Research Journal of Applied Sciences, Engineering and Technology 8(11): 1350-1354, 2014
DOI:10.19026/rjaset.8.1106
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 14, 2014 Accepted: July 19, 2014 Published: September 20, 2014

Research Article

A Methodology for Heart Disease Diagnosis Using Data Mining Technique

R. Kavitha and E. Kannan
Department of CSE, Vel Tech Rangarajan Sagunthala R and D Institute of Science and Technology (Vel Tech RR and SR Technical University), Chennai-62, Tamil Nadu, India

Abstract: Heart Disease diagnosis is done typically by doctor's knowledge and training. But even then patients are requested to take more number of medical tests for diagnosis, in which all the tests does not contribute towards effective diagnosis of heart disease. There are nearly 15 attributes which are involved in the heart disease diagnosis process. The objective of this study is to identify the key patterns and feature subsets from the heart disease data set using the Naive Bayes classifier model. The proposed system identifies feature subsets of critical data instances in data sets. It identifies and removes the redundant attribute and inter correlated attribute. The 15 is reduced to 5 attribute using our diagnosis approach by which we can naturally reduce the computational time and cost of the process. In our proposed work we also find the critical nugget. Critical Nugget is a small collection of records or instances that contain domain-specific important information. It helps to reduce the irrelevant attribute and to find the top critical nuggets. The experimental results have validated to reduce the attribute and significantly improve the accuracy of the classification task.

Keywords: Data mining, heart disease diagnosis system, navie bayes classification

INTRODUCTION

Heart disease (Health India, http://health.india.com/topics/heart-disease/) refers to a group of diseases or problems in which the heart or the vessels supplying blood to the heart are damaged and are not able to function in a normal way. It was believed that heart diseases occur in older people. But nowadays, heart diseases are quite common in young adults, mainly because of lifestyle and poor eating habits. Heart diseases take years to progress and may begin to develop at a very young age. However, most people do not show any symptoms of heart diseases before they reach their 50s or 60s. There are several factors that increase the risk of heart diseases and associated conditions. These include age, gender, obesity, high blood pressure, high cholesterol levels and stress.

The study (World Heart Federation, 2012) took place over a five-year period which involved people from 11 cities across various regions of India were conducted under the chairmanship of Professors Prakash and Rajeev Gupta. They said "India has the dubious distinction of being known as the coronary and diabetes capital of the world". These results showed the government to develop public health strategies that will change lifestyles if the risk factors are to be controlled. It is dedicated to leading the global fight against cardiovascular disease including heart disease and stroke with a focus on low and middle income countries. They work with ministries of health, members, health practitioners, partners and the World Health Organization to establish best-practice models for cost effective prevention and control.

Feature subset selection is the process of selecting a subset of relevant features for use in model construction Clinical diagnosis (World Heart Federation, 2012) is through doctors rather than patterns hidden in medical data base. Hence there is a chance of wrong diagnosis and treatment. Patients are advised to take more number of tests for diagnosis of a disease. In most of the case, not all the tests contribute towards effective diagnosis of a disease. Medical data bases are high volume in nature. Classification (Jabbar *et al*., 2013) may yield a smaller amount of accurate results if medical data consists of irrelevant and redundant features. The central assumption while using a feature selection technique is that the data contains many redundant or irrelevant features. It is been active and rich field of research in machine learning and data mining. Feature selection is a dimensionality reduction technique used to reduce irrelevant data and to increase accuracy (Kotsiantis *et al*., 2006; Jabbar *et al*., 2012).

The unavailability of experts and incorrectly identified cases has demanded the need to develop a debauched and efficient diagnosis system. Many researchers have been felt on this area and researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart

Corresponding Author: R. Kavitha, Department of CSE, Vel Tech Rangarajan Sagunthala R and D Institute of Science and Technology (Vel Tech RR and SR Technical University), Chennai-62, Tamil Nadu, India

disease. In heart disease database there exists several features out of which only few are relevant or critical features. The feature which contributes towards effective diagnosis is termed as critical feature.

The objective of our study is to predict the diagnosis of heart disease with reduced number of attributes and also to find the critical nuggets and improve the prediction accuracy. By finding the critical nuggets we can dramatically improve the efficiency and accuracy level of the heart disease diagnosis process. Then we find feature subset selection by applying the Pearson co-efficient correlation measure along with standard deviation and conditional entropy to find the relevance among the attributes in the dataset.

After finding the relevancy level and co-relation co-efficient value we find the feature subset. Then we analyse the accuracy level between the predicted features by using Weka tool. In Weka, by using Navie Bayes classification algorithm, we predict the relevant and irrelevant classified instances among the feature selection. After finding the classification accuracy, we find the critical nuggets. To calculate the critical score, we find the average weighted value of each attribute in the dataset by using class variable. Based on the weighted value we evaluate the critical score, based on which we get the critical nuggets (Mansur and Md. Sap, 2005).

