Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 65 (2015) 459 – 468

Procedia Computer Science

International Conference on Communication, Management and Information Technology (ICCMIT 2015)

Feature Analysis of Coronary Artery Heart Disease Data Sets

Randa El-Bialy[1] , Mostafa A. Salamay[2], Omar H. Karam[3] and M.Essam Khalifa[4]

[1]*British University in Egypt (BUE), Cairo, Egypt, Randa.Elbialy@Bue.edu.eg*
[2]*British University in Egypt (BUE), Cairo, Egypt, mostafa.salama@bue.edu.eg*
[3]*British University in Egypt (BUE), Cairo, Egypt, omar.karam@bue.edu.eg*
[4]*Ain Shams University, Cairo, Egypt, esskhalifa49@gmail.com*

Abstract

Data sets dealing with the same medical problems like Coronary artery disease (CAD) may show different results when applying the same machine learning technique. The classification accuracy results and the selected important features are based mainly on the efficiency of the medical diagnosis and analysis. The aim of this work is to apply an integration of the results of the machine learning analysis applied on different data sets targeting the CAD disease. This will avoid the missing, incorrect, and inconsistent data problems that may appear in the data collection. Fast decision tree and pruned C4.5 tree are applied where the resulted trees are extracted from different data sets and compared. Common features among these data sets are extracted and used in the later analysis for the same disease in any data set. The results show that the classification accuracy of the collected dataset is 78.06% higher than the average of the classification accuracy of all separate datasets which is 75.48%.

Peer-review under responsibility of Universal Society for Applied Research

Keywords: Data Mining, Fast Decision Tree Learning Algorithm, Decision Trees.

1. Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. In healthcare, Data mining is a field of high importance and has become increasingly effective and essential. The healthcare industry today generates large amounts of complex data concerning patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction and enabling support for cost-savings and decision-making. Data mining provides a set of tools and techniques that can be applied to the data to achieve these goals. Coronary heart disease is considered a fatal illness that causes death to over a million patients every year. Nearly half the patients diagnosed with CAD will eventually die from the disease. Devastatingly, 335,000 of CHD patients will die of a heart attack in an emergency department or before they even reach the hospital. According to the American Heart Association, over 7 million Americans have suffered a heart attack in their lifetime. When plaque is built inside the coronary arteries, they narrow these arteries making them unable to carry oxygenated blood to the heart muscle causing the well-known symptoms of CAD such as chest pain (angina) and shortness of

Peer-review under responsibility of Universal Society for Applied Research
doi:10.1016/j.procs.2015.09.132

breath [1, 2]. A system to identify the probability of the existence of coronary heart disease was presented in [3] by Patel et al. In that work, the parameters were divided into two levels, each with its weight. In order to derive a decision, a neuro-fuzzy integrated approach was implemented taking into consideration both parameter levels to reach a final decision. The choice of a fuzzy approach is advantageous in that it reduced error rate and in turn enhanced efficiency. In [4] Chitra R. et al. developed a computer aided heart disease prediction system. The system is intended to help physicians diagnose heart disease. Their conclusion was that, data mining can play a major role in heart disease classification. It is further noted that due to the nature of the issue, feature reduction techniques may be needed in order to enhance the efficiency and search time of the classifier in question. Nidhi Bhatlaet al. [5] discussed the results of applying various data mining techniques that have been closely associated with heart disease diagnosis in recent years. The techniques that were surveyed by that work include Neural Networks, Decision Trees and Genetic Algorithms. The work of Srinivas K. et al. [6] explored various data mining aspects that went beyond the scope of classification; namely clustering, association rule mining and time series analysis. The work focused on the prediction of various combinations of certain heart disease attributes. This work has a high potential for further expansion and enhancement. In recent decades, many experts have tried to make CAD diagnosis using computer-aided techniques, such as neural networks [7], the Bayesian model and decision tree 6, support vector machine [8], and the naive Bayes classifier [9]. The aim of this work is to apply an integration of the results of the machine learning analysis applied on different data sets targeting the CAD disease. This will avoid the missing, incorrect, and inconsistent data problems that may appear in the data collection, finding a set of attributes which are really important for this disease predication in order to improve the performance of the classifier, help physicians to successfully diagnose CAD. This current paper is organized as follows. Section 2 contains previous work related to the topic. In Section 3 the Decision Tree Integration Model / Methodology is presented. Results and Discussion are in Section 4. The conclusion is provided in Section 5.

