Available online at www.banglajol.info

Bangladesh J. Sci. Ind. Res. **50(4)**, 293-296, 2015

BCSIR

BANGLADESH JOURNAL
OF SCIENTIFIC AND
INDUSTRIAL RESEARCH

E-mail: bjsir07@gmail.com

A comparison of three discrete methods for classification of heart disease data

D. Chaki[1]*, A. Das[1] and M. I. Zaber[2]

[1]*Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh.*
[2]*Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh.*

Abstract

The classification of heart disease patients is of great importance in cardiovascular disease diagnosis. Numerous data mining techniques have been used so far by the researchers to aid health care professionals in the diagnosis of heart disease. For this task, many algorithms have been proposed in the previous few years. In this paper, we have studied different supervised machine learning techniques for classification of heart disease data and have performed a procedural comparison of these. We have used the C4.5 decision tree classifier, a naïve Bayes classifier, and a Support Vector Machine (SVM) classifier over a large set of heart disease data. The data used in this study is the Cleveland Clinic Foundation Heart Disease Data Set available at UCI Machine Learning Repository. We have found that SVM outperformed both naïve Bayes and C4.5 classifier, giving the best accuracy rate of correctly classifying highest number of instances. We have also found naïve Bayes classifier achieved a competitive performance though the assumption of normality of the data is strongly violated.

Keywords: Classification; Comparison; C4.5 Classifier; Naïve bayes classifier; SVM Classifier; Heart disease data

Introduction

Now-a-days, the number of people suffering from heart disease is increasing drastically. According to the World Health Organization (WHO) causes of death summary tables 2008, the total number of deaths due to cardiovascular disease has reached to 17.3 million in a year (World Bank, 2008). However, precise diagnosis at an initial phase followed by appropriate treatment can save huge amount of lives(Yan *et al.,* 2003). Unfortunately, correct diagnosis of heart disease at a primary phase is quite a challenging task because of complex interdependence on various factors (Yan *et al.,* 2003). Hence, there is a demanding need to develop medical diagnosis systems in such a way that can assist medical practitioners in the diagnostic process.

Precise prediction of risk factors which are associated with cardiovascular disease is critically important for the diagnosis and treatment of heart disease. In order to acquire appropriate information from the databases, biologists are using up-to-date machine learning techniques enormously. Among the existing techniques, supervised learning methods are the most popular in heart disease diagnosis (Kumaravel *et al.,* 1996). Statistical analysis has identified some risk factors related with heart disease to be age, blood pressure, smoking habit (Heller *et al.,* 1984), total cholesterol (Wilson *et al.,* 1998), diabetes (Simons *et al.,* 2011), hypertension, family history of heart disease (Din *et al.,* 2007), obesity and lack of physical activity (Shahwan, 2010). Various data mining techniques have been used by the researchers to aid medical practitioners through better accurateness in the diagnosis of heart disease. Decision Tree, Naïve Bayes, Neural Network, Genetic Algorithm, Support Vector Machine, and direct kernel self-organizing map are some techniques used so far in the diagnosis of heart disease (Shouman *et al.,* 2011).

In this paper, we have shown a comparison of three discrete classifiers that may be used in machine learning, namely the naïve Bayes algorithm, the C4.5 decision tree and the Support Vector Machine (SVM) classifier. Our study was inspired by the need to discover an automated method to find the most suitable machine learning technique for predicting survivability rate of heart disease patients. We used naïve Bayes, C4.5 and SVM keeping into account the high non-normality of our data (Fig. 1 and Fig. 2). Here, Fig. 1