LITERATURE REVIEW

In David and Evangelos (2013) a new metric was introduced called CR score, which is used for measuring the criticality of a subset or nugget. A critical nugget is a small collection of records or instances that contain domain-specific important information. The work identified certain properties of nuggets and gave an experimental analysis of the properties. The main goal of the study is to derive an accurate data model that classifies the new test data instances. It also helped to validate that critical nugget help in increasing the classification accuracies in the heart disease data set.

In Jabbar *et al*. (2013) the author applied feature subset selection on medical data to determine that attributed which contributes more towards the disease which indirectly reduce the number of clinical tests to be taken by a patient. He applied K nearest neighbour with feature subset selection in the diagnosis of heart disease. The results showed that the accuracy is increased in the diagnosis of heart disease.

In Mai *et al*. (2012) the author investigate diagnosis of heart disease by applying KNN technique. The results showed that by applying the KNN we could achieve higher accuracy than neural ensemble in the diagnosis of heart disease patients. The result also showed that applying voting could not enhance the KNN accuracy in the heart disease diagnosis.

Table 1: Heart disease attributes

Sl. No	Attribute name	Properties
1	Age	Year
2	Sex	1-Male 0-Female
3	Chest pain type	1-Angina type 1 2-Angina type 3-Non-angina 4-Asymptomatic
4	Trestbps	Mmhg on admission to the hospital
5	Chol	mg/dL
6	Fasting blood sugar	1: >120 mg/dL 0: <120 mg/dL
7	Restecg	0: Normal 1: 1 having ST-T 2: Showing probable
8	Thalach	Maximum heart rate achieved
9	Exang	1: Yes 0: No
10	Oldpeak	ST depression induced
11	Slope	1: Unsloping 2: Flat 3: Downsloping
12	Ca Tha	Value 0-3) 3: Normal 6: Fixed defect 7: Reversible defect
13	Num	Predicted attribute
14	Abpre	Distinct class attribute

In Jabbar *et al*. (2012) and Koteeswaran *et al*. (2012) associative rule mining is used to model a prediction system. The author proposes an efficient associative classification algorithm using genetic approach for heart disease prediction. The experimental results showed that most of the classifier rules help in the best prediction of heart disease which even help doctor's in the diagnosis decisions.

In Syed *et al*. (2013), the author compares the performance and working of clinical decision support system which uses different data mining techniques for heart disease prediction and diagnosis.

In Chaitrali and Sulabha (2012), a heart disease prediction system is developed using neural network. The system predicts the chance of patient getting a heart disease. For the prediction the system used 13 attributes and for the better results they added two more attributes with it.

In Anbarasi *et al*. (2010), genetic algorithm is used to determine the attributes which contributes more towards the diagnosis of heart ailment which indirectly reduces the number of test which are needed to be taken by a patient.in which 13 attributes are reduced to 6 attributes. Also the author showed that decision tree data mining technique and Navies Bayes performed consistently well.

Data set: A dataset of 597 records are taken with 15 attributes. They are listed in Table 1. In the attribute selection the distict class variable is found as Abpre with values 1.0 and 2.0.

Proposed work: Using the real time heart disease database, the outliers are detected as the initial step from Fig. 1. It is the primary step in the data mining