2. Related Work

A Decision Tree is a decision support system that uses a tree-like graph where each node denotes a test on an attribute value and each branch represents an outcome of the test while tree leaves stand for classes or class distributions. Decision trees have many advantages such as i) construction of decision trees is unique for every dataset (not complicated), ii) end users can understand them easily, iii) a variety of input data can be handled e.g. nominal, numeric and textual, iv) the ability to process missing values or invalid/erroneous datasets and v) high performance is achieved with a little effort. The classification is performed by starting at the root node for each new record to be classified, and depending on the results at each consecutive node, a leaf node is reached thus determining the class for that record or determining a probability distribution for the possible classes [10, 11]. There are many algorithms for building decision trees such as ID3 and C4.5. There is also the Fast Decision Tree. Quinlan Ross introduced the ID3 (Iterative Dichotomiser) in 1986 [12]. It determine the splitting attribute according to the information gain measure; for each and every attribute in the data set the information gain the attribute with the highest information gain is identified and assigned as a root node for the tree. Declaring the selected attribute as a root node and representing the possible attribute values as arcs, all possible outcome instances are tested to check whether they fall under the same class or not. The node is represented with a single class name if the instances are falling under the same class; otherwise the splitting attribute to classify the instances is chosen. The main disadvantages facing ID3 are: accepting only categorical attributes, giving inaccurate results when noise exists, testing only one attribute at a time for making decisions and pruning is not supported [13,14].

2.1. C4.5 Algorithm

C4.5 is an extension and improved version of the ID3 algorithm developed by Quinlan Ross 12. Pruning is the key feature/step to overcome the over fitting problem in the ID3 algorithm. Both categorical and continuous attributes in the data set are handled. A Gain Ratio is used as the splitting criterion, and then according to the selected threshold the attribute values are split into two partitions: all the attributes with

values above the threshold are dealt with at one child and the remaining at another child. For each attribute the gain ratio is calculated. The attribute with the maximum gain ratio will be the root node. In order to improve the accuracy of the C4.5 algorithm, pessimistic pruning is used to eliminate unnecessary branches from the decision tree [13, 14]. The time complexity of C4.5 is O($m.n^2$), where m is the size of the training data and n is the number of attributes [15].

2.2. Fast Decision Tree

The Fast Decision Tree algorithm is an improved version of the C4.5 algorithm developed by Jiang Su and Harry Zhang. The Fast decision tree algorithm has a time complexity of O(m.n), where m is the size of the training data and n is the number of attributes. According/ with respect to accuracy fast decision tree algorithm and c4.5 performs competitively. Using a large number of textual data sets allows the fast decision tree algorithm to perform significantly better and faster than C4.5. In the Fast Decision Tree algorithm, the independent information gain (IIG) is used as a splitting criterion, the IIG must be calculated for each candidate attribute. The attribute with the maximum IIG will then be the root node. According to the maximum IIG, the attribute values are split into two partitions: all the attributes with values Section headings above the threshold is dealt with at one child and the remaining at another child. The time complexity for the Fast Decision Tree using IIG, O(m.n), is lower than the time complexity of C4.5 which is O($m.n^2$) [15].

