Journal of Scientific & Industrial Research
Vol. 76, November 2017, pp. 720-724

Optimal and Novel Hybrid Feature Selector for Accurate Prediction of Heart Disease

B Amarnath[1]* and S A A Balamurugan[2]

*[1]Department of Computer Science & Engineering, Veerammal Engineering College, K.Singarakottai, Tamil Nadu, India
[2]Department of Information Technology, K L N College of Information Technology, Pottapalayam, Tamil Nadu, India

Received 04 September 2016; revised 22 April 2017; accepted 05 September 2017

Heart disease prediction is designed to support clinicians in their diagnosis. We proposed a method for classifying the heart disease data. The patient's record is predicted to find if they have symptoms of heart disease through data mining. It is essential to find the best fit classification algorithm that has greater accuracy on classification in the case of heart disease prediction. Since the data is huge attribute selection method used for reducing the dataset. Then the reduced data is given to the classification. In the investigation, the hybrid attribute selection method combining CFS and Filter Subset Evaluation gives better accuracy for classification. We also propose a new feature selection method algorithm which is the hybrid method combining CFS and Bayes Theorem. The proposed algorithm provides better accuracy compared to the traditional algorithm and the hybrid algorithm CFS and Filter Subset Evaluation.

Keywords: Data Mining, Feature Selection, Classification

Introduction

Data mining is defined as the process of extracting data, analyzing it then discovering knowledge in a useful form that identifies correlation within the data. Data mining is classified into descriptive and predictive methods. The heart disease data prediction is designed to support clinicians in their diagnosis for heart disease prediction[1,2]. They typically work through an analysis of medical data and a knowledge base of clinical expertise. The quality of medical diagnostic decisions for heart disease can be increased by improvements to these predicting systems. Data mining provides a way to get the information buried in the data. Data mining methods may aid the clinicians in the predication of the survival of patients and in the adaptation of the practices consequently. To build models for prediction of the class based on selected attributes the J48, Bayes Net, Naive Bayes, Simple Cart and REPTREE algorithms are to classify and develop a model to diagnose heart attacks in the patient data set from medical practitioners[3]. An embedded hybrid intelligent classification solution approach based on dynamically reduced subsets of features is validated by using multi-classifier RF classification technique to identify the CHD attack cases[4]. Statistical and classification techniques were utilized to develop the multi-parametric feature of HRV (Heart Rate Variability). Besides, they have assessed the linear and the non-linear properties of HRV for three recumbent positions, to be precise the supine, left lateral and right lateral position. Numerous experiments were conducted by them on linear and nonlinear characteristics of HRV indices to assess several classifiers. E.g.: Bayesian classifiers, Classification based on Multiple Association Rules, Decision Tree, Support Vector Machine and Multilayer Perceptron[5-9]. Intelligent Heart Disease Prediction System built with the aid of data mining techniques like a Decision Tree, Naive Bayes, and Neural Network[10]. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. With the aid of neural networks, the experiments were carried out on a sample database of patient's records of blood pressure and sugar[11]. The Coactive Neuro-Fuzzy Inference System (CANFIS) model diagnosed the presence of disease by merging the neural network adaptive capabilities and the fuzzy logic qualitative approach and further integrating with genetic algorithm. The CANFIS model is assured in the prediction of heart disease[12]. A decision rule-based method incorporates domain-specific definitions of high, medium and low correlations techniques and the proposed algorithm

*Author for Correspondence
E-mail: amars_88@yahoo.co.in

conducts a heuristic search for the most relevant features for the prediction task[13].

The following classification algorithms such as Naive Bayes, Decision Tree, Multilayer Perceptron, KNN are applied on heart disease dataset to predict whether the patient is affected by heart disease or not. Here, we propose the hybrid algorithm combining the best feature selection methods for classification.

