with high construction time, a better attribute reduction with considerably same construction time has been achieved by Naive Bayes as compared to the other two classifiers. We have intended to broaden our work to predict the severity of the disease by using fuzzy numerous techniques of data mining such as association rules, cluster analysis, classification, fuzzy systems [10].

3 Data Source

To predict the possibility of heart disease in an individual, we use health profiles like age, blood pressure, blood sugar, sex, etc. It helps in the exploration of different important aspects like patterns and structures within the health factors related to heart disease. We have used open-source heart disease database to predict the possibility of heart disease in an individual based off of his health attributes as shown in Table 1.

4 Proposed Work

The proposed work as illustrated in Fig. 1 constitutes the significance of feature optimization on the classification performance of heart disease prediction. Here, classifiers used are decision tree (DT), artificial neural networks (ANN), Naïve Bayes (NB), support vector machine (SVM). The heart disease data samples are gathered from UCI repository and the anomalies are eliminated. It is divided into the respective training and testing set of data. On the one hand, the input data is directly subjected to classification through the above classifiers and their performance is determined by using classification rate metric. On the other hand, the dataset is implemented using genetic algorithm as the attribute optimizer. It identifies and removes the less relevant attributes and the result is an optimized dataset. This optimized data is applied to classifiers for performance analysis. It is observed that genetic algorithm optimizes the heart disease dataset and the performance of classification is enhanced.

A popular classifier that is quite simple and easy in terms of implementation is decision tree. High-dimensional data can be easily handled with no requirement of

Table 1 Heart disease dataset

Attribute	Description
Age	Age in years
Sex	Sex (1 = male; 0 = female)
cp	Chest pain type
trestbps	Resting blood pressure
chol	Serum cholestoral in mg/dl
fbs	(fasting blood sugar >120 mg/dl)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
exang	Exercise induced angina
oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	Number of major vessels (0–3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect
num	Diagnosis of heart disease

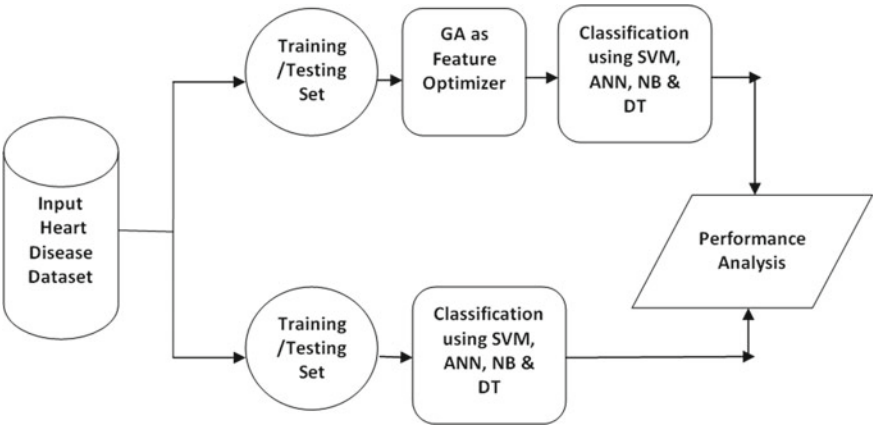

Fig. 1 Proposed system model for heart disease prediction

domain or parameter knowledge. The results produced are very easy to interpret and readable. Decision trees are used to go through the features to analyze the patient's profile. Naïve Bayes is a classifier model that considers zero dependency among attributes. Naive Bayes assumes that the value of attributes in a given class is independent of other attribute values which is also known as conditional independence. We do not require any Bayesian methods in order to work with Naïve Bayes. This is also considered to be the advantage of Naïve Bayes.

Genetic Algorithm [5] is a natural evolutionary methodology. The genetic algorithm is a population-based search which starts initially with zero number of attributes and an initial population which is created by random generation rules. It is based on the idea of reproduction, natural selection, and survival of the fittest. Parents produce off-springs using genetic cross-over, mutation, and selection. The process continues to a point where it produces a population P with evolution where each and every rules of P is satisfied by the fitness value threshold. We take initial population size of 20 instances, probability of cross-over as 0.6, and probability of mutation as 0.033 and the process continues for twentieth generation. We get a total of six attributes from an initial 13 number of attributes by using genetic algorithm. Finally, after getting six attributes from initial 13 attributes by attribute reduction, we use different classifiers on the dataset to predict the heart disease.