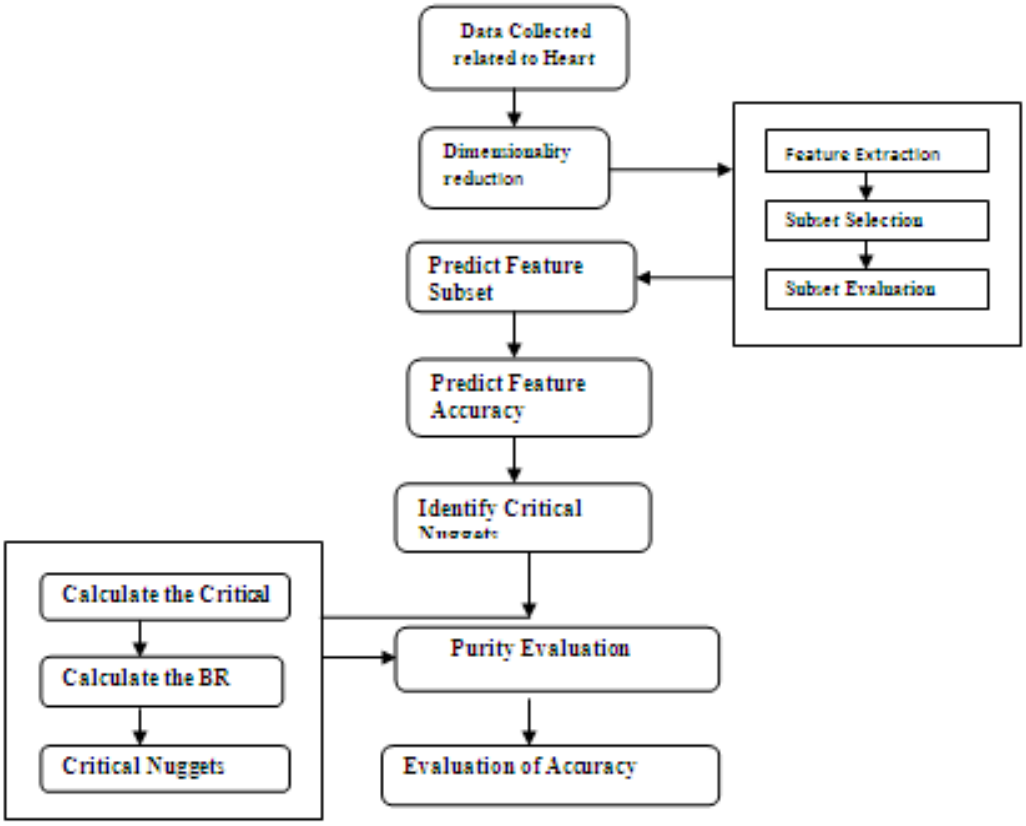

Fig. 1: Heart disease detection system

application. An outlier is an observation point that is distant from other observation. During the knowledge discovery the presence of irrelevant and redundant information creates difficulty in training phase. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, featur e extraction and selection etc. The product of data pre-processing is the final training set. This method is used for the noise removal from the real time dataset.

The next step is feature selection. Initially there are 15 attributes considered in common for the heart disease database. From the attributes the distinct class variable is taken for the feature extraction. On finding the class variable standard deviation is applied on both the classes. The standard deviation (SD) (represented by the Greek letter sigma, σ) shows how much variation or dispersion from the average exists. A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value); a high standard deviation indicates that the data points are spread out over a large range of values.

For Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

where,

$(x_i-\mu)^2$

μ = Mean value

where,

x_i = Individual values in dataset

N = Total number of values

$\sum_{i=1}^{N}(x_i - \mu)^2$ = Sum of all values

It is applied to find the most important attribute. The entropy and the conditional entropy is calculated for both of the distinct class variables.

For Entropy:

$$E = \sum_{i-1}^{k} p_i \, log_2 p_i$$

where, p_i is the probability density value.

After doing the pre-processing the Pearson correlation coefficient is used to find the subset selection from the dataset.

The Pearson correlation coefficient measures the strength and direction of the relationship between two variables. Where the value r = 1 means a perfect positive correlation and the value r = -1 means a perfect negative correlation.

Formula:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

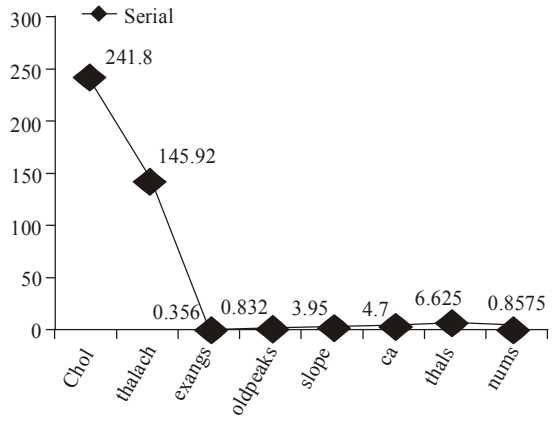

Fig. 2: Line graph which shows the critical score of each attribute

Table 2: Attributes with their critical score

Attribute	Critical score
Chol	241.8
Thalach	145.92
Exangs	0.356
Oldpeaks	0.832
Slope	3.95
Ca	4.7
Thals	6.625
Nums	0.8575

Table 3: list of reduced attributes

Critical nuggets based on critical scores
Chol
Thalach
Slope
Ca
Thal

Table 4: Before extracting the relevant attributes

Correctly classified	259	43.0%
Incorrectly classified	338	56.0%

Table 5: After extracting the relevant attributes

Correctly classified	333	55.0%
Incorrectly classified	264	44.0%

Then the relevant data and the irrelevant data from the heart disease data is identified. Feature selection, also known as subset selection, is the process of selecting a subset of relevant features. In the relevancy process the true positive and true negative values are calculated for before extracting the relevant attributes and after extracting the relevant attributes. The sensitivity and the specificity values of the dataset are calculated as 55 and 44%, respectively:

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{Positive}}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{Negative}}$$

Feature selection is also useful as part of the data analysis process. Subset selection evaluates a subset of features as a group for suitability.