Since, we have large collections of data which consumes more time for classification. The patient's record is classified and predicted if they have the symptoms of heart disease. It is essential to find the best-fit algorithm that has greater accuracy on classification in the case of heart disease classification. A large number of data can be reduced that using hybrid attributes selection methods. In order to find the best two algorithms of attribute selection method, the attribute selection method which gives higher accuracy after removing attributes for both classification and clustering are identified and combined to form the hybrid attribute selection method. Then the reduced data are fed into a sequence of classifiers classified to attain better accuracy.

Experimental method

The Correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated with each other. Chi Square is a feature selection method which is used to test whether the occurrence of a specific term and the occurrence of a specific class are independent. Relief is a feature selection algorithm used in binary classification (generalisable to polynomial classification by decomposition into a number of binary problems). Its strengths are that it is not dependent on heuristics, runs in low-order polynomial time, and is noise-tolerant and robust to feature interactions, as well as being applicable for binary or continuous data; however, it does not discriminate between redundant features, and low numbers of training instances fool the algorithm.

CFS+Bayes Theorem are a feature selection method which uses the concept of correlation analysis and conditional independency to select the best subset of original features. Information gain is a ranking based feature selection method which measures the amount of information in bits about the class prediction to decide most relevant features.

CFS and Bayes theorem based feature selector

A new hybrid feature selection method by combining CFS and Bayes Theorem is proposed. The CFS algorithm reduces the number of attributes based on the SU measure, In CFS each attribute are compared pairwise to find the similarity and the attributes are compared to class attribute to find the amount of contribution it provides to the class value, based on these the attributes are removed. The selected attributes from the CFS algorithm is fed into Bayes theorem for further reduction. Bayes theorem calculates the conditional probability for each attribute and the attribute which has highest conditional probability is selected. Both the algorithms CFS and Bayes theorem works on the conditional probability measure.

Proposed algorithm

input: S(F1; F2; :::; FN;C) // a training data set
// a predefined threshold
output: Sbest {Abest(highest IG)} // an optimal subset

1 begin
2 for i = 1 to N do begin
3 calculate SUi;c for Fi;
4 if ($Su_{i,c} \geq \delta$)
5 append F_i to S'_{list};
6 end;
7 order S'_{list} in descending SUi,c value;
8 F_p = getF irstElement(S'_{list});
9 do begin
10 F_q = getNextElement(S'_{list} , F_p);
11 if (F_q <> NULL)
12 do begin
13 $F'_q = F_q$;
14 if ($SU_{p,q}$, $SU_{q,c}$)
15 remove Fq from S'_{list} ;
16 Fq = getNextElement(S'_{list} , F'_q);
17 else Fq = getNextElement(S'_{list} , F_q);
18 end until (Fq == NULL);
19 F_p = getNextElement(S'_{list} , F_p);
20 end until (Fp == NULL);
21 $S_{best} = S'_{list}$;
22 Sbest={X1,X2,..XN}
23 for j=1 to N begin
24 for k=j+1 to N begin
25 $P[C_m/(X_j,X_k)]=P[(X_j,X_k)/C_m]*P(C_m)$
26 $P[C/(X_j,X_k)]= [C_1/(X_j,X_k)]+P[C_2/(X_j,X_k)]+\ldots$ .... $+P[C_n/(X_j,X_k)]$
27 If(P[C/Xj,Xk)]>Ω)
28 {
29 if((P[C/Xj]>(P[C/Xk])
30 Remove Xk from Sbest

```
31  Sbest=IG(X)
32  Else
33  Remove Xj from Sbest
34  Sbest=IG(X)
35  }
36  end;
```