3. Decision tree integration model/ Methodology

The heterogeneity datasets produce heterogeneity data types. These datasets provide information complementary to the medical diagnosis, and so it can enhance the classification accuracy percentages [16]. The diagnosis of patients is based on the accuracy of the medical data collected. Different problems exist for the medical data which constitutes a real challenge in the field of medical informatics. These problems can be summarized as follows [17,18]: a) Incorrect, sparse and temporal information, b) Small sized samples, c) Measuring attribute errors, d) Manual data collection inconsistencies, e) Missing values, f) Professionalism of medical analysts / practitioners/ technicians in diagnosing the disease, g) Accuracy of machines or instruments used in the diagnosis [19]. The proposed model: In this work, the required is to use the complementary information existing in different data sets to deal with these problems. The model of integrating the different data sets is proposed as shown in Figure 1 as follows:

1. Data sets that deal with the same medical problem are gathered from different resources. The number and type of attributes may differ from one data set to the other, but the classification category or target class label must be the same.
2. The data sets collected in the current work, are four datasets for coronary artery heart disease: Cleveland Heart disease, Hungarian heart disease, V.A. heart disease and statlog project heart disease which consists of 13 features. All were downloaded from the UCI repository [20].
3. Two classification techniques are applied; C4.5 and the fast decision tree, to extract the corresponding decision tree for each of the data sets and then to compare and detect the classification accuracy percentages.
4. Five Common features among all the decision trees are extracted. For example, the Cleveland dataset as shown in figure two and figure three for C4.5 and fast decision trees classifiers. After building the C4.5 and Fast decision Tree significant difference in tree size and processing time between the fast decision tree and the C4.5 are shown in table 2, therefore the slight difference in accuracy between them can be ignored.
5. These five most common features are used to build a new integrated data set from all the input data sets. The new pruned collected dataset includes only these extracted common features. These features shows highest information gain values that are selected to avoid over-fitting in the new generated and integrated decision tree.
6. C4.5 and the fast decision trees are applied to this integrated data set, and a new decision trees are generated. In the processes of creating these new trees, another layer of pruning or feature selection, is applied. Only four features frequently appear in the new decision tree as shown in figure 8. The fast decision tree selects only ca, age, cp and thal.

7. It may be noticed that for small number of instances, the classification accuracy of the decision tree applied on all features will be higher than that of the four-featured decision tree. But for big data of large number of instances, the tree size and so the processing time will be apparently low in the case of the four-featured data sets.
8. One last step is applied to show the efficiency of this proposed model, which is the comparison between the average of classification accuracies of all data sets in separate and the classification accuracy of the integrating decision tree.

