ISSN 2201-2796

Data Mining with Decision Tree to Evaluate the Pattern on Effectiveness of Treatment for Pulmonary Tuberculosis: A Clustering and Classification Techniques

Babu C Lakshmanan,
Cognizant Technology Solutions
Chennai, India

Valarmathi Srinivasan,
Department of Epidemiology,
The TamilNadu Dr.MGR Medical University,
Chennai, India.

Chinnaiyan Ponnuraja
Department of Statistics,
National Institute for Research in Tuberculosis (ICMR),
Chennai, India.
cponnuraja@gmail.com

Abstract- **Data mining is a process which helps in uncovering interesting data patterns in large volume of data. This procedure has become an ever more activity in all areas of medical science research especially in healthcare circumstances. Data mining has resulted in the innovation of useful hidden patterns from enormous databases. In this paper, a methodology is proposed for the programmed exposure and classification to evaluate the pattern on effectiveness of treatment for Pulmonary Tuberculosis (PTB) patients. Tuberculosis is a disease caused by mycobacterium which spreads through the air and hits low immune bodies easily. Our methodology is based on clustering and classification that classifies the success rate of Tuberculosis treatment based on the two broad classifications of the drug susceptibility testing (DST) namely, sensitivity to all drugs and resistance to any one drug. Age and weight are the main influencing factors for PTB patients, Two Step Clustering(TSC) is used to group data into different clusters and assign classes based on age and weight besides, The same procedure is being compared between with and without clusters of age and weight. Subsequently multiple different classification algorithms are trained on the result set to build the final classifier model based on decision tree along with K-fold cross validation method. The best obtained treatment effectiveness was 97.9% on a specified pattern from Classification and Regression Trees (CART). The proposed approach helps clinicians in their treatment planning procedures for different categories (through decision trees) to discover relationships which are currently hidden in the data.**

***Index Terms*— Data Mining, Decision Tree, CART, CHAID, Clinical Trial**

I. Introduction

Data mining is the technology that recommends the potential means to discover the unidentified knowledge in the large databases. However, since the performance of a data mining technique is dependent on the underlying problem and also it is imperative to analyze their relative performances for any given task. Data mining has been identified as the technology that offers the possibilities of discovering the hidden knowledge from these accumulated databases [1]. The uninterrupted development of more and more complicated classification models through business-related and software packages have turned out to provide various benefits only in detailed problem domains where some prior background knowledge or new evidence can be exploited to further improve classification performance [2]. This script explores the data mining techniques in order to identify the one that will offer the best performance in application to classification of success rate of treatments for TB patients. Classification is one of the data mining tasks that are commonly used to analyze medical data [3]. On the other hand, there is related research that proves no individual data mining technique has been shown to deal well with all kinds of classification problems [4, 5]. Some researchers compared several methods in order to obtain the highest accuracy in diagnosing tuberculosis. CART methodology was developed during 1980s in the paper entitled “Classification and Regression Trees” [6]. Classification and regression trees are becoming increasingly popular for partitioning data and identifying pattern in both small and large datasets. For building decision trees, CART uses the so-called learning sample as a set of historical data with pre-assigned classes for all observations. Classification trees include those models in which the dependent variable (the predicted variable) is categorical. Regression trees include those in which it is continuous.The Decision Tree procedure creates a tree-based classification model. It classifies cases into groups or predicts values of a dependent (target) variable based on the values of independent (predictor) variables. The procedure provides validation tools for exploratory and confirmatory classification analysis. CART is used in clinical trial tuberculosis data with its two splitting rules for impurity measures namely *Gini* Index and *Twoing*. These procedures are being compared between with and without clusters methods to identify the pattern prediction for treatment effectiveness to propose an approach which helps clinicians in their treatment planning management. Additionally, it is however another goal to bring out as well as to discover the hidden information visually (decision trees) in the data. Ultimately and hoping that it is being achieved with some extent.

II. MATERIALS AND METHODS

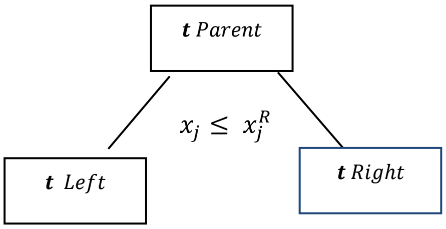
Chi-squared Automatic Interaction Detection (CHAID) and Classification And Regression Trees (CART) are giving different trees. Though they are working for the same purpose, there are a number of differences between these two tree structures. CART gives a better tree than compared to CHAID even in a small sample [7]. CHAID [8] was intended to work with categorical and discretized targets. CHAID uses multi-way splits; it means that the current node is split into more than two nodes by default. The CART manuals [9,10] were provided for understanding CART of a comprehensive background and conceptual basis and discussed it further the art of tree-structured data analysis, provides detailed listings and explanations of CART. CART does binary splits and each node split into two daughter nodes by default and it can definitely do regression and classification also grows a large tree and then post-prunes the tree back to a smaller version. On the other hand this allows CART to perform better than CHAID in and out-of-sample for a given tuning parameter combination. CART handles missing values with surrogate splits; it means that with missing values for predictor variables the algorithm uses predictor variables that are not as good as the primary split variable but mimic the splits produced by the primary splitter. The most important difference is that split variable and split point selection in CHAID is less strongly confounded as in CART. CHAID has no such obsession as like CART. CART algorithm will itself identify the most significant variables and eliminate non-significant ones. CART is a well working prediction machine learning method, so if prediction is the main aim, hence the choice is for prediction using CART.

III. ADVANTAGES OF CART

CART easily handles both continuous and categorical variables and other advantages of CART method is its robustness to outliers. Usually the splitting algorithm will isolate outliers in individual node or nodes. An important practical property of CART is that the structure of its classification or regression trees is invariant with respect to monotone transformations of independent variables. One can replace any variable with its logarithm or square root value, the structure of the tree will not change. CARTs are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. Therefore, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values and it was described in detail [11]. Building tree structure without any biases and unbiased approach on an idea that originated in the CART and uses a two-step approach based on significance tests to split each node without affecting the integrity of inferences [12-17]. CART consists of two splitting rules for impurity measures called *Gini* index, and *Twoing*, distinctive loss functions are the *phi* coefficient for classification trees. The phi coefficient is for a 2 x *k* table formed by the split on *k* categories of the dependent variables. The splitting under *Gini* index is a variance estimate based on all comparisons of possible pairs of values in a subgroup. , The *Twoing* is to describe splitting *k* categories as if it were a 2 category splitting problem [6].

IV. CLASSIFICATION TREES AND SPLITTING RULES

Classification trees are used when for each observation of learning sample, knows the class in advance. Classes in learning sample may be provided by user in accordance with some exogenous rule. Let tp be a parent node and tl, tr- respectively left and right child nodes of parent node tp. Consider the learning sample with variable matrix, which consists of M number of variables xj and Nobservations. Let class vector Yconsist of Nobservations with total amount of Kclasses. Classification tree is built in accordance with splitting rule - the rule that performs the splitting of learning sample into smaller parts.



$$\Delta i(t) = i(t_p) - E[i(t_c)] \tag{1}$$

where t_c is left and right child nodes for the parent node t_p.

Assume P_l, P_r are probabilities of left and right nodes respectively and we will get the

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r) \tag{2}$$

CART solves the maximization problem for each node

$$\underset{x_j \le x_j^{R, j=1,\ldots,M}}{\arg\max} [i(t_p) - P_l i(t_l) - P_r i(t_r)] \tag{3}$$

Gini Index is the most commonly used rule and it uses the following impurity function

$$i(t) = \sum_{k \ne 1} p(k/t) p(l/t) \tag{4}$$

where *k* , *l* are index classes ranges from *1, ...,K,* $p(k/t)$ is conditional probability of class k in node t. In applying the above impurity to maximization problem we will get the following changes of impurity measure

$$\Delta i(t) = -\sum_{k=1}^{K} p^2(k/t_p) + P_l \sum_{k=1}^{K} p^2(k/t_l) + P_r \sum_{k=1}^{K} p^2(k/t_r) \qquad (5)$$

The following equation is obtained using the *Gini* algorithm

$$\underset{x_j \le x_j^R, j=1,\dots\dots,M}{arg\,max} \left[-\sum_{k=1}^{K} p^2(k/t_p) + P_l \sum_{k=1}^{K} p^2(k/t_l) + P_r \sum_{k=1}^{K} p^2(k/t_r) \right] \qquad (6)$$

Gini algorithm will search in learning sample for the largest class also it works well for noisy data.

Twoing splitting rule slightly differs from *Gini*. *Twoing* will search for two classes that will make up together equally [6]. This rule will maximize the following change of impurity measure

$$\Delta i(t) = \frac{P_l P_r}{4} \left[\sum_{k=1}^{K} \left| p(k/t_l) - p(k/t_r) \right| \right]^2 \qquad (7)$$

The above equation implies the following maximization problem

$$\underset{x_j \le x_j^R, j=1,\dots\dots,M}{arg\,max} \left(\frac{P_l P_r}{4} \left[\sum_{k=1}^{K} \left| p(k/t_l) - p(k/t_r) \right| \right]^2 \right) \qquad (8)$$

Decision trees are represented by a set of query which splits the learning sample into smaller and smaller parts. CART algorithm will search for all possible variables and all possible values in order to find the best split and the question that splits the data into two parts with maximum homogeneity. The process is then repeated for each of the resulting data fragments.

V. DATA

Here is in the simple classification tree used for a randomized controlled clinical trial data from National Institute for Research in Tuberculosis (ICMR) for classification of their patients to different levels. The eligible patients were randomly allocated into three different regimens (treatments) including a control treatment as in a revised form of RNTCP (Revised National Tuberculosis Control Programme) treatment with two more trial regimens. All these treatments were administered for six months duration each [18]. All patients were assessed up clinically and bacteriologically every month up to six months. In this application there were 1237 patients (after few more exclusions) with five core variables are included: such as *age* (years) and *weight* (kg) at baseline as in a continuous form, *sex* (male-1, female-0), *drug susceptibility test* (PreRxDSTstatus: sensitive to all drugs-0 and resistant any one drug-1) and *treatment group* (treatment A-1, treatment B-2, Control-0; included as a influencing variable for both CART methods) as in a categorical form, included with an outcome variable of *status* having two levels (sputum culture conversion: converted-1 and not converted-0). The Decision tree comparison is based on with and without clustering of two variables age and weight by the two-step cluster (TSC) method. The corresponding decision trees and their levels of variables are illustrated in Figs.1 and 2. Fig.1 shows that the decision tree is performed by using CART without TSC of age and weight; which is being acknowledged that there is no additional split after the child node of Weight under the initial node from *PreRxDSTstatus;* that means patients who were having initial resistant will obviously having very slow response than compared to the left hand side under sensitivity group of patients who responded well with all drugs.

VI. RESULTS AND DISCUSSION

Figs. 1 and 2 are proving huge differences in their sputum culture conversion rate between patients who had initial drug resistant (at least resistant to any one drug) and patients who sensitivity to all drugs. From Table.1, we can see that the maximum gain of sputum culture conversion rate for node number 4 is 100% which gives the pattern like patients should have sensitivity to all drugs at the beginning and age should be more than 50 years; and the minimum gain is 50% for node 5 under drug resistant group which gives the pattern like patients who had initial drug resistant and underweight may cause very less in their disease conversion. There is no further split under the resistant group and this conveys heterogeneity in the split. Gains are arrived under *Gini Index splitting rules*; ranging from 78.2% to 98.3% for nodes 11 to 16 respectively.

Table 1. Decision Tree without Two Step Clustering Gains for Nodes

Node	Node		Gain		Response (%)
	N	%	N	%	
4	66	5.3	66	6.2	100.0
14	121	9.8	119	11.2	98.3
15	98	7.9	95	8.9	96.9
13	90	7.3	83	7.8	92.2
12	151	12.2	138	13.0	91.4
16	362	29.3	322	30.3	89.0
11	119	9.6	93	8.8	78.2
6	170	13.7	116	10.9	68.2
5	60	4.9	30	2.8	50.0

The highest prediction pattern identified is 100% of node 4 which may not be generalized on its belief, because it does not conveys much information. The next highest predictions are 98.3% and 96.9% of node 14 and 15 respectively under gain index. The patters for 14 and 15 have the next highest conversion rates and they give the most convincing patterns in Fig.1. Though the node 4 has highest prediction pattern, there is no multiway split. Under the resistant arm the maximum prediction is 68.2% of node 6. In fact, it not conveys the homogeneity in their splits. Moreover, the other remarkable gains range from 78.2% to 98.3% for nodes 11 to 16 respectively in Table 1.

Fig. 1. Decision Tree without Two Step Clustering

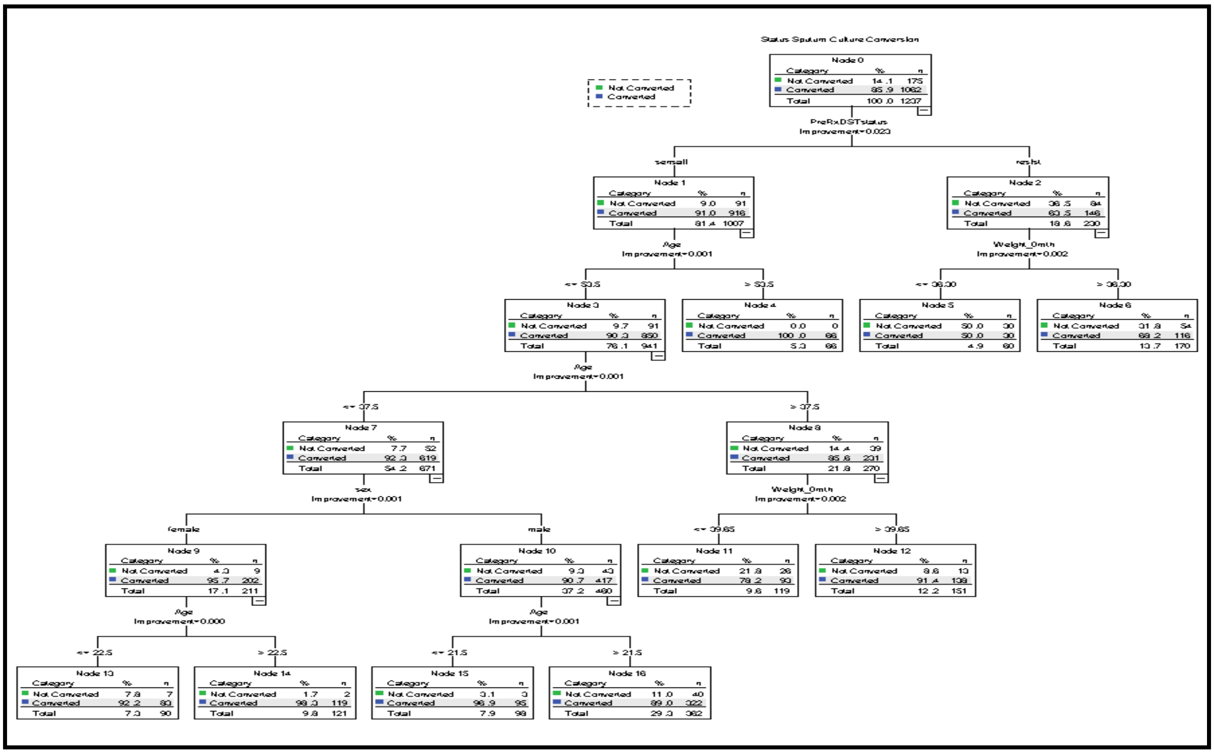
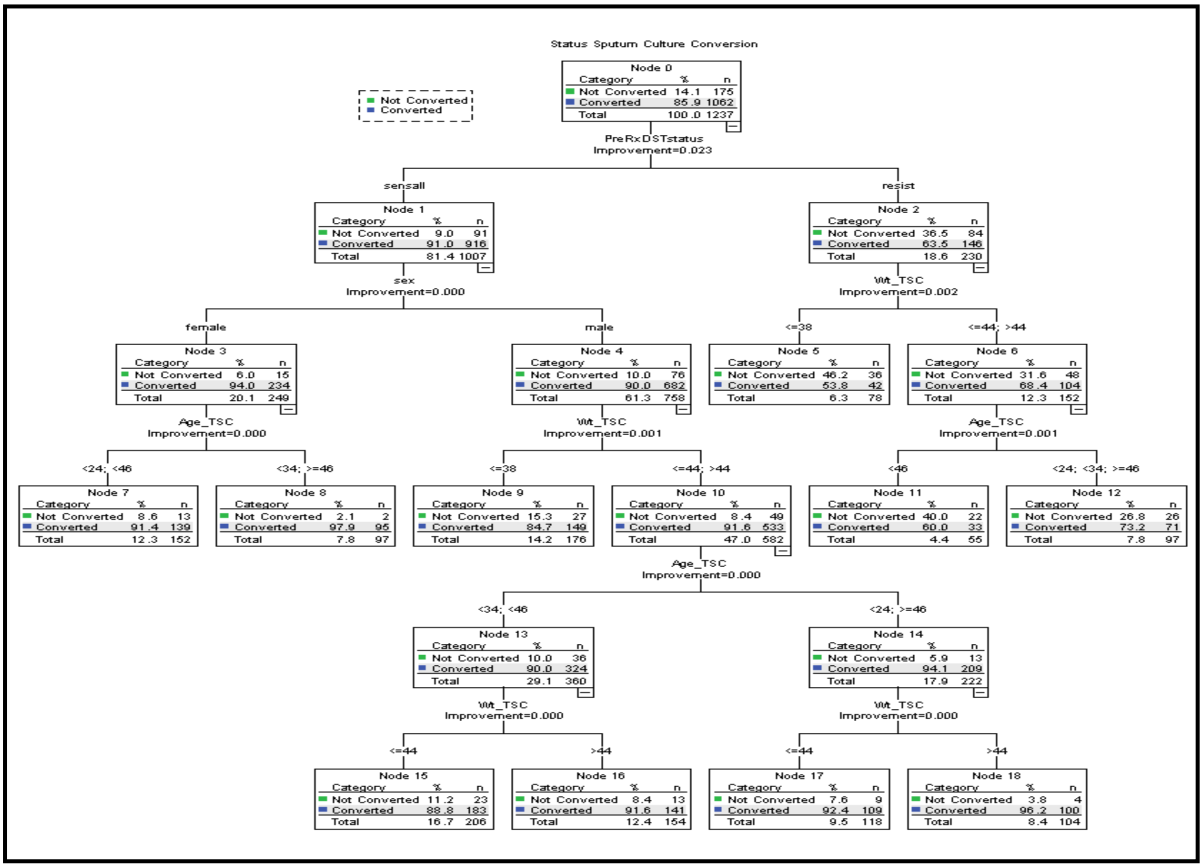


Figure.2, the decision tree performed using CART with clusters based on TSC for age (four clusters: <24, 25-33, 34-45, >=46) and for weight (three clusters: <=38, 39-44, >44). The major difference between Fig.1 and Fig.2 is additional splits under PreRxDSTststus of resistance arm. There is an additional split under the group of resistant patients than compared to the decision tree without clusters based on TSC. The sputum culture conversion rate for node 5 has very less especially for patients who have initial drug resistant (Patient resistant for at least one drug at base line called initial drug resistant). But this is slightly higher than the decision tree without clustering method.

Fig. 2. Decision Tree with Clustered variables using TSC



From Table 2., the maximum sputum culture conversion rate is 97.9% in node 8 under the drug sensitivity group and the minimum gain is 53.8% and also it gets further split with age classification under the PreRxDSTstatus of drug resistant group. Moreover the maximum gain in sputum culture conversion rate is 73.2% for node 12 under the PreRxDSTstatus of drug resistant group. There is a further split after weight classification in drug resistant group (Refer node 11 and 12 in Fig.2) which is statistically significant. The highest prediction pattern is identified for patients with initially sensitive to all drugs, female in gender and age lies between 25 to 33 years. The other pattern, if gender is male then, they should have sensitive organs for all drugs, age should be either less than 24years or greater than or equal to 46years and weight should be larger than 38kg having prediction about 96.2%. These patterns are really conveys some message as well as imitating the reality of the original findings of the report [18].

Table 2. Decision Tree without Two Step Clustering:Gains for Nodes

Node	Node-by-Node						Cumulative				
	Node		Gain		Response	Index	Node		Gain		Response
	N	Percent	N	Percent			N	Percent	N	Percent	
8	**97**	**7.8**	**95**	**8.9**	**97.9**	**114.1**	97	7.8	95	8.9	97.9
18	104	8.4	100	9.4	96.2	112.0	201	16.2	195	18.4	97.0
17	118	9.5	109	10.3	92.4	107.6	319	25.8	304	28.6	95.3
16	154	12.4	141	13.3	91.6	106.6	473	38.2	445	41.9	94.1
7	152	12.3	139	13.1	91.4	106.5	625	50.5	584	55.0	93.4
15	206	16.7	183	17.2	88.8	103.5	831	67.2	767	72.2	92.3
9	176	14.2	149	14.0	84.7	98.6	1007	81.4	916	86.3	91.0
12	97	7.8	71	6.7	73.2	85.3	1104	89.2	987	92.9	89.4
11	55	4.4	33	3.1	60.0	69.9	1159	93.7	1020	96.0	88.0
5	**78**	**6.3**	**42**	**4.0**	**53.8**	**62.7**	1237	100.0	1062	100.0	85.9

VII. SUMMARY

The results patterns arrived under decision tree of CART procedure of with and without TSC are highly important. In fact the results presented in the original paper [18] for culture conversion at end of treatment was reported as 91%, 94% and 89% in Regimen-A(Split1), Regimen-B(Split2) and control regimen respectively. Since the treatment group is identified as an influencing Variable, under the CART procedure it achieves the highest predicted conversion rates (97.9% and 96.2% among female and male respectively) between genders using *Gini Index*. It confirms the objective as well as with the prediction pattern to discover the hidden information visually and numerically. The *Twoing* method is also tried but gives similar results. Thus, it is expected to have more branches for all variables, but the tree which is presented as Figure 1 and 2 are the optimum as well as, are persisting with the homogeneous according to sources presented here. [6] Showed that this method tends to yield trees with too many branches and can also fail to pursue branches which can add significantly to the overall fit. They also advocate, as an alternative, pruning the tree. This procedure is to be considered subsequently in the future work to try with added variables to have more branches. On the other hand, this permits CART to perform better than CHAID in and out-of-sample for a given tuning parameter combination. Since the aim of the paper is prediction as well as identifying pattern for efficacy of treatment, it is further concluded that CART is a glowing effective prediction mechanism as a prediction point of view especially in lesser sample size and ultimately it proves to bring out the hidden information showing in which the course of increasing the accuracy on its prediction aspects.

ACKNOWLEDGEMENTS

The authors acknowledge the director, National Institute for Research in Tuberculosis, Indian Council of Medical Research, Chennai, India for the generous support, of providing facilities and data to complete this work successfully.

REFERENCES

[1] Two Crows Corporation. Introduction to Data mining and Knowledge discovery. 3rd ed. [online]. Available from: http://www.twocrows.com/intro-dm.pdf.

[2] Goebel M, Gruenwald. A survey of data mining and knowledge discovery software tool. ACM SIGKDD Explorations Newsletter. 1999; 1:20-23.

[3] Bakar AA, Febriyani F. Rough Neural Network Model for Tuberculosis Patient Categorization. In Proceedings of the International Conference on Electrical Engineering and Informatics. 2007; 1:765–768.

[4] Caruana R, Mizil AN. Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004; 69 - 78

[5] Han J, Kamber M. Data mining concept and techniques. London: Morgan Kaufmann Publishers: 2001.

[6] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.

[7] Hastie Trevor, Tibshirani Robert, Friedman Jerome. The element of statistical learning Data Mining, Inference, and Prediction, Second Edition, Springer: 2013.

[8] Kass GV. An exploratory technique for investigating large quantities of categorical data. Appl Stat. 1980; 29:119–127.

[9] Steinberg D, Colla P. CART: