Treatment Response Classification in Randomized Clinical Trials: A Decision Tree Approach

P Venkatesan, N R Yamuna

Department of Statistics, National Institute for Research in Tuberculosis, (ICMR), Chennai-600 031, India
venkaticmr@gmail.com, nryamuna@yahoo.co.in

Abstract

Decision Trees are a subfield of machine learning technique within the larger field of artificial intelligence. It is a supervised learning technique for classification and prediction. The decision trees are widely used for outcome prediction under various treatments for disease cure, prevention, toxicity and relapse. The aim of the paper is to compare the decision tree algorithms in classifying tuberculosis patient's response under randomized clinical trial condition. Classification of patient's responses to treatment is based on bacteriological and radiological methods. Three decision tree approaches namely C4.5, Classification and regression trees (CART), and Iterative dichotomizer 3 (ID3) methods were used for the classification of response. The result shows that C4.5 decision tree algorithm performs better than CART and ID3 methods.

Keywords: Decision tree. Data mining.CART.C4.5. ID3.

1. Introduction

Tuberculosis diseases are an infectious disease caused by Mycobacterium tuberculosis, which normally affects the lungs. One-third of the world population is presently infected with tuberculosis, and 5–10% of these can be expected to develop active illness at some point of time in their lives (Schlipköter and 2010). Short-course chemotherapy is a well-known method for the treatment of pulmonary tuberculosis (Jawahar 2004; Tuberculosis Research Centre Madras 1983). Patient classification is a decision-making method that has been applied widely to the identification and diagnosis of illness (James 2005). Decision tree is one of the most widely used supervised classification technique(Utgoff and Brodley, 1990). It has been used in medical and health care applications for more than 3 decades and has been shown to be a powerful classification tool (Podgorelec et al. 2002). Some of the classification techniques are used to identify the groups of individuals with particular outcomes, while other techniques identify groups of individuals who are at risk of developing specific outcomes. Compared to other classification methods, the decision tree technique is attractive because it clearly shows how to reach a decision, and also it is easy to construct automatically from labeled instances. Decision tree has many different approaches and algorithms to deal with the problem of building a decision tree model. However, the way of selecting splitting attributes and splitting criterion are different in decision trees (Han and Kamber 2006). For example CART uses binary recursive partitioning concept whereas C4.5 and ID3 use non-binary concept.Numerous authors have published their results on comparison of the classification techniques in several areas of medicine and others (Li et al. 2010; Kim 2010). Ture et al. (2005) compared the various classification techniques to predict hypertension groups and controls.

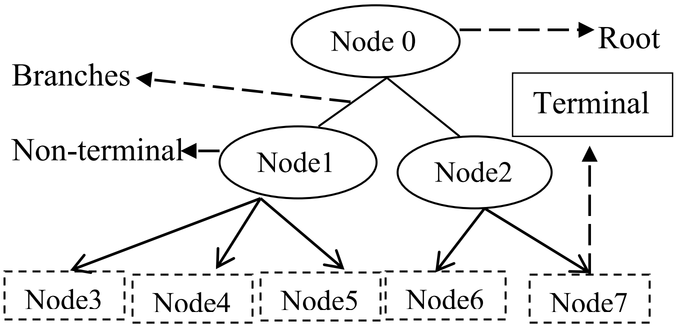
In this paper, three well known decision tree methods namely ID3 (Quinlan 1979) C4.5 (Quinlan 1993) and CART (Breiman et al. 1984) were used to classify treatment response under control clinical trials.This organization of the paper is as follows: Section 2 reviews briefly the decision tree methods (ID3, C4.5 and CART) for the classification. It also deals with the important aspects of the database considered. Section 3 presents the application of the classification method and comparison of the results. Section 4 deals with the discussion and conclusion.

2. Material and methods

2.1 Decision trees

Classification is the most familiar and most popular data mining technique. In data mining, a decision tree is a predictive model which can be used to represent both classification and regression tree. Decision tree used a "divide and conquer" technique to split the data into subsets. The result of decision treeis in the form of rule-based or tree-based. A simple construction of a decision tree is shown in Figure 1.