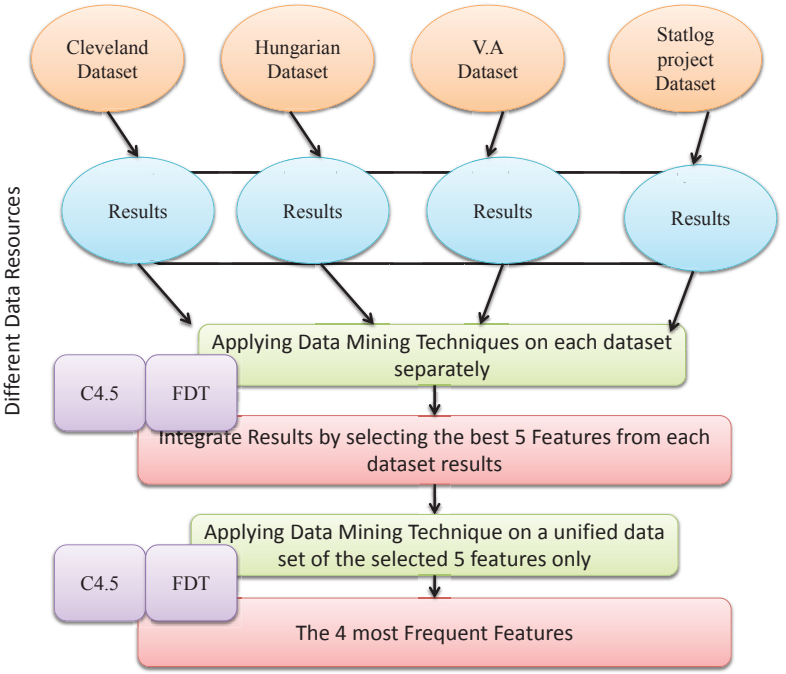

Fig. 1. C4.5 decision Tree for Cleveland dataset

4. Results and Discussions

This paper shows the analysis of different decision trees such as: C4.5 and Fast Decision Tree, which can be helpful for medical analysts or practitioners for accurate diagnosis by carefully selecting the associated features. The results were analyzed using Weka tool on the data using 10- fold cross validation to test the accuracy and time complexity of C4.5 and Fast Decision Tree classifiers. The following table shows the characteristics of selected datasets related to the heart disease domain. The data was collected from the four following locations: Cleveland Clinic Foundation , Hungarian Institute of Cardiology, V.A. Medical Center, Long Beach, CA and statlog project. The instances number for each Dataset as follows: Cleveland:303, Hungarian:294, Long Beach VA:200 and Statlog project:270. These datasets contains 14 attributes which have been extracted from a larger set of 75 attributes.

Table 1. Dataset characteristics

Dataset	# of attributes	# of classes	# of instances	Missing values
Cleveland heart disease	14	2	303	No
Hungarian heart disease	14	2	294	yes
V.A heart disease	14	2	200	yes
Statlog project	14	2	270	yes

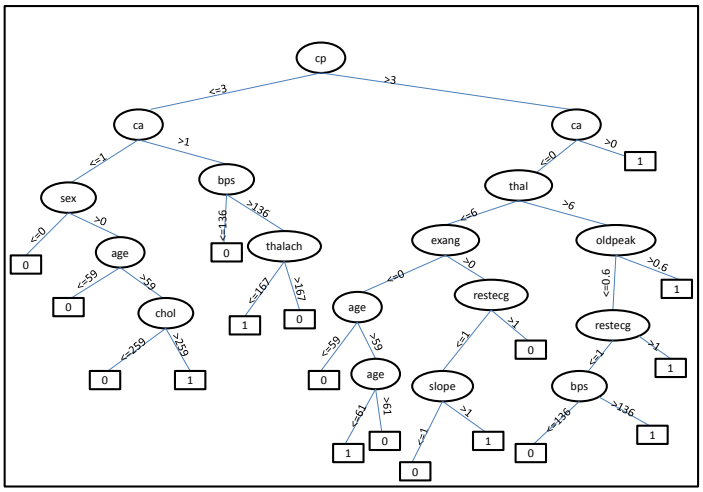

Fig. 2. C4.5 decision Tree for Cleveland dataset

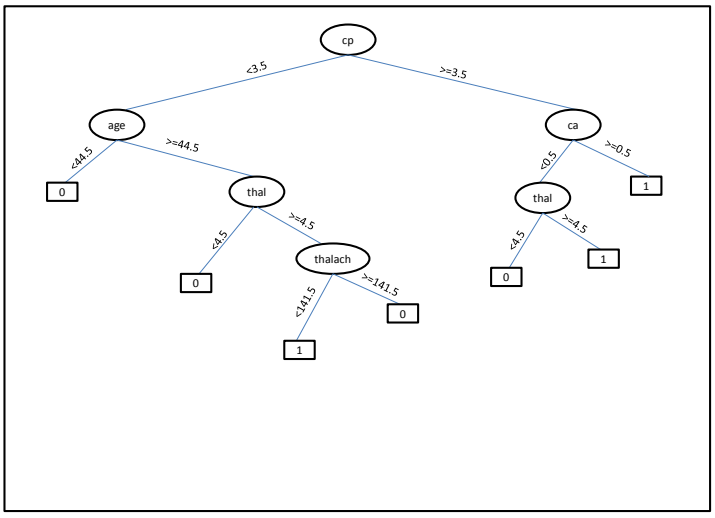

Fig. 3. Fast Decision Tree (REP Tree in Weka) for Cleveland dataset

In Figures 2 and 3 there are 5 common features between the two decision trees: cp, age, ca, thal, and