Dataset used in the experiment

The following is the sample of the Heart Disease Data.arff

```
@relation heart-stat log
@attribute age real
@attribute sex real
@attribute chest real
@attribute resting_blood_pressure real
@attribute serum_cholestoral real
@attribute fasting_blood_sugar real
@attribute resting_electrocardiographic_results real
@attribute maximum_heart_rate_achieved real
@attribute exercise_induced_angina real
@attribute old peak real
@attribute slope real
@attribute number_of_major_vessels real
@attribute thal real
@attribute class {absent, present}
@data
70,1,4,130,322,0,2,109,0,2.4,2,3,3, present
67,0,3,115,564,0,2,160,0,1.6,2,0,7, absent
57,1,2,124,261,0,0,141,0,0.3,1,0,7, present
64,1,4,128,263,0,0,105,1,0.2,2,1,7, absent
74,0,2,120,269,0,2,121,1,0.2,1,1,3, absent
65,1,4,120,177,0,0,140,0,0.4,1,0,7, absent
56,1,3,130,256,1,2,142,1,0.6,2,1,6, present
59,1,4,110,239,0,2,142,1,1.2,2,1,7, present
```

The Heart Disease data after applying traditional method in WEKA, The original attributes is reduced significantly then these attributes can be fed to various classifiers. The CFS+Bayes theorem algorithm is implemented where the attribute after CFS is 4 and the selected attributes after Bayes theorem is only 3. CFS feature selection method which selects the attributes based on the symmetrical uncertainty reduces the number of attributes from 13 to 4. The reduced attributes is fed to Bayes theorem for further reduction. We applied four classification algorithms for heart disease data such as NB, J48, KNN, and NN. First, we applied these algorithms for the whole data set. The whole data in ARFF document is given to WEKA and classification algorithm is applied to it. The heart.arff will contain a large quantity of data and apply classification algorithms to this dataset are time-consuming and also give a result with less accuracy. Hence we have to reduce the data set by using attribute selection method. Then this reduced dataset is fed into the four classification algorithm and which algorithm is best fit for this prediction is investigated. Likewise, all other attribute selections and classification algorithms are applied to heart disease dataset. From that, we found that NB classification algorithm gives better accuracy after applying the CFS attribute selection method.

Hybrid feature selector

The two best feature selection methods are applied in sequence. (i.e.) CFS followed by Filtered Subset Evaluation. In this method, the reduced number of attributes after CFS is 7 and these 7 attributes are fed to Filtered Subset Evaluation which reduces as 6 attributes. After applying the hybrid feature selector, the data is applied to the classification algorithm in which Naive Bayes gives higher accuracy compared to the other classifiers. After applying into Naive Bayes, the incorrectly classified instances are separated. The correctly classified samples are kept as training set and the incorrectly classified samples as test set are fed into various other classifiers, where the J48 gives greater accuracy. We proposed a new hybrid algorithm that is CFS+Bayes Theorem. When applying this feature selection algorithm, the attributes are reduced as 3. Then reduced dataset is given to classifiers. Naive Bayes gives the greater accuracy compared to other classifiers.

Result and Discussion

The dataset used in the experimentation is collected from UCI machine learning repository[14]. This dataset contains 270 Instances and 14 Attributes. All the Attributes are numeric-valued. The hybrid feature selector method is implemented using CFS and Bayes theorem. The accuracy of classification algorithms on original features set is shown in Table 1. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data. The measure of a model's ability is to correctly label a previously unseen test case. If the label is categorical (classification), accuracy is commonly reported as the rate which a case will be labeled with the right category. If the label is continuous, accuracy is commonly reported as the average distance between the predicted label and the correct value. A confusion matrix displays the number

of correct and incorrect Predictions made by the model compared with the actual classifications in the test data. The matrix is *n*-by-*n*, where *n* is the number of classes. From that, we calculated the accuracy of each classification algorithms.

Attribute selection

The dataset contains a large number of features. The irreverent and redundant features are removed using the attribute selection method. Number of selected attributes by each feature selection method is shown in Table 2.