Fig.1. *Structure of decision tree*



The top most nodes in the decision tree is called root node (Node 0). The root node contains the complete data set and other nodes correspond to subgroups of the data set. The root node forms the basis of building a decision tree, which consists of two essential nodes such as non-terminal nodes (Node 1 and 2) and terminal nodes or leaf nodes (Nodes 3, 4, 5, 6 and 7). Non-terminal nodes represent tests on one or more attributes and terminal nodes return the decision outcomes.

To construct a decision tree, choosing splitting attributes plays a major role. The choice of attribute involves not only an examination of the data in the training set but also the informed input of domain experts. To improve the performance of applying the tree for classification, a balanced tree with the fewest levels is desirable. The creation of the tree definitely stops when the training data are perfectly classified. Once the tree is constructed, some modifications to the tree might be needed to improve the performance of the tree during the classification phase. The pruning phase might remove redundant comparisons or remove subtrees to achieve better performance.

There are many advantages to the use of decision trees for classification. Decision trees are easy to use and efficient. Rules can be generated that are easy to interpret and understand. They scale well for large databases because the tree size is independent of the database size. Each record in the database must be filtered through the tree. Trees can be constructed for data with many attributes. Disadvantages also exist for decision tree algorithms. First, they don't easily handle continuous data. These attributes domain must be divided into categories to be handled. Handling missing data is difficult because correct branches in the tree could not be taken. Since the decision tree is constructed from the training data, overfitting may occur. This can be overcome via tree pruning.

2.2 Iterative dichotomizer 3 (ID3)

The ID3 algorithm for building a decision tree was first developed by Quinlan (1979). It is a top-down approach starting with selecting the best attribute to test at the root of the tree. The selection of the best attribute in ID3 is based on an information theory approach or entropy (Quinlan 1986). Entropy is a measure of the amount of uncertainty present in a set of data. When all data in a set belong to a single class, there is no uncertainty and hence, the entropy is zero. In general,the value of entropy falls between 0 and 1and reaches a maximum when the probabilities are all the same.Given a set S, containing two examples ('positive' and 'negative') of target concept, the entropy of a set S relative to this binary classification is defined as:

$$E(S) = -p_{(positive)} \log_2 p_{(positive)} - p_{(negative)} \log_2 p_{(negative)} \qquad (1)$$

Where $p_{(positive)}$ *and* $p_{(negative)}$ are the fraction of positive and negative examples in S.

Information gain calculates the expected reduction in entropy. Gain (S, X) of an attribute X, relative to a collection of examples S is,

$$gain(S, X) = E(S) - \sum_{v \in values\,(X)} \frac{|S_v|}{|S|} \times E(S_v) \qquad (2)$$

where,
values (*X*) = set of all possible values for attribute *X*.
S_v= subset of S for which attribute *X* has value *v* (ie.,S_v = $\{s \in S \mid X(s) = v\}$)
E(S) = entropy of the original collection *S*.

ID3 select splitting attributes with the highest information gain. It can handle only nominal attributes. Modifications and improvements on the ID3 algorithm culminated into the popular C4.5 algorithm.

2.3 C4.5

Quinlan (1993) proposed C4.5 decision tree algorithm which depends on ID3 algorithm. It can perform test on both nominal and numerical attributes. The use of the gain ratio was one of various developments that were made to ID3 over a number of years. Further improvements include methods for dealing with numeric attributes, missing values, noisy data and generating rules from trees (Quinlan 1996). In general, when a decision tree is built, missing data are simply ignored. The gain ratio is calculated by looking only at the other records, which have a value for that attribute. In order to classify a record with a missing attribute value, the attribute values for the other records can be used to predict the same.

If S is the set of training data denoting a concept with *c* classes, *f(C$_j$, S)* is the frequency of class *C$_j$* occurring in that set, then the expected information required to classify a given class in S is:

$$Info(S) = -\sum_{j=1}^{c} \frac{f(C_j,S)}{|S|} \log_2\left(\frac{f(C_j,S)}{|S|}\right) \tag{3}$$

when an attribute, A, with *v* values, has been selected as a test attribute, then the expected information needed to identify a class under that test is:

$$Info_A(S) = \sum_{i=1}^{v} \frac{|S_i|}{|S|}\, info(S_i) \tag{4}$$

where$S_1, S_2, \ldots, S_v$ is the subset of S all of whose instances possess value *i* for attribute A. The information gain is the difference between the expected information needed to identify a class with and without the test on attribute A:

$$gain(A) = Info(S) - \sum_{i=1}^{v} \frac{|S_i|}{|S|} \times Info(S_i) \tag{5}$$