thalach.
Association rules for figure 2:
If cp <= 3 and ca ¡= 1 and sex <= 0 then 0.
If cp<=3 and ca<=1 and sex>0 and age<=59 then 0.
If cp<=3 and ca<=1 and sex>0 and age>59 and chol<=259 then 0.
If cp<=3 and ca<=1 and sex>0 and age>59 and chol>259 then 1.
If cp<=3 and ca>1 and bps<=136 then 0.
If cp<=3 and ca>1 and bps>136 and thalach<=167 then 1.
If cp<=3 and ca>1 and bps>136 and thalach>167 then 0.
Association rules for figure 3:
If cp<3.5 and age<44.5 then 0.
If cp<3.5 and age>=44.5 and thal<4.5 then 0.
If cp<3.5 and age>=44.5 and thal>=4.5 and thalach<141.5 then 1.
If cp<3.5 and age> =44.5 and thal>=4.5 and thalach>=141.5 then 0. As noticed from the association rules from the C4.5 DT, there exists a high correlation between cp, ca and age. These three attributes appear together more than any other sub-set of features. And for the association rules from fast DT, there exists a high correlation between cp, age, and thal features. The intersection between both cases shows that the correlation between ca and age is more highlighted. Later on in the results, the correlation between ca and age appears to be dominant in the decision trees generated from the integrated data set as appears in figure 8. All experiments were performed on a core i5 with 2.4GHZ CPU and 4.00 G RAM.

Table 2. Classifier accuracy, execution time and tree size to build the decision trees

Dataset	Decision Tree	Accuracy %	Time/sec	Tree size
Cleveland	C4.5	78.54	0.01	43
	FDT	77.55	0.01	13
Hungarian	C4.5	78.57	0.03	27
	FDT	78.23	0	5
V.A	C4.5	71.5	0.13	23
	FDT	69.5	0	11
Statlog	C4.5	76.6	0.01	35
	FDT	76.6	0.01	11