Performance evaluation for classifiers

After reducing the number of irrelevant and redundant features, the resulting features is given as input to various classifiers and the performance evaluation of each classification algorithm is shown in Table 2. A new hybrid Feature selector by combining CFS and Bayes theorem is proposed in this work. We found that CFS and Bayes theorem based feature will able to improve the detection performance of all the classification algorithm in the domain of heart disease prediction with minimal feature sets and less computational time than other methods. In order to evaluate the prediction rate, there are several indices such as specificity, sensitivity, precision, and accuracy to assess to assess the models' validity. These indices are calculated by the confusion matrix as shown in Figure 1. This matrix is a useful tool for analyzing the performance of classification method in data diagnosis or observations of various categories. The ideal state, most parts of the relevant data with the observations should be located on the main diagonal of the matrix, and the remaining values of the matrix are zero or near zero.

- FN= The number of positively labeled data, which falsely have been classified as "Negative".
- TN= The number of negatively labeled data, which have been classified as "Correct".
- TP= The number of positively labeled data, which have been classified as "Correct".
- FP= The number of negatively labeled data, which falsely have been classified as "Positive".

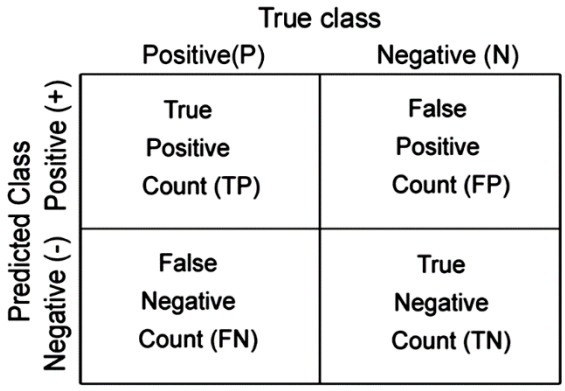

Fig.1 — Confusion Matrix

Table 1 — Classifiers accuracy on full feature set

Classifiers	Correctly Classified Samples	Incorrectly Classified Samples	Accuracy (%)
Naive Bayes	226	44	83.70
J48	207	63	76.66
KNN	203	67	75.18
Multilayer Perceptron	211	59	78.148

Table 2 — Number of selected features and classifiers performance on selected features by each feature selection method

Attribute Selection Methods	No of Attributes Selected	NB	KNN	J48	ANN	Avg.
CFS	7(3,7,8,9,10,12,13)	85.5	78.14	81.11	82.22	81.74
Chi-Squared	13(1,2,3,4,5,6,7,8,9,10,11,12,13)	83.70	75.18	76.66	80.37	78.97
Consistency Subset	10(1,2,3,7,8,9,10,11,12,13)	84.07	78.14	78.88	81.11	80.55
Filtered Attribute Eval.	13(1,2,3,4,5,6,7,8,9,10,11,12,13)	83.70	75.18	76.66	80.37	78.97
Filtered Subset Eval.	6(3,8,9,10,12,13)	85.18	80	79.60	78.88	80.91
Gain Ratio	13(1,2,3,4,5,6,7,8,9,10,11,12,13)	83.70	75.18	76.66	78.88	78.60
Information Gain	13(1,2,3,4,5,6,7,8,9,10,11,12,13)	83.70	75.18	76.66	80.37	78.97
Latent Semantic Analysis	1(1)	54.07	51.11	55.55	52.96	53.42
One R Attribute Eval.	13(1,2,3,4,5,6,7,8,9,10,11,12,13)	83.70	75.18	76.66	79.25	78.69
Relief Attribute Eval.	13(1,2,3,4,5,6,7,8,9,10,11,12,13)	83.70	75.18	76.66	78.14	78.42
CFS+Filtered Subset	6(3, 8, 9, 12, 13)	85.18	80.74	79.62	78.88	81.10
CFS+Bayes Theorem	3(3, 12, 13)	80.37	85.55	85.18	85.18	84.07

Table3 — Comparison on model fitness and model accuracy of four various applied machine learning algorithms

Algorithms	Specificity	Sensitivity	Sensitivity	Precision	Accuracy (%)
DT	0.89	0.84	0.76	0.85	76.66
NB	0.84	0.79	0.84	0.79	83.70
KNN	0.91	0.79	0.75	0.87	75.18
Multilayer Perceptron	0.91	0.78	0.78	0.87	78.15