5 Results Analysis

Here, various classification algorithms are presented and implemented for determining classification performance of heart disease data samples. The first analysis comprises of using classifiers alone without using any attribute optimization method like genetic algorithm. It is observed that classification with artificial neural network gives the best classification accuracy rate of 89.7% while support vector machine produces the least accuracy rate of 84.2% as shown in Fig. 2. In the second analysis as shown in Fig. 3, genetic algorithm is used as an attribute optimization tool. Here, again it is observed that using genetic algorithm with artificial neural network produces the optimal classification performance with 90.6% accuracy rate. Also with support vector machine, it is seen that the classification efficiency is the minimum with 87.8%.

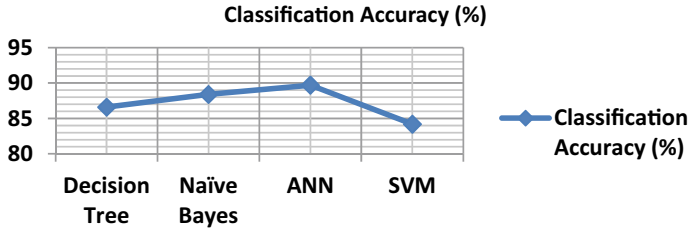

Fig. 2 Classification without attribute optimization

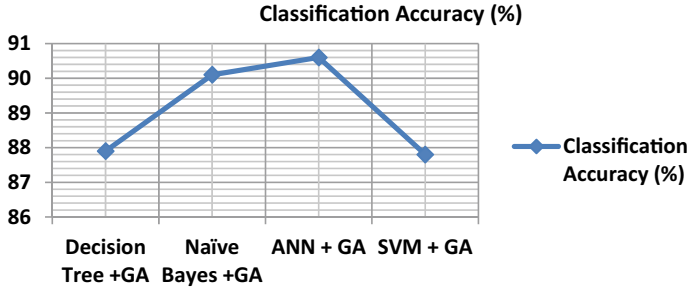

Fig. 3 Classification with genetic algorithm as attribute optimization

6 Conclusion

The objective of this paper is to discuss the different techniques of data mining that could be used to predict the presence of heart disease. Analysis suggests that different technologies that use different number of attributes reach results of varying accuracy. The accuracy depends on tools used for implementation. Additionally, genetic algorithm has been used as the attribute optimization agent. We observed that the use of genetic algorithm is helpful in optimization of classification performance of heart diseases. Although there are evidences of successful application of techniques of data mining to help the healthcare professionals or doctors to diagnose heart diseases, this paper provides a faster and simpler data mining model. The model has been developed by thoughtful analysis of different prediction models in data mining with the objective of finding the greatest model for further work.

References

1. Dent, T. (2010). Predicting the risk of coronary heart disease. PHG foundation publisher.
2. World Health Organization. (2010). Global status report on no communicable diseases..
3. Ratnam, D., HimaBindu, P., MallikSai, V., Rama Devi, S. P., & Raghavendra Rao, P. (2014). computer-based clinical decision support system for prediction of heart diseases using naïve bayes algorithm. *International Journal of Computer Science and Information Technologies, 5*(2), 2384–2388.
4. Purusothaman, G., & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. *Indian Journal of Science and Technology, 8*(12).
5. Srinivas, K., Raghavendra Rao, G., & Govardhan, A. (2010). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In *Proceedings of 5th International Conference on Computer Science & Education* (pp. 24–27). China.
6. Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science, 2*(6), 6–11.

7. Peter, J., & Somasundaram, K. (2012). An empirical study on prediction of heart disease using classification data mining techniques. In *Proceedings of IEEE International Conference on Advances In Engineering, Science and Management (ICAESM)*, pp. 514–518.
8. Masethe, H. D., & Masethe, M. A. (2014). Prediction of heart disease using classification algorithms. In *Proceedings of the World Congress on Engineering and Computer Science (WCECS)*, San Francisco, USA.
9. Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology, 2*(10), 5370–5376.
10. Vijayarani, S., & Sudha, S. (2012) A study of heart disease prediction in data mining. *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, *2*(5), 2249–9555.