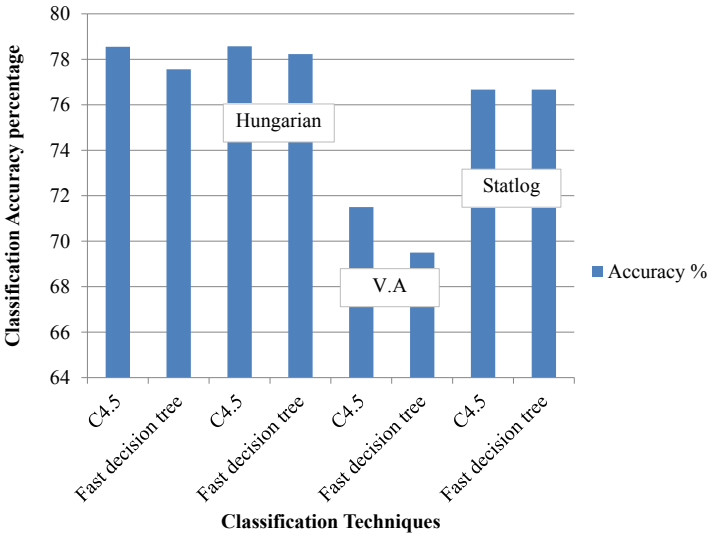

Fig. 4. Classification technique accuracy

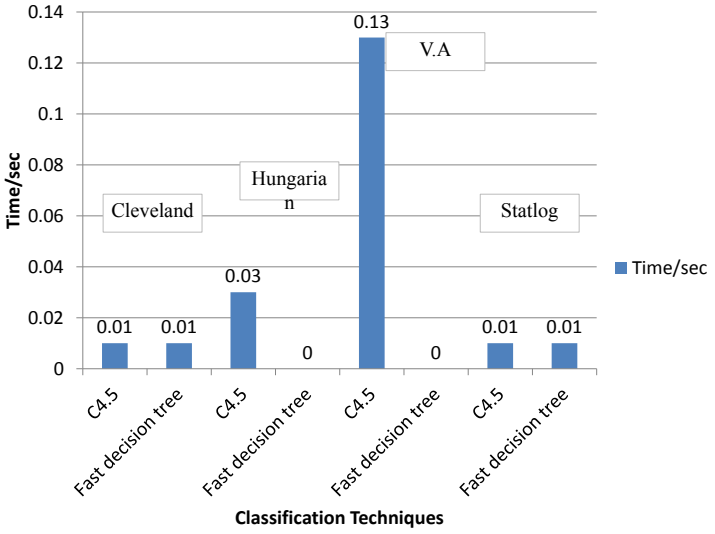

Fig. 5. Classification technique execution time

Table 3. Best selected features for heart diseases

Dataset	Decision Tree	Best selected feature
Cleveland	C4.5	Cp, Thal, Ca, Thal
	FDT	Cp, Age, Ca, Thal, Thalach
Hungarian	C4.5	Exang, Oldpeak, Sex, Cp, Slop, Thalach
	FDT	Cp, Oldpeak
V.A	C4.5	Cp, Age, Exang, Fbs, Sex
	FDT	Cp, Age, Chol
Statlog	C4.5	Thal, Cp, Ca, Exang, Age
	FDT	Ca, Thal, Cp,Thalach, Oldpeak

The abbreviations used in table 3 are: cp: chest pain type (4 types): Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic trestbps: resting blood pressure (in mm Hg on admission to the hospital), chol: serum cholestoral in mg/dl, fbs: fasting blood sugar ¿ 120 mg/dl) restecg: resting electrocardiographic results thalach: maximum heart rate achieved , exang: exercise induced angina , oldpeak = ST depression induced by exercise relative to rest, slope: the slope of the peak exercise ST segment, ca: number of major vessels (0-3) colored by flourosopy and thal: 3 = normal; 6 = fixed defect; 7 = reversable defect.

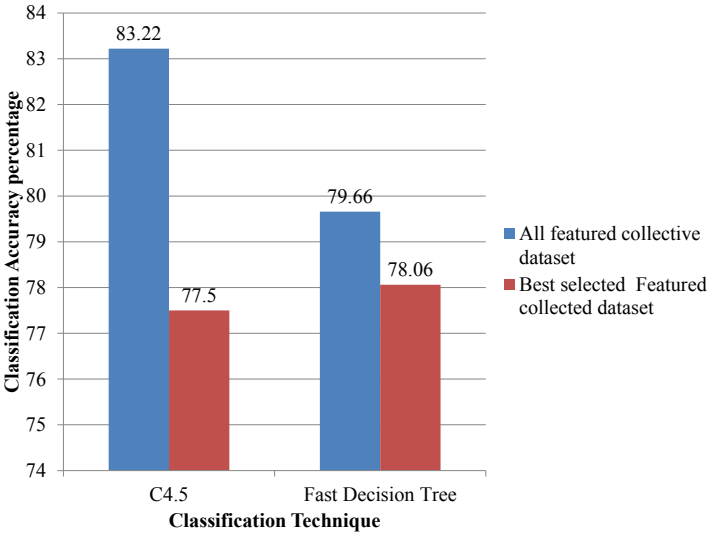

Fig. 6. Results for the comparison between best feature and collective dataset

Table 4. Results for the comparison between best feature and average of collected dataset

Decision Tree	Average of each separate dataset accuracy %	Best selected featured collected dataset accuracy %
C4.5	76.30	77.50
FDT	75.48	78.06

The collected dataset includes the data of the selected attributes from all the datasets. The classification accuracy of this collected dataset (77.5%, 78%) is higher than the average of classification accuracy of all datasets (76.3%, 75.4%) as shown in fig 6. This shows that the accumulation of knowledge resulted from the

integration of all datasets could enhance the diagnosis process. The results in table 4 prove that the proposed model solves the problem of inaccuracy that appears in different dataset resources.

