Biomedical Research 2016; Special Issue: S107-S111 *ISSN 0970-938X*
www.biomedres.info

C4.5 classification algorithm with back-track pruning for accurate prediction of heart disease.

Jothikumar R[1], Siva Balan RV[2*]

[1]Department of Computer Science and engineering, Noorul Islam University, Kumaracoil, Thuckalay, Kanyakumari Dt-629180, Tamil Nadu, India

[2]Department of Master of Computer Application, Noorul Islam University, Kumaracoil, Thuckalay, Kanyakumari Dt-629180, Tamil Nadu, India

Abstract

In Machine Learning, Decision tree is the mostly used classifier for predictive Modeling. The C4.5 classifier suffers from overfitting; poor attribute split technique, inability to handle continuous valued and missing valued attributes with high learning cost. Among all, overfitting and split attribute has high impact on the accuracies of prediction. The Efficient Back-track pruning algorithm is introduced here to overcome the drawback of overfitting. The proposed concept is implemented and evaluated with the UCI Machine Learning Hungarian database. This database having 294 records with fourteen attributes were used for forecasting the heart disease and relevant accuracies were measured. This implementation shows that the proposed Back-track pruned algorithm is efficient when compared with existing C4.5 algorithm, which is more suitable for the application of large amounts of healthcare data. Its accuracy has been greatly improved in line with the practical Health care Historical data. The result obtained proves that the performance of Back-track pruned C4.5 algorithm is better than C4.5 algorithm.

Keywords: Data mining, Heart disease, Decision Tree, C4.5 algorithm, Overfitting, Back-track pruning.

Accepted on August 09, 2016

Introduction

Heart disease has become the major challenge for health care industries. This illness is one of the most leading reasons of death all over the world in the past decade [1]. Cardiovascular disease or heart disease is a class of disease that involves the heart, blood vessels (arteries, capillaries and veins). It is hard for health care professionals to forecast the heart attack as it is a difficult task [2]. The American heart association has estimated that 17.3 million people die because of cardiovascular disease every year, particularly heart attacks, strokes, coronary heart disease and pulmonary heart disease etc. This global cause of death can increase the number to grow more than 23.6 million by 2030. The populations affected by heart diseases are mostly in low and middle-income countries, Where 80% of these deaths occur usually at younger ages than in higher income countries.

Health care industry contains huge volume of data and this can be used for effective analysis and diagnosis of many diseases by several data mining algorithms [3]. The medical industry is highly information rich, but knowledge poor [4]. However, there is a lack of effective analysis tools to discover hidden relationships in data [5]. Like for instance the symptoms from clinic, the practical and pathological symptoms of heart diseases are linked with the human organs including heart, which shows signs of different diseases in human body. Perhaps these signs have similar symptoms of heart diseases as well. To prevent the cause of death and reduce in number, analysis and forecast is very important, but it has never been an easy task for accurate diagnosis of heart diseases. Lots of research is being done for diagnosis of heart disease, but still the complications in various factors are causing delay in diagnosis of the heart related diseases and deciding the accuracy. Researchers are facing difficulties to find accuracy in diagnosis. Prediction is the knowledge to predict future data from historical and current data according to time-series data [6].

Data mining has been extensively implemented by research professionals to assist medical professionals to enhance accuracy for the finding of heart disease [7]. Disease diagnosis is one of the applications where data mining tools are proving successful results. C4.5 classification is one such algorithm playing major role in prediction with few drawbacks as mentioned. The Modified Back-track pruned C4.5 classification algorithm is introduced which can take huge volume of data as input from medical domain to identify and foretell the diseases better than C4.5. C4.5 Classification algorithm has been used by many researches in different application areas. Even they provide good results it suffers from the problem of overfitting. The tree works well for

learned instances and not for new real-time instances called overfitting. The proposed Back-track pruned algorithm uses Back-tracking concept to reorganize the nodes of the constructed tree to overcome the drawback. The proposed Back-track pruned algorithm behaves well for the given dataset than C4.5.

