Research Journal of Applied Sciences, Engineering and Technology 11(1): 10-13, 2015
DOI: 10.19026/rjaset.11.1669
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 25, 2014 Accepted: February 27, 2015 Published: September 05, 2015

Research Article

A Framework for Heart Disease Prediction Using K nearest Neighbor Algorithm

R. Kavitha and E. Kannan
Department of CSE, VEL Tech University, Chennai-62, Tamil Nadu, India

Abstract: Heart disease prediction is an area where many researchers are working using different data mining techniques. This study proposes a framework to develop a heart disease prediction process using k-nearest neighbor with wrapper filter. Heart disease diagnosis is mostly done with doctor's knowledge and practice. But the cost spent by the patients are more in order to take test in which all the test does not contribute towards effective diagnosis of disease. The patient's record is predicted to find if they have symptoms of heart disease through data mining techniques. Many researches have been undergone and researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. In heart disease database there exists several features out of which only few are critical features. The feature which contributes towards effective diagnosis is termed as critical feature. Our study proposes a framework to find the subset of critical feature using K nearest neighbor and wrapper filter. This in turn produces a prediction model. Finally we exhibit the ideas of diagnosing heart disease with critical feature.

Keywords: Classification, feature, heart disease, nearest neighbor, prediction

INTRODUCTION

In recent days, (Jabbar *et al*., 2012) the emphasis of the World Heart Day this year is on women and children, as heart diseases have emerged as the number one killer for Indian women, according to doctors. In current times, it was said that heart diseases are in general narrowed to males, whereas the study from various report and survey says that it is more attainment to females. Indian women account for 15% of the global burden of heart disease which kills about 15 million people every year. Heart disease (Times of India, 2009) is actually the number one killer of women where if suffer the first heart attack she is under greater risk of losing her lives as compared to men. In 'The Hindu' it is noted that deaths due to cardio vascular diseases now stood at 24% which might go up to 30 to 40% in the coming years. In the Times of India (2009), it was apprehended that by next ten years, India might have the maximum prevalence of diabetes and heart diseases in the world. The mortality rate due to cardio vascular diseases in Tamil Nadu was 360-430 per 1 Lakh population, the highest in the country. So developing an expert system or prediction of the heart disease has emerged as an important area of work in the field of data mining. Hence the necessity of expert systems and to predict risk factors for these diseases in order to take preventive measures is primitive. Much research work has been carried out using different data mining techniques in this area. Medical data comprises of a number of test essential to diagnose a particular disease (Jabbar *et al*., 2013a).

Medical diagnosis is a cognitive process. An expert uses several medical data to make a diagnostic process. The initial step is to describe a set of disease instead of a specific disease. The next step is, the expert gets the follow up test to more support. The initial diagnostic impression can be a broad term describing a category of diseases instead of a specific disease or condition. After the initial step, the expert obtains follow up tests and procedures to get more data to support or reject the original diagnosis and will attempt to narrow it down to a more specific level. A diagnostic procedure may be performed by various health care professionals such as doctor, nurse, dentist, pediatric etc., (Koteeswaran *et al*., 2012).

Clinical diagnosing (Jabbar *et al*., 2013b) is completed by consultants instead of patterns hidden in medical information base. Thus there's an opportunity of wrong diagnosing and treatment. Patients are suggested to require variety of tests for diagnosing of a sickness. In most of the case, not all the tests contribute towards effective diagnosing of a sickness. Medical information bases are high volume in nature. Classification might manufacture less correct results if medical information consists of irrelevant and redundant options. Feature subset selection is the process of selecting a subset of relevant features for use in system model. The inner assumption when using this technique is that the data contains many redundant or irrelevant features. It has been active and rich field of

Corresponding Author: R. Kavitha, Department of CSE, VEL Tech University, Chennai-62, Tamil Nadu, India

research in machine learning and data mining. Feature selection is a dimensionality reduction technique used to reduce irrelevant data and to increase accuracy. K-nearest neighbor is one of the most widely used classifier. Classification is achieved by identifying the nearest neighbor to determine the class of a sample.

In this study we propose a framework to find the subset of critical feature using K nearest neighbor and wrapper filter. This in turn produces a classification model. This classification model is implemented using different classifiers using voting technique in order to verify the accuracy of the classifier. Finally we exhibit the ideas of diagnosing heart disease with critical feature alone. This study is been carried out with the 13 attributes of the heart disease attributes.