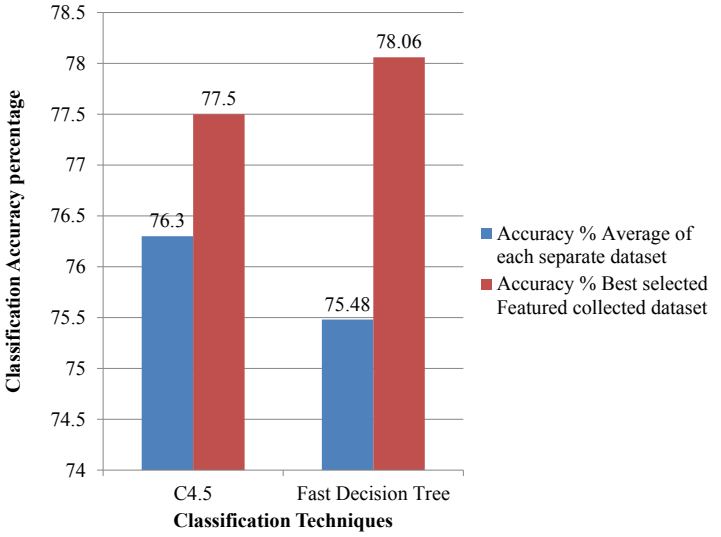

Fig. 7. Results for the comparison between best feature and AVG of collective dataset

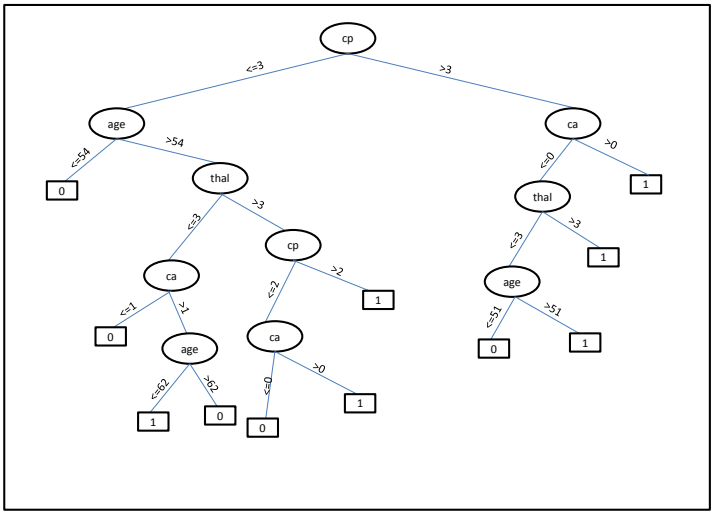

Fig. 8. C4.5 for the selected best feature

5. Conclusions

Data sets dealing with the same medical problems like Coronary artery disease (CAD) may show different results when applying the same machine learning technique. The classification accuracy results and the selected important features are based mainly on the efficiency of the medical diagnosis and analysis. The aim of this work is to apply an integration of the results of the machine learning analysis applied on

different data sets targeting the CAD disease. Fast decision tree and pruned c4.5 tree are applied where the resulted trees are extracted from different data sets and compared. Common features among these data sets are extracted and used in the later analysis for the same disease in any data set. The results show that the classification accuracy of the collected dataset is higher than the average of the classification accuracy of all separate datasets. The future work is to try different ensemble techniques on different data sets targeting the same problem. The diversity of resources will provide better performance in knowledge extraction and a clear understanding of the measuring and collecting data problems.