Materials and Methods

Literature survey

Recent Techniques by Bouali and Akaichi like Artificial Neural Netowrk, Support Vector Machine, and Bayesian Network, Decision Trees, J48 classification, ID3, Decision Table and C4.5 were studied in depth and the drawbacks were considered to overcome the in C4.5 Back-Track Pruning algorithm [8]. Also the data mining techniques discussed by Krishnaiah, V., Narsimha (2016) has been studied in depth for better understanding [9].

Decision tree construction

There is a growing interest nowadays to process large amounts of data using the well-known decision-tree learning algorithms. Building a decision tree as fast as possible against a large dataset without substantial decrease in accuracy and using as little memory as possible is essential [10]. One of the most challenging tasks for data mining community is to develop classifiers that can mine very large datasets [11]. A decision tree is used to determine the optimum course of action, in situations having several possible alternatives with uncertain outcomes [12]. C4.5 Decision Tree is the Very First Fundamental Supervised Machine Learning classification algorithm which is extensively implemented and typically achieves very good performance in prediction.

The decision tree-based algorithms do not perceive noise and contradictions in data [13]. After getting the class labels of a record set, decision tree technique is applied to find out the method the attributes behaves, to predict the class labels for the latest instances that can be alike. It learns through the underlying structure of training data when bunch of historical records with answers or class labels is fed as Input. The machine learning happens here. If new records from test data were given as input, it can predict the class label. The generated tree contains Root Node, Branch Node and Leaf or class Node. The root node represents test on an Attribute, branch node represents outcome of the test and leaf node represents the decision made after testing all attributes called class labels. The path generated from root node to leaf node represents the classification rules. In the construction of the tree, mainly two flavors needs to be considered, i.e., first is splitting criterion used and the second is overfitting technique which is applied here [14].

However the Decision trees can be an excellent tool in Machine Learning, it suffers from variety of disputes when implementing [15]. Analyzing those issues will help out to find the ways to improve the performance and accuracy of prediction. This paper analyses the issue called overfitting which have major impact on the performance and accuracy of the Decision Trees and provides excellent solution.

1. Decision tree building is in a top to bottom fashion, using recursive divide and conquer way. Initially, all the training instances were at the root of the tree.
2. Attributes needs to be categorical (if the attributes were continuously valued, they need to be discretised).
3. Input data is portioned recursively based on selected attribute.
4. Testing instances at every node were chosen based on the heuristic measure or with the statistical measure.

Proposed back-track pruned algorithm

Back-track pruned C4.5 technique is applied to construct a decision tree, with the available non-categorical attributes C1, C2 up to Cn, the categorical attribute C, and learning dataset T of instances.

Input:

1. The learning dataset ‘D’, and its set of training observations with respective class values.
2. Attribute list represented as A, which is the set of relevant candidate attributes.
3. Chosen splitting criteria technique.

Output: A decision tree.

Method:

STEP 1: Construct a node called ‘N’.

STEP 2: Return ‘N’ as a leaf node labeled with C, if all observations in the training dataset have the same class output value C.

STEP 3: Return N as leaf node labeled with majority class output value in training dataset, if attribute list is empty.

STEP 4: To find the ‘‘best” splitting criterion attribute, use selected splitting criteria method to training dataset in the form of order.

STEP 5: Name the node ‘N’ with the splitting criterion attribute.

STEP 6: Eliminate the splitting criterion attribute from the attribute list.

STEP 7: Loop: Each value of ‘j’ in the attribute splitting criterion.

- Consider Dj, the observations in training dataset fulfilling the attribute value j.
- If Dj has no observations (empty), then attach the leaf node with the popular class output value to the node ‘N’.
- Otherwise, Attach the node returned by the generate decision tree (‘Dj’, -attribute list, selected splitting criteria method) to the node ‘N’.

STEP 8: End for Loop.

STEP 9: Return node 'N'.