*Corresponding author: E-mail: dipankar@bracu.ac.bd

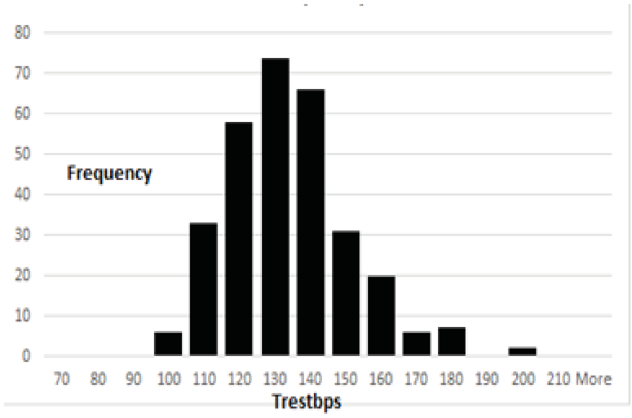

Fig. 1. Histogram of variable Trestbps

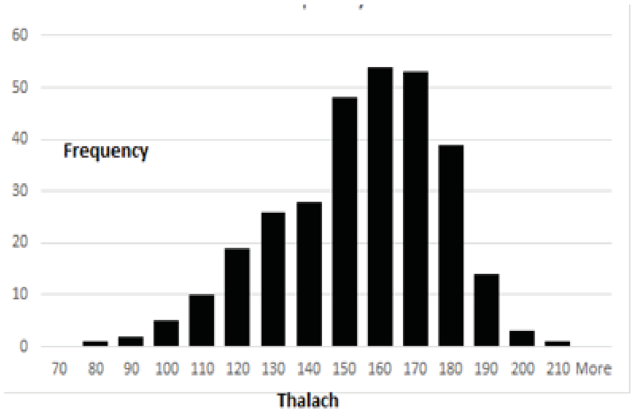

Fig. 2. Histogram of variable Thalach

Table I. Selected cleveland heart disease data set attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non -anginal pain 4 = asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic result: 0 = normal 1 = having ST-T abnormality 2 = showing probable or define left ventricular hypertrophy by Estes'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 to 3
Thal	Discrete	3 = normal 6 = fixed defect 7 = reversible defect
Diagnos is (num)	Discrete	Diagnosis classes: 0 = healthy 1 = patient who is subject to possible heart disease

represents the histogram of variable trestbps and Fig. 2 represents the histogram of variable thalach. Basically, these two variables are two attributes of our data set. And these two figures state that our data set has non-normal distribution. For this reason, we started using the C4.5 and the naïve Bayes classifiers and then we compare results with the SVM. Astonishingly, we have found that the naïve Bayes classifier does perform better than the C4.5, although the assumption of normality of the data is strongly violated. However, we have obtained that SVM outperformed both naïve Bayes and C4.5 classifier, giving the best accuracy rate of correctly classifying highest number of instances.

Materials and methods

C4.5 forms decision trees from a set of training data. It uses the concept of Information Entropy. To make a decision that splits the data into smaller subsets, each attribute of the data is used. C4.5 examines the difference in entropy that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then recurs on the smaller sub-lists.

A Bayesian classifier is a fast-supervised classification technique and this is the appropriate classifier for extensive prediction and classification tasks on composite and incomplete data sets. Naïve Bayesian classification works better when the attributes' values for the sessions are self-determining. The naïve Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f(x) can take on any value from same finite set V (Rish, 2011).

Support Vector Machine (SVM) is a class of supervised learning algorithms first introduced by Vapnik (Vapnik, 1995). Given a set of training samples, each marked for belonging to one or two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the samples as points in space are mapped in such a way that the samples of the distinct categories are separated by a clear gap (i.e., as wide as possible) (Vapnik, 1995). The data used in this study is the Cleveland Clinic Foundation Heart Disease Data Set available at UCI Machine Learning Repository (Lichman, 2013 and Detrano, 1988). This data set has 76 raw attributes, but all published experiments refer to using a subset of 14 of them. In particular, Cleveland data set is the only one that has been used Machine Learning researchers to this date. Consequently, to allow comparison with the literature, we restricted testing to these 13 attributes and 1 goal attributes which are listed in Table 1. The "goal" field refers to the presence of heart disease in the patient. The data set consists of 13 numeric attributes including age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, oldpeak, slope, number of vessels coloured and thal. The classes comprise of integers valued 0 (no presence of heart disease) and 1 (presence of heart disease).

Results and discussion

For clustering analysis, we had used the data set which contains 303 instances of patients and a panel of 14 attributes including the target attribute. There is basically two categories of population among 303 cases. One of them are healthy and the number of this class is 165. The other one is the patient who is subject to possible heart disease and they are in number of 138.

After loading our data in the WEKA software (Witten and Frank, 2005), we chose the C4.5 classifier algorithm. As it can handle continuous attributes, there was no need to discretize any of the attributes and in our experiments, we accepted the default values for the parameters. We chose to run the classifier 10 times using the 10-fold cross validation option and evaluate the accuracy of the obtained classification simply by looking at the percentage of the corrected classify instances. Later, we used the same 'initial conditions' and repeated the experiments for the same number of times for running other classifiers. After that, we observed the returning results and the results we got were quite good. Precisely, we obtained 235 cases correctly classified (77.56%) and 68 (22.44%) incorrectly classified.