References

[1] Webmd.com. "Risk factors for heart disease," retrieved 20/8/2014 from http://www.webmd.com/heart-disease/risk-factors-heart-disease.
[2] National Heart, Lung, and Blood Institute, "What is coronary heart disease?" retrieved 20/8/2014 from http://www.nhlbi.nih.gov/health/healthtopics/ topics/cad/.
[3] Bahadur Patel, Ashish Kumar Sen D P, Shamsher Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level", International Journal Of Engineering And Computer Science, 2013, 2319-7242, p. 2663-2671.
[4] Chitra R and Seenivasagam V, "Review of Heart Disease Prediction System Using Data Mining Technique and Hybrid Intelligent Techniques", ICTACT journal on Soft Computing , 2013, 2229-6956(ONLINE), p. 605-608.
[5] Nidhi Bhatla and Kiran Jyoti, "An Analysis of Heart Disease Prediction Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), 2012, 2278-0181.
[6] Srinivas K, Raghavendra Rao R, Govardhan A, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", The 5th International Conference on Computer Science & Education Hefei, 2010, China. August 24-27.
[7] S. B. Patil and Y. S. Kumaraswamy, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," European Journal of Scientific Research, 2009, p. 642-656.
[8] I. Babaoglu, O. K. Baykan, N. Aygul, K. Ozdemir, and M. Bayrak, "Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization," Expert Systems with Applications,2009,p. 2562-2566.
[9] M. C. Tu, D. Shin, and D. Shin, "Effective diagnosis of heart disease through bagging approach," in Proceedings of 2nd International Conference on Biomedical Engineering and Informatics,2009 ,Tianjin, China, pp. 1-4.
[10] Neeraj Bhargava, Girjas Sharma, Ritu Bhargava, Manish Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering , 2013, 2277128x, p. 114-1119.
[11] Nikita Jain, Vishal Srivastava," Data Mining Techniques: a survey paper", IJRET: International Journal of Research in Engineering and Technology, 2013, 2321-7308, p. 116-119.
[12] Quinlan, J.R, "Induction of decision trees". Journal of Machine Learning ,1986, pp. 81-106.
[13] D.Lavanya, K.Usha Rani, "Performance Evaluation of Decision Tree classifiers on medical datasets", International Journal of Computer Applications, 2011 , 0975-8887, p. 1-4.
[14] Venkatadri, & Lokanatha, "A Comparative Study of Decision Tree Classification Algorithms in Data Mining." International Journal of Computer Applications in Engineering, Technology and Sciences, 2011, p. 230-240.
[15] Jiang Su and Harry Zhang, "a Fast Decision Tree Learning Algorithm", AAAI'06 Proceedings of the 21st national conference on Artificial intelligence, 2006 , 978-1-57735-281-5, p.500-505.
[16] Jiang Su and Harry Zhang, "a Fast Decision Tree Learning Algorithm", AAAI'06 Proceedings of the 21st national conference on Artificial intelligence, 2006 , 978-1-57735-281-5, p.500-505.
[17] Inci Aksoy, Bertan Badur, Sona Mardikyan, "Finding hidden patterns of hospital infections on newborn: A data mining approach", journal of Istanbul University Journal of the School of Business Administration, 2010, 1303-1732, p. 210-226.
[18] V. Speckauskiene, A. Lukosevicius, "Methodology of Adaptation of Data Mining Methods for Medical Decision Support: Case Study", journal of Electronics and Electrical Engineering, 2009 , 1392-1215, p. 25-28.
[19] Erik Liljegren,, "Usability in a medical technology context assessment of methods for usability evaluation of medical equipment", International Journal of Industrial Ergonomics,2006, p. 345-352.
[20] The UCI machine learning repository [online]. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.