Back-track pruned C4.5 algorithm builds a decision tree from a set of training data using the concept of information entropy [16]. Building of the decision tree is to use the highest measure of entropy-based information gain for the heuristic information [17]. Back-track pruned C4.5 builds decision trees from a set of training data using the concept of information entropy [18]. The Entropy and Information Gain are calculated for pruning the decision tree. The attribute having maximum gain yields in nodes with the smallest entropy.

Conditions for stopping partitioning

- If every samples for a known node fit in the identical class.
- If there are no outstanding attributes for additional partitioning-mass selection is employed for classifying the leaf.
- If no samples absent.

Classification techniques have been concerned substantial attention together in machine learning and in the data mining investigate areas [19]. Even the tree works well on training set, it may not perform well on real data sets as few of its parts may unknown during training. This occurrence is called overfitting the training data which needs to be eliminated.

Steps in pruning:

Step 1: Deduce the decision tree from the training set, increasing the tree awaiting the training data is fit as well as likely and permitting overfitting to happen.

Step2: Alter the knowledgeable tree into an equivalent set of rules by creating one rule for each path from the root node to a leaf node.

Step 3: Trim all rule by eliminating any preconditions that result in enhancing its predictable accuracy.

Step4: Arrange the pruned rules by their predictable accuracy, and judge them in this succession while classifying the instances.

Overfitting can diminish the accuracy of a decision tree on real instances considerably. For example, in a trial using C4.5 performed on noisy training data, the final tree was found to be 10%-20% less accurate because of overfitting. Pruning with Back-track is a technique used to discover towering accuracy of prediction. This technique prunes the tree as fine in hopes of dropping overfitting. The approach for handling overfitting is Back-track pruning. The steps introduced in Back-track pruning are as follows

Steps in Proposed Back-track Pruning

Step 1: The complete tree is constructed perfectly by allowing overfitting to occur.

Step 2: The algorithm starts from bottom of the tree and evaluates each non-leaf node towards root node.

Step 3: Identify and eliminate the nodes or sub trees with no negative result on the correctness of the decision tree. Removing such type of nodes or sub trees is called Back-track pruning.

Step 4: The resultant tree obtained by eliminating these unnecessary subtrees functions like original tree and yields high accuracy.

UCI Hungarian database

The Hungarian database with 294 instances and fourteen mentioned attributes were applied for implementation. The detailed information about the 14 attributes has been given below:

1. Age in terms of years (age).

2. Sex in terms of Male and Female.

3. Chest pain type referred as chest_pain.

Value -1 means typical angina.

Value- 2 means atypical angina.

Value -3 means non-anginal pain.

Value -4 means asymptomatic.

4. Resting BP (Blood Pressure) means (trestbps) (in terms of mm Hg while appointing to the hospital)

5. Serum cholesterol referred as (chol) in mg/dl

6. Fasting blood sugar (fbs) greater than 120 mg/dl where (1=true; 0=false)

7. Restecg referred as (restecg) resting ecg results

8. Maximum heart rate referred as (thalach)

9. Exercise induced angina referred as (exang)

10. ST depression referred as (oldpeak)

11. Slope referred as (slope)

Value 1 means up sloping

Value 2 means flat

Value 3 means down sloping

12. The number of major vessels (0-3) colored by flourosopy referred as (ca)

13. (thal) 3 -normal; 6 -fixed defect; 7 - reversable defect

14. Predicted attribute, diagnosis of heart disease (angiographic disease status) referred as num

Value -0: less than 50 percent diameter narrowing.

Value -1: greater than 50 percent diameter narrowing.

Sample data set

The sample test dataset is given below in Table 1. It is obtained from the Hungarian database having 294 instances and 14

attributes. Here the dataset is partitioned in to two halves while the first half is used as Training set and the second half is used as testing data set.