LITERATURE REVIEW

In David and Evangelos (2013) proposed a system for identifying a critical bunch of information using classification tasks. It introduces the ideas of innovative domain-independent method to measure criticality, suggests a heuristics to reduce the search space for finding critical information. In Jabbar *et al.* (2013a) proposed a classification using KNN with feature subset selection in the diagnosing of the art disease. The experimental results show that applying feature subset selection to KNN will enhance the accuracy in the diagnosis of heart disease for AP population. In Mai *et al.* (2012), the researchers have shown that combining different classifiers through voting is outperforming other single classifier. It also investigates that integrating voting technique enhance the accuracy of the classification model that is being built. In Sidahmed *et al.* (2013) considers a final set of attributes that contains the most relevant feature model that increases the accuracy. The algorithm in this case produces 85.5% classification accuracy in the diagnosis of CAD.

In Jyoti *et al.* (2011) and Rajeswari *et al.* (2011) paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Number of test has been taken to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like Neural Networks, KNN, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. In Sidahmed *et al.* (2013), Data mining techniques have been used in medical diagnosis for many years and have been known to be effective. In this study we proposed a new method to discover association rules in medical data to predict heart disease. This approach is expected to help doctors to make accurate decisions (Kavitha and Kannan, 2014).

METHODOLOGY

Stage 1: Finding critical feature set: Consider (David and Evangelos, 2013) a heart disease data set t, combining of m instances, n attributes and two classes namely heart disease data set and non-heart disease data set, denoted a positive '+' and negative '-' otherwise called as true positive and true negative. Also consider a subset t_s, consisting of d instances, where,

T_r^+ : Subset of '+' classes
T_r^- : Subset of '-' classes
- Set of attributes t_r $\{A_1, A_2, \ldots\ldots A_n\}$

D : Data instances $(D_1, D_2, \ldots\ldots D_m)$
B^+ : Positive points that are true
B^- : Negative points that are true
$B^{+'}$: Positive points that are false
$B^{-'}$: Negative points that are false:

$$|t_r^+| = |B^+| + |B^{+'}|$$

$$|t_r^-| = |B^-| + |B^{-'}|$$

$$|t_r| = |t_r^+| + |t_r^-|$$

where, |x| is cardinality of set.

Using the following notations on the medical data set prescribed by Cleveland database (Nidhi and Kiran, 2012) a classification model is defined called M_1, using a classification algorithm (K-Nearest Neighbor) on the training data T_r, then the sample instances from Ts are removed by creating a new classification model M_2. The M_1-M_2 is called as Critical Score (CS-this metric is used to validate the critical attributes of the features in the heart disease data set. Also the assumption is that all the values of attributes are numeric and not categorical. The same process is applied by increasing the data instances of original data and sample data (David and Evangelos, 2013):

$$CS = \frac{\sum_{i=1}^{n} wj}{n} \quad w_j = \frac{w_j^+ + w_j^-}{2} \quad w_j^+ = \frac{d_j^+}{d} \quad w_j^- = \frac{d_j^-}{d}$$

The important property of critical score is that, to find as many attributes that are sensitive to small changes. The greater the number of attributes that are sensitive to such changes, the higher is the resulting Critical Score (CS). The attribute with the high critical score is considered as critical feature. The feature is identified as critical and also the priority of the feature is selected by the ranking method. CS (t) >CS (t_s) only when the t is more critical than t_s.

Critical score algorithm for feature selection:

M1 = Model resulting
m = Data instances in T_r
n = Number of attributes
P_0 = Query M_o to obtain new class labels for N1
Δ_j = A measure of how much attributes values can increase or decrease
Δ_j = max (t_s [A_j] - min (t_s [A_j]) to find the max and min values of vectors for training sample t_s [A_j]
If Δj = 0 then no change
Else
V = t_s [A_j] + Δ_j - increment all values by Δ_j
Genetare new matrix
Obtain new class labels for N_1
V = t_s [A_j] - Δ_j -decrement all values in column by Δj
Generate new matrix
Obtain new class labels for N_2
w_j^+ = positive samples,
w_j- = negative samples these are the number of switched classes:

Calculate $w_j = \frac{w_j^+ + w_j^-}{2}$

Calculate CS = $\frac{\sum (wj)}{n}$ Return CS

List of attributes considered:

- Sex (value 1: Male; value 0: Female)
- Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
- Fasting Blood Sugar (value 1: >120 mg/dL; value 0: <120 mg/dL)
- Restecg-resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
- Exang-exercise induced angina (value 1: yes; value 0: no)
- Slope-the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
- CA-number of major vessels colored by floursopy (value 0-3)
- Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
- Trest Blood Pressure (mm Hg on admission to the hospital)
- Serum Cholesterol (mg/dL)
- Thalach-maximum heart rate achieved
- Oldpeak-ST depression induced by exercise relative to rest
- Age in Year

KNN is a method for classifying objects based on closest training data in the feature space. It is a type of instance based learning (Jyoti *et al*., 2011).