Later, we considered the naïve Bayes classifier and again we used the same default parameters which is based on the assumption that numeric attributes are conditionally independent. This method performed better than the previous one. It classified properly 253 instances (83.5%) out of 303 instances; 50 instances were misclassified (16.5%). However, one must be aware that naïve Bayes relies on two fundamental assumptions: the first one is the complete independence of features and the second is that the attributes should follow a normal distribution, which is not always true (Soria *et al.,* 2008). Considering the later assumption, it can be easily state that our data does not have a normal distribution. Though the violation of its assumption, the naïve Bayesian classifier is strangely effective on our data set, showing a good performance.

Table II. Comparison of results

Method	Classified	Misclassif ied
C4.5	235 (77.56%)	68 (22.44%)
Naïve Bayes	253 (83.5%)	50 (16.5%)
SVM	255 (84.12%)	48 (15.84%)

We finally applied the SVM classifier using the same default parameters leaving kernel as polykernel and tolerance parameter as 0.001. The default sequential minimal optimization algorithm was used. Comparison of alternative learning algorithms is outside the scope of this study. This method outperformed the other two significantly.

Of the 303 cases, 255 (84.12%) were classified perfectly using this method; just 48 cases (15.84%) were misclassified. The result is summarised in Table II.

Conclusion

In this paper, we studied three different machine learning techniques and used them over a novel data set of heart disease for classification. From our experiments, we obtained different results for each of them. Using the whole dataset (14 attributes × 303 instances), we got the best performance from the SVM classifier: in fact only 48 cases were incorrectly classified. The naïve Bayes and C4.5 returned similar results but worse than the SVM.

From the results, it can be stated that all classifiers achieved a reasonable performance. However, we found that, SVM performed significantly better than both C4.5 and naïve Bayes classifier on our data set.Future research involves more intensive testing using a larger heart disease database to get more accurate results.

References

Detrano R (1988), Heart Disease Data Set, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Din S, Rabbi F, Qadir F and Khattak M (2007), Statistical Analysis of Risk Factors for Cardiovascular Disease in Malakand Division, *Pakistan Journal of Statistics and Operation Research* **3:** pp 107-110.

Heller RF, Chinn S, Pedoe HD and Rose G (1984), How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Preventin Project, *British Medical Journal* **288:** pp 1409-1411.

Kumaravel N, Sridhar KS and Nithiyanandam N (1996), Automatic diagnoses of heart diseases using neural network. *Proceeding of the 15th Biomedical Engineering Conference.* pp 319-322.

Lichman M (2013), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Rish I (2001), An empirical study of the naïve Bayes classifier, *IJCAI 2001 workshop on empirical methods in artificial intelligence* **3:** pp 41-46.

Shahwan L (2010), Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne, *International Journal of Research in Nursing* **6:** pp 1-7.

Shouman M, Turner T and Stocker R (2011), Using Decision Tree for Diagnosing Heart Disease Patients. *Proceedings of the 9th Australasian Data Mining Conference.*

Simons LA, Simons J, Friedlander Y, McCallum J and Palaniappan L (2011), Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study, *Medical Journal of Australia* **178:** pp 113-116.

Soria D, Garibaldi JM, Biganzoli E and Ellis IO (2008), A Comparison of Three Different Methods for Classification of Breast Cancer Data.*Proceedings of the 7th International Conference on Machine Learning and Applications*.

Vapnik VN (1995), The nature of statistical learning theory. *New York: Springer*.

Wilson PWF, D'Agostino RB (1998), Prediction of Coronary Heart Disease Using Risk Factor Categories, *American Heart Association Journal* **97**(18): 1837-1847

Witten IH and Frank E (2005), Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, *Morgan Kaufmann Publishers*, San Francisco, USA.

World Bank (2012), Health statistics and information systems, Retrieved from http://www.who.int/healthinfo/global_burden_disease/estimates.

Yan H, Zheng J, Jiang Y, Peng C and Li Q (2003), Development of a decision support system for heart disease diagnosis using multilayer perceptron, *IEEE Symposium on Circuits and Systems* **5:** pp 709-712.

Received: 8 September 2015; Revised: 29 September 2015;

Accepted: 19 October 2015.