Table 1. *Sample data set.*

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	male	typ_angina	145	233	t	left_vent_hyper	150	no	2.3	down	0	fixed_defe	'<50'
67	male	asympt	160	286	f	left_vent_hyper	108	yes	1.5	flat	30	normal	'>50_1'
67	male	asympt	120	229	f	left_vent_hyper	129	yes	2.6	flat	2	reversable	'>50_1'
37	male	non_anginal	130	250	f	normal	187	no	3.5	down	0	normal	'<50'
41	female	atyp_angina	130	204	f	left_vent_hyper	172	no	1.4	up	0	normal	'<50'
56	male	atyp_angina	120	236	f	normal	178	no	0.8	up	0	normal	'<50'
62	female	asympt	140	268	f	left_vent_hyper	160	no	3.6	down	2	normal	'>50_1'
57	female	asympt	120	354	f	normal	163	yes	0.6	up	0	normal	'<50'
63	male	asympt	130	254	f	left_vent_hyper	147	no	1.4	flat	1	reversable	'>50_1'
53	male	asympt	140	203	t	left_vent_hyper	155	yes	3.1	down	0	reversable	'>50_1'

Results and Discussion

Results of C4.5 and back-pruned C4.5 algorithm

Table 2. *Result obtained from back-pruned C4.5 and C4.5 algorithm.*

Measures	Existing C4.5 algorithm		Proposed Back-Track Pruned C4.5 algorithm	
Correctly Classified	60%		66.66%	
Incorrectly Classified	40%		33.33%	
Kappa statistic	0.2		0	
Mean absolute error	0.4233		0.4444	
Root mean squared error	0.6144		0.4714	
Relative absolute error	77.6111%		85.7143%	
Root relative squared error	112.6336%		90.4534%	
Coverage of cases (0.95 level)	70%		100%	
Mean rel. region size (0.95 level)	65%		100%	
Total No. of Instances	10		10	
Class	Values		Values	
	<50	>50_1	<50	>50_1
TP Rate	0.4	0.8	1	0
FP Rate	0.2	0.6	1	0
Precision	0.67	0.571	0.67	0
Recall	0.4	0.8	1	0
F-Measure	0.5	0.667	0.8	0
ROC Area	0.64	0.064	0.5	0.5

The Training and Test datasets belongs to Hungarian database with 294 records and fourteen attributes were applied on C4.5 and Back-Track Pruned C4.5 Classification algorithm. It is proved that that Back-Track Pruned C4.5 Classification algorithm is performing better accuracy than C4.5. The Back-Track Pruned C4.5 yields the accuracy of 66.66% while the C4.5 gives 60%. The Incorrectly classified Instances of Back-Track Pruned C4.5 is 33.33% and C4.5 is 40%. The other measures of Back-pruned C4.5 and C4.5 Algorithm is also Tabulated in above Table 2.

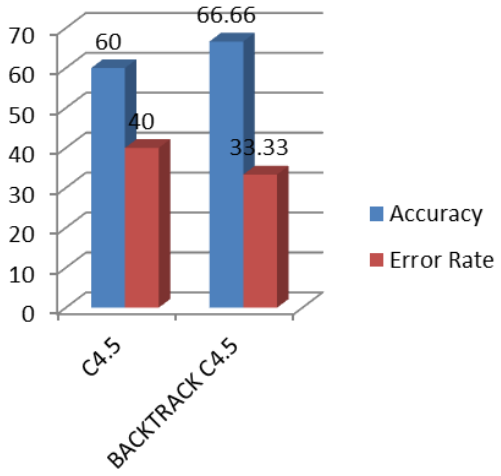

Figure 1. *Accuracies and error-rate of back-track pruned C4.5 and C4.5.*

The Figure 1 compares the accuracies of C4.5 and Back-Track Pruned C4.5 algorithms. It is proved that the Accuracy of Back-Track Pruned C4.5 is better than C4.5 Classification

algorithm. Also the incorrectly classified instances were minimum in Back-Track Pruned C4.5 than C4.5. The other measures like Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, Coverage of cases (0.95 level), Mean rel. region size (0.95 level), Total Number of Instances, Class, TP Rate, FP Rate, Precision, Recall, F-Measure and ROC Area were measured for analysis of performance.