The critical nugget is found by using the critical nugget score. Using the nugget score more relevant attribute is calculated. Finally top k nugget score is calculated and the CN (critical nugget) score is calculated to find the critical nugget. After calculating the critical score value we sort it in descending order, based on which we can get the top K nuggets value:

$$\text{CR}_{\text{score}} = \frac{\sum_{j=1}^{n}(w_j)}{n}$$

where,

$$\text{w}_\text{j} = \frac{\left(w_j^+ + w_j^-\right)}{2}, w_j^+ = \frac{d_j^+}{d}$$

and,

$$w_j^- = \frac{d_j^-}{d}$$

The critical score value produced after this process is tabulated in Table 2.

A line graph in Fig. 2 shows the data labels of each attribute and also depicts the critical attribute which have the high score.

Finally we produce only 5 relevant attribute which are considered as most critical attribute in Table 3.

EXPERIMENTAL RESULTS

The experimental results were conducted with weka tool. A 597 record heart disease data set is taken for the experimental purpose. The data set consist of 15 attributes. The sensitivity and specificity of the dataset are calculated for both Table 4 and 5 with relevant attribute and irrelevant attribute.

$$\text{Precision} = \frac{t-pos}{(t-pos+f-pos)}$$

$$\text{Accuracy} = \text{sensitivity}.\frac{pos}{(pos-neg)} + specificity\ \frac{neg}{(pos-neg)}$$

The accuracy and precision are calculated as 99 and 55%, respectively. In Anbarasi *et al*. (2010) the accuracy of the Naive Bayes is 96.5% whereas for our system it is 99%.

CONCLUSION

This heart disease detection system is developed in order to reduce the total number of test prescribes by the doctor to the patients. The heart disease detection system is developed with the reduced attributes. It is been proposed that we can diagnose the heart disease with the 5 attribute only from the 15 attributes. The critical attribute are produced by calculating the critical score for each attribute in the data set. This approach

also reduces the total cost spent by the patient in the diagnosis of the heart disease. The precision and the accuracy of our detection system are calculated. Also it is shown that our system has better accuracy compared with other models.

REFERENCES

Anbarasi, M., E. Anupriya and N.C.H.S.N. Iyengar, 2010. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int. J. Eng. Sci. Technol., 2(10): 5370-5376.

Chaitrali, S.D. and S.A. Sulabha, 2012. A data mining approach for prediction of heart disease using neural networks. Int. J. Comput. Eng. Technol., 3(3): 30-40.

David, S. and T. Evangelos, 2013. On identifying critical nuggets of information during classification tasks. IEEE T. Knowl. Data En., 25(6): 1354-1367.

Jabbar, M.A., D.L. Deekshatulu and P. Chandra, 2012. Heart disease prediction system using associative classification and genetic algorithm. Proceeding of the International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies (ICECIT, 2012), 1: 183-192.

Jabbar, M.A., D.L. Deekshatulu and P. Chandra, 2013. Heart disease classification using nearest neighbor classifier with feature subset selection. Annals, Computer Science Series, 11th Tome 1st, Fasc, pp: 47-54.

Koteeswaran, S., J. Janet, E. Kannan and P. Visu, 2012. Terrorist: Intrusion monitoring system using outlier analysis based search knight algorithm. Eur. J. Sci. Res., 74(3): 440-449.

Kotsiantis, S., D. Kanellopoulos and P. Pintelas, 2006. Data preprocessing for supervised leaning. Int. J. Comput. Sci., 1(2): 111-117.

Mai, S., T. Tim and S. Rob, 2012. Applying k-earest neighbour in diagnosing heart disease patients. Int. J. Inform. Educ. Technol., 2(3): 220-223.

Mansur, M.O. and M.N. Md. Sap, 2005. Outlier detection technique in data mining: A research perspective. Proceeding of the Postgraduate Annual Research Seminar, pp: 23-30.

Syed, U.A., A. Kavita and B. Rizwan, 2013. Data mining in clinical decision support systems for diagnosis, prediction and treatment of heart disease. Int. J. Adv. Res. Comput. Eng. Technol., 2(1): 219-223.

World Heart Federation, 2012. Dubai. Retrieved from: http://www.world-heartfederation.org/press/releases/detail/article/reasons-for-indias-growing-cardiovascular-disease-epidemic-pinpointed-in-largest-ever-risk-factor/. (Accessed on: April 20, 2012)