The attribute giving the maximum information gain is selected as the current split. ID3 used information gain criterion (Equation.6) to select the test for partition. However, the gain criterion is biased towards the high frequency data. To restructure this problem, C4.5 normalizes the information gain by the amount of the potential information generated by dividing T into *v* subsets:

$$split\,info(A) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \log_2\left(\frac{|S_i|}{|S|}\right) \tag{6}$$

C4.5 selects the test to partition the set of available cases is defined as:

$$gain\,ratio(A) = \frac{gain(A)}{split\,info(A)} \tag{7}$$

C4.5 selects the test that maximizes gain ratio value. The difference between ID3 and C4.5 algorithm is that ID3 uses binary splits, whereas C4.5 algorithm uses multi-way splits. In order to reduce the size of the decision tree, C4.5 uses post-pruning technique; whereas an optimizer combines the generated rules to eliminate redundancies. The improved version of C4.5 is C5.0, which includes cross-validation and boosting capabilities.

2.4 Classification and Regression Trees (CART)

CART is a non-parametric decision tree algorithm developed by Breiman et al. in 1984(Breiman et al. 1984). CART produces either classification or regression trees, based on whether the response variable is categorical or continuous. This methodology is a binary recursive partitioning procedure (Lewis 2000), which always split the node into only two nodes. The partitioning procedure is repeated for every node of the data until it becomes the terminal node. There are three important steps in CART. (i)Tree growing process: it is based on the recursive partitioning algorithm to select the variables using splitting criterion. In CART,gini criterion is used for determining the best split (James et al. 2005). Let *i(T)* denote the impurity at node(*T*), then *i(T)* must be zero when node(*T*) is pure and a maximum when the categories are equally represented. The gini impurity for node *T* is defined as:

$$i(T) = \left\lfloor 1 - \sum_j P^2(c_j) \right\rfloor \tag{8}$$

where, *P (c$_j$)* is the fraction of rows in node *T* with class *c$_j$*.
The reduction in impurity of node *T* is given by:

$$\Delta i(T) = i(T) - P_L\, i(T_L) - P_R\, i(T_R) \tag{9}$$

where, T_L and T_R - Left and Right child node of nodes, $i(T_L)$ and $i(T_R)$ are their impurities, and P_L and P_R are proportion of examples in the child node T_L and T_R. Maximum reduction in impurity is chosen as the split point. The splitting process will continue until no further split is possible and the maximal tree is obtained. The process is stopped when there is only single case in each of the terminal nodes or all cases within each terminal node have the same distribution of predictor variables, making splitting impossible.(ii) Tree pruning: There are several reasons involved, which may lead to overfit the data.When a tree is overfitted, it will lead to inaccuracy in estimating prediction errors, which can be overcome by pruning.The type of pruning differs depending upon the application type, i.e., whether the decision tree is used for classification or for prediction or clustering. The popular pruning techniques include cost-complexity pruning, reduced error pruning, pessimistic error pruning, minimum error pruning and minimum description length, bootstrapping etc. Breiman et al.

(1984) recommended the minimal cost complexity method for pruning the maximal tree. (iii) Optimal tree selection: During which the tree that fits the information in the learning dataset, but does not overfit the information, is selected from among the sequence of pruned trees (Kohavi 1995).

CART works better than discriminant analysis when the variables are uncorrelated. Surrogate variables can be used at a node for missing data cases. It can deal with large datasets of high dimensionality. The CART tree is insensitive to explanatory variable transformations. Outliers are easily handled by CART.

3.Application to clinical trial response classification

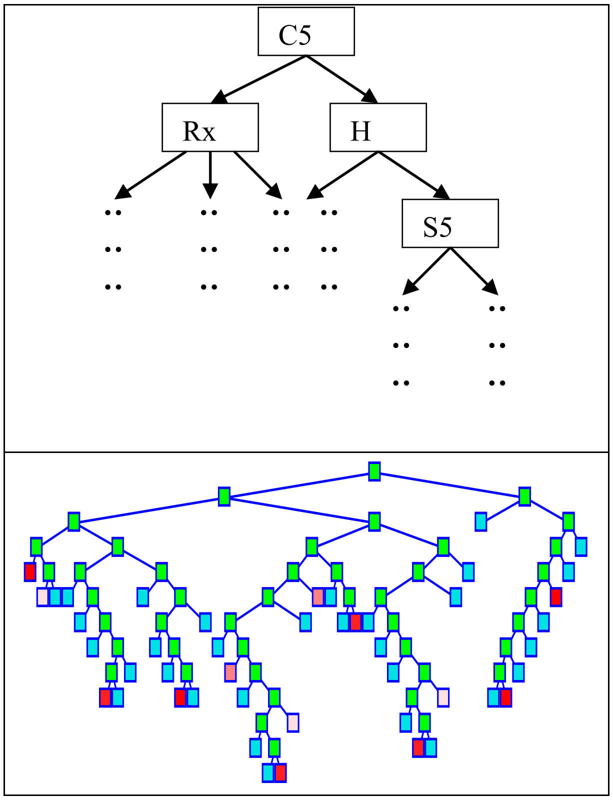
The database consists of 686 cases of pulmonary tuberculosis treated under clinical trial at the Tuberculosis Research Centre, Chennai (Tuberculosis Research Centre 1983). The information on demograph, bacteriological and biochemical investigator at the start of the treatment along with response to treatment at the end of treatment were collected. The sputum culture at fifth month was used as the response variable for fitting CART, C4.5 and ID3 classification. The Waikato Environment for Knowledge Analysis (WEKA) software was used to generate the ID3 and C4.5. The CART 6.0 (Steinberg and Colla 1997) software was used for classification tree. The important attributes used were treatment, sputum culture, sensitivity tests to various anti-TB drugs,sex and percentage of treatment received.Using these attributes, we constructed a decision tree. The description of the cases according to the disease and demograph variables is given in Table 1.

Table 1. *Description of the attribute Information*

Variable	Description
Treatment (R_x)	3 levels (1- R_5, 2- R_7, 3- Z_7)
Inactivation (Int)	2 levels (1- Slow, 2- Rapid)
Sex	2 levels (0-Female, 1- Male)
Sensitivity,Ref (SenR)	2 levels (0 – Sensitivity, 1- Resistance)
Sensitivity,Str (Sen S)	2 levels (0 – Sensitivity, 1- Resistance)
Sensitivity,Iso (Sen H)	2 levels (0 – Sensitivity, 1- Resistance)
Smear (S_5)	2 levels (0 – Neg, 1- Pos)
Culture (C_5)	2 levels (0 – Neg, 1- Pos)
Percentage, R_x received	2 levels (0 – <80%, 1- >80%)
Response at end	2 levels (0 – fav, 1- Unfav)

The unpruned ID3 decision tree is shown in Figure 2.

Fig.2.*Output of Unpruned ID3 decision tree*



It is complicated and it is not easy to understand. ID3 algorithm cannot handle continuous attributes, so we discretized the attributes. The maximal tree overfit the data. In order to avoid over-fitting the data, all methods try to limit the size of the resulting tree. The tree pruning is done by examining the performance of the tree on a holdout dataset. The graphical C4.5 pruned decision tree is shown in Figure 3.

Culture, sensitivity (Iso), smear and treatment received are the important variables in C4.5. It is easier to understand and implement, when a decision tree is converted into rules, which makes it simple. An initial rule is created by considering every path from the root to a leaf by concerning all the test conditions appearing in the path because the conjunctive rule antecedents while concerning the class label held by the leaf as the rule consequence.From Figure 3, rules can be derived from the decision tree, such as (i) If C_5 = positive and Sen H = sensitive and S_5 = negative and $R_x = R_7$, Z_7 then favourable. (ii) If C_5 = positive and Sen H = sensitive and S_5 = negative and $R_x = R_5$ then unfavourable (iii) If C_5 = positive and Sen H = resistance then unfavourable (vi) If C_5 = positive and Sen H = sensitive and S_5 = positive then unfavourable. Figure 4 shows the output of CART decision tree with four leaf nodes based on culture results, treatment and sensitivity tests to various anti-TB drugs. To evaluate the performance of the algorithms we employed accuracy, sensitivity and specificity. Sensitivity, specificity are statistical measures of the performance of a binary classification tests.

Fig.3. *C4.5 pruned tree for Tuberculosis data*

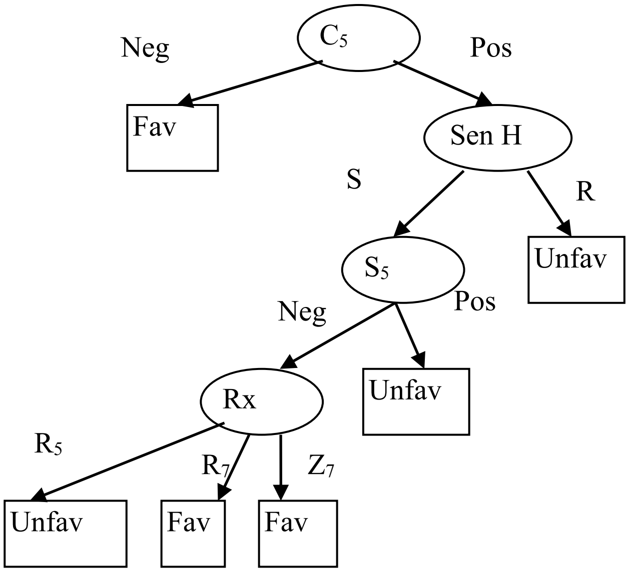


Fig.4. *CART Tree for Tuberculosis data*

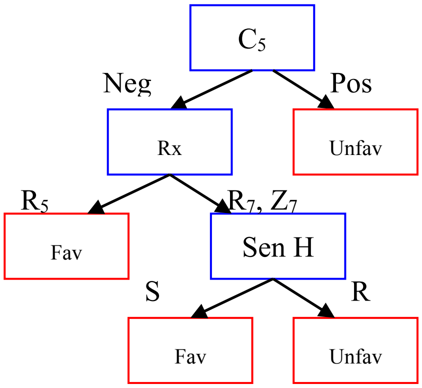


Table 2 shows the summary of results on three decision tree classifiers using selected variables. In ID3, 641 cases were correctly classified (93.4%) and 44 (6.4%) incorrectly classified. C4.5 performed better than ID3 algorithm succeeding correctly classifying 647 cases (94.3%) out of 686, and just 39 (5.6%) were incorrectly classified.C4.5 decision tree has six leaf nodes of size ten with higher accuracy 94.3%, specificity 95.91% and sensitivity 94.19%.The accuracy of CART is lower than ID3 and C4.5 algorithm. Only 90.5 % of cases were correctly classified and the sensitivity is 94% and specificity is 62.3%.

Table2. Comparison of the performances of three methods

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
ID3	93.4	94.27	85.71
C4.5	94.3	94.19	95.91
CART	90.5	94.08	62.33

ID3, CART and C4.5 algorithms create classification rules by constructing a tree-like structure of the data. However, they are differing in splitting criteria and pruning method. From figure 3, we found that the culture, smear and treatment were the strongest predictors. The main difference between CART and the other two methods is that the CART splitting rule allows only binary splits whereas other method allows multiple splits. From Table 2, we can conclude that C4.5 algorithm has highest accuracy compared to other two algorithms because of its simplicity, robustness and effectiveness. The predictive performance of ID3 is slightly lower than the performance of C4.5 algorithm and better than CART algorithm.

4. Conclusions

We have presented the results of three different decision tree algorithms. In data mining, decision tree is the effective classification technique. It is more useful in medical research to construct algorithms for disease classification and prediction. Several published works in the medical field have demonstrated the success of decision tree methods (Mello et al. 2006; Gerald et al. 2002; Das 2010). There are lot of decision tree methods are available for classification. Among all others, C4.5 and CART are popular technique for classification. This work compared the effectiveness of the three popular classification algorithms namely C4.5, ID3 and CART to classify Tuberculosis dataset. C4.5 decision tree has the ability to handle data with missing attribute values better than ID3 decision tree algorithm. It also avoids overfitting the data and reduces error pruning.Experiments and analysis on the tuberculosis database has found some interesting rules.Pramanik et al. (2010) compared ID3, C4.5 and CART algorithm using biomedical heart disease dataset and confirmed that the accuracy of ID3 algorithm is greater than C4.5 algorithm, and CART better than both ID3 and C4.5. However, Anyanwu and Shiva (2009) have reported that the classification accuracy of ID3 is better than that of CART for a large dataset because ID3 has a high accuracy for large data that have been preprocessed and loaded into the memory at the same time. Our results also indicate that the C4.5

classifier performs better in performance of rules generated and accuracy than ID3 and CART. CART produced fewer rules than the other two algorithms. C4.5 had the highest accuracy rate and also it had the highest specificity compared to the other decision tree methods.Although the classification accuracy between C4.5,ID3 and CART are little bit similar, the computational performance differs significantly.

5. Acknowledgements

This research has been supported by a grant from the University Grants Commission. We wish to thank Director, Tuberculosis Research Centre (ICMR), Chennai for permitting us to use randomized clinical trial data.

6. References

1. Anyanwu M N,Shiva SG (2009) Comparative analysis of serial decision treeclassification algorithms. International Journal of Computer Science and Security 3(3): 230-240
2. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and RegressionTrees. Wadsworth International Group, Belmont, CA
3. Das R (2010) A comparison of multiple classification methods for diagnosis ofParkinson disease. Expert Systems with Applications 37: 1568–1572
4. Gerald LB, Tang S, Bruce F, Redden D, Kimerling ME, Brook N, Dunlap N, BaileyWC (2002) A Decision Tree for Tuberculosis Contact Investigation. American Journal of Respiratory and Critical Care Medicine166: 1122-1127
5. Han J, Kamber M (2006) Data Mining: Concepts and techniques. Morgan Kaufmann,2nd edition
6. James KE, White RF, Kraemer HC (2005) Repeated split sample validation to assesslogistic regression and recursive partitioning: an application to the predictionof cognitive impairment. Statistics in Medicine 24(19): 3019-3035
7. Jawahar MS (2004) Current trends in chemotherapy of tuberculosis. Indian Journal ofMedical Research 120: 398-417
8. Kim YS (2010) Performance evaluation for classification methods: comparativesimulation study. Expert Systems with Applications 37(3): 2292-2306
9. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimationand model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence2 (12): 1137–1143(Morgan Kaufmann, San Mateo)
10. Lewis RJ (2000) An introduction to classification and regression tree (CART)analysis. Paper presented at the annual meeting of the Society for Academic Emergency Medicine, San Francisco, CA
11. Li H, Sun, J and Wu, J. (2010) Predicting business failure using classification andregression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. Expert Systems with Applications 37 (8): 5895-5904
12. Mello FCQ, Bastos LGV, Soares SLM, Rezende VM, Conde MB, Chaisson R E,Kritski AL, Netto AR, Werneck GL (2006) Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study.BMC Public Health6(43): 1-8
13. Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision trees: An overview andtheir use in medicine.J Med Syst26(5): 445-463
14. Pramanik S, Md. R Islam, Md. J Uddin (2010) Pattern Extraction, Classification andcomparison between attribute selection measures. International Journal of Computer Science and Information Technologies 1(5): 371-375
15. Quinlan JR (1979) Discovering rules by induction from large collections of examples.Expert Systems in the Micro Electronic Age, Edinburgh University Press, 168–201
16. Quinlan JR (1986) Induction of Decision Trees. Machine Learning 1: 81-106
17. Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan KaufmannPublishers, San Mateo, CA
18. Quinlan JR (1996) Improved use of continuous attributes in C4.5. Journal of ArtificialIntelligence Research 4: 77-90
19. Schlipköter U, Flahault A (2010) Communicable diseases: achievements andchallenges for public health. Public Health Reviews32:90-119
20. Steinberg D, Colla P (1997) CART - Classification and Regression Trees, San Diego,California: Salford Systems, http://www.salfordsystems.com/cart.php.
21. Tuberculosis Research Centre, Madras (1983) Study of chemotherapy regimens of 5and 7 months duration and the role of corticosteroids in the treatment of sputum-positive patients with pulmonary tuberculosis in south India. Tubercle 64: 73-91
22. Ture M, Kurta I, Kurumb AT, Ozdamarc K (2005) Comparing classificationtechniques for predicting essential hypertension. Expert Systems with Applications 29: 583-588
23. Utgoff P, Brodley C (1990) An Incremental Method for Finding Multivariate Splitsfor Decision Trees. Machine Learning: Proceedings of the Seventh International Conference, pp.58
24. Weka reference WEKA: http://www.cs.waikato.ac.nz/~ml/weka/