Conclusion

The heart disease dataset from UCI Machine learning Hungarian database with 294 records and 14 attributes has been implemented with C4.5 and Back-track Pruned C4.5 Classification algorithm. Back-pruned C4.5 performs better than C4.5 with 66.66% of improved accuracy. Here, an optimal solution has been provided for overfitting of the decision Tree. It supports both continuous and discrete instances. The Missing attributes are not included for gain and entropy computations, so that reliability is maintained. This approach minimizes the overhead, memory required, size of the tree, time taken and results in improved accuracy by removing the branches not useful with minimized learning cost. As the paper focus is completely on prediction of the heart disease further, the same can be extended to predict the survival rate of the heart attack patients.

References

1. Chakrabarti S. Data mining: know it all. Burlington, MA: Elsevier/Morgan Kaufmann Publishers, 2009.
2. Xiaoliang Z, Jian W, Hongcan Y, Shangzhuo W. Research and application of the improved algorithm C4. 5 on decision tree. InTest and Measurement, 2009. ICTM'09. International Conference on, 184-187.
3. Jothikumar R, Sivabalan RV. Performance Analysis on Accuracies of Heart Disease Prediction System Using Weka by Classification Techniques. AJBAS 2015; 9: 741-749.
4. Jothikumar R, Sivabalan RV, Sivarajan E. Accuracies of j48 weka classifier with different supervised weka filters for predicting heart diseases. ARPN J Eng Applied Sci 2015; 10: 7788-7793.
5. Jothikumar R, Sivabalan RV, Kumarasen AS. Data Cleaning Using Weka For Effective Data Mining In Health Care Industries. Int J Appl Eng Res 2015.
6. Xiaohu W, Lele W, Nianfeng L. An application of decision tree based on id3. Physics Procedia 2012; 25: 1017-1021.
7. Jabbar MA, Deekshatulu BL, Chndra P. Alternating decision trees for early diagnosis of heart disease. In Circuits, Communication, Control and Computing (I4C), 2014 International Conference on, 322-328.
8. Bouali H, Akaichi J. Comparative Study of Different Classification Techniques: Heart Disease Use Case. In Machine Learning and Applications (ICMLA), 2014 13th International Conference on, 482-486.
9. Krishnaiah V, Narsimha G, Subhash N. Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review. Int J Comput Appl 2016; 136: 43-51.
10. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, 108-115.
11. Masethe HD, Masethe MA. Prediction of heart disease using classification algorithms. In Proceedings of the World Congress on Engineering and Computer Science 2014.
12. Boryczka U, Kozak J. Enhancing the effectiveness of Ant Colony Decision Tree algorithms by co-learning. Appl Soft Comput 2015; 30: 166-178.
13. Hacibeyoglu M, Arslan A, Kahramanli S. A hybrid method for fast finding the reduct with the best classification accuracy. Adv Electric Comput Eng 2013; 13: 57-64.
14. Mahmood AM, Kuppa MR. Early detection of clinical parameters in heart disease by improved decision tree algorithm. In Information Technology for Real World Problems (VCON), 2010 Second Vaagdevi International Conference on, 24-29.
15. Ruggieri S. Efficient C4. 5 [classification algorithm]. IEEE Transact Knowledge Data Eng 2002; 14: 438-444.
16. Idri A, Kadi I. Evaluating a decision making system for cardiovascular dysautonomias diagnosis. SpringerPlus 2016; 5: 81.
17. Shi G. Data mining and knowledge discovery for geoscientists, Elsevier Amsterdam, Netherlands 2013.
18. Shouman M, Turner T, Stocker R. Using data mining techniques in heart disease diagnosis and treatment. In Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on, 173-177.
19. Purdilă V, Pentiuc ŞG. Fast decision tree algorithm. Adv Electric Comput 2014; 14: 65-68.

***Correspondence to**

Siva Balan

Department of Master of Computer Application

Noorul Islam University

Tamil Nadu

India