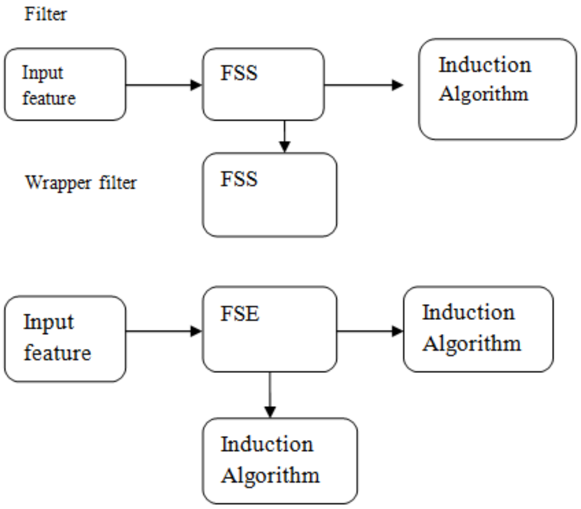

Fig. 1: Feature evaluation

Algorithm KNN (Jabbar *et al*., 2013b):

- Let k be the number of nearest neighbor and T be the set of training example
- For each test sample Z = {x', y'} do
- Do compute distance
- Select k closest training example
- y'-argmax £ (xi, yi) $\sum T_z (v = y_i)$
- End

Stage 2: Feature subset selection: Preprocessing (Jabbar *et al*., 2013a) used in machine learning, where subset of features available for the data is selected for learning algorithm. The subset containing the critical attributes that contributes to accuracy for the prediction system is shown in the Fig. 1.

Subset evaluation: Primary risk Factors for heart disease are:

- Use of tobacco
- Alcohol
- Hyper tension
- Physical inactivity
- Cholesterol
- Obesity unhealthy diet
- Raised blood glucose

Algorithm for selecting feature subset:

- Choose the heart disease data set
- Apply CS measure on the dataset
- Identify the critical feature and remove the rest
- Apply K nearest neighbor algorithm
- Find the accuracy and correct prediction of the classifier

RESULTS AND CONCLUSION

This method indirectly reduces the number of tests to be taken by the patients. This model given the feature extraction and the ranking of the features using wrapper method. The Information gain of each attribute is taken and the raking of the features are found and predicted with the ranking. The attribute with the highest rank is taken and considered as the relevant or critical feature. So by the use of this method the number of test taken for the diagnosis of the patient will be reduced greatly. This prediction model helps the experts in efficient decision making process with fewer attributes to diagnose the heart disease. In this study, we have presented an intelligent and effective heart disease prediction system using data mining techniques:

```
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 14 num):
Information Gain Ranking Filter
Ranked attributes:
0.25393          3 chest_pain
0.24723          9 exang
0.22447          10 oldpeak
0.073            8 thalach
0.05913          2 sex
0.0297           11 slope
0.01737          6 fbs
0.00269          7 restecg
0.00207          13 thal
0                4 trestbps
0                5 chol
0                12 ca
0                1 age
Selected attributes: 3, 9, 10, 8, 2, 11, 6, 7, 13, 4, 5, 12, 1: 13
```

Firstly, we have provided an efficient approach for the extraction of features from the heart disease data repository. The preprocessed heart disease data repository was extracted using KNN algorithm.

REFERENCES

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, J.K., C. Priti and B.L. Deekshatulu, 2012. Knowledge discover from mining association rules for heart disease prediction. J. Theor. Appl. Inf. Technol., 41(2).

Jabbar, M.A., B.L. Deekshatulu and P. Chandra, 2013a. Heart disease classification using nearest neighbor classifier with feature subset selection. Anale Seria Inform., 9: 47-54.

Jabbar, M.A., B.L. Deekshatulu and P. Chandra, 2013b. Classification of heart disease using artificial neural network and feature subset selection. Glob. J. Comput. Sci. Tech. Neural Artif. Intell., Version 1.0., 13(3).

Jyoti, S., A. Ujma, S. Dipesh and S. Sunita, 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. Int. J. Comput. Appl., 17(8): 975-8887.

Kavitha, R. and E. Kannan, 2014. A methodology for heart disease diagnosis using data mining technique. Res. J. Appl. Sci. Eng. Technol., 8(11): 1350-1354.

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.

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

Nidhi, B. and J. Kiran, 2012. An analysis of heart disease prediction using different data mining techniques. Int. J. Eng. Res. Technol., 1(8): 1-4.

Rajeswari, K., V. Vaithiyanathan and P. Amirtharaj, 2011. Prediction of risk score for heart disease in India using machine intelligence. Proceeding of the International Conference on Information And Network Technology IACSIT Press, Singapore, 4: 18-22.

Sidahmed, M., A. Baghdad and M. Mostéfa, 2013. Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm. In: Sundarapandian *et al*. (Eds.), CSE, CICS, DBDM, AIFL, SCOM, 1: 41-51.