This section presents the experimental results and analysis done for this study. In this work, four classifiers including DT, KNN, NB and Multilayer Perceptron are conducted. Data divided into training set and test set (70% and 30% respectively). The training set is used to build the classifier and test set used to validate it. Model development is conducted in two main steps including model fitness and model accuracy. To calculate the model fitness criteria we used the data of training set; however, to compute the model accuracy measurements, data of testing set is applied which is merely much more valuable to judge about our model's accuracy. Related results of these experiments are demonstrated in Table 3. NB classifier has been able to build a model with the greatest accuracy since the model prediction accuracy is 83.70%. Model accuracies obtained from other classifiers are different as this value for DT, KNN and Multilayer Perceptron have been 76.667%, 75.18%, and 78.15% respectively.

Conclusion

This paper investigated the significance of feature selection methods for improving the performance of classification methods. The experimentation is conducted on the dataset of health care domain. It is found that the CFS and Filter Subset Evaluation reduces the number of irrelevant and redundant attributes thereby increases the performance of classifiers. In addition the new feature selection namely CFS and BT was proposed. The proposed algorithm gives better accuracy for NB and KNN classifier. We conclude that CFS and Bayes Theorem based feature selector is best suitable for heart disease data prediction.

References

1 Suganya R, Rajaram S, Sheik Abdullah A & Rajendran J, A novel feature selection method for predicting heart diseases with data mining techniques, *Asian J of Info Tech*, **15** (2016) 1314-1321.

2 Anbarasi M, Anupriya E & Iyengar N CH S N, Enhanced prediction of heart disease with feature subset selection using genetic algorithm, *Int J of Engg Sci Tech*, **2** (2010) 5370-5376.

3 Walid Moudani, Feature selection for heart disease classification, *Int J of Med, Hlth, Biomed, Bioengg and Pharm Engg*, **7** (2013) 105-110.

4 Hlaudi Daniel Masethe & Mosima Anna Masethe, prediction of heart disease using classification algorithms, *Proc of the world cong on engg and comp sci*, **2** (2014).

5 Heon Gyu Lee, Ki Yong Noh & Keun Ho Ryu, Mining Biosignal Data: coronary artery disease diagnosis using linear and nonlinear features of HRV, *Emer Tech in Know Dis and Data Min*, (2007) 218-228.

6 Wenmin Li, Jiawei Han & Jian Pei, CMAR: Accurate and efficient classification based on multiple association rules, *In Proc of the IEEE Int Conf on Dat Min*, (2001) 369-376.

7 R. Agrawal & R. Srikant, Fast algorithms for mining association rules in large databases, *In Proc of the 20th Int Conf on Very Large Databases*, (1994).

8 Cristianini N, Shawe Taylor J, An Introduction to support vector machines: and other kernel-based learning Methods C*am Univ Press New York*, (2000).

9 Sarah Ashoori & Shahriar Mohammadib, Compare failure prediction models based on feature selection technique: empirical case from iran, *Proc Com Sci*, **3** (2011) 568–573.

10 Sellappan Palaniappan & Rafiah Awang, Intelligent heart disease prediction system using data mining techniques, *Int J of Comp Sci and Netw Secur*, **8** (2008) 343-350.

11 Niti Guru, Anil Dahiya & Navin Rajpal, Decision support system for heart disease diagnosis using neural network, *Delhi Busi Rev*, **8** (2007), 1-6.

12 Latha Parthiban & Subramanian R, Intelligent heart disease prediction system using CANFIS and genetic algorithm, *Int J of Biolog, Biomed and Med Sci*, **3** (2008) 157-160.

13 Andreeva P, Dimitrova M & Gegov A, Information representation in cardiological knowledge based system, *SAER'06*, (2006) 23-25.

14 UCI Repository: *http://archive.ics.uci.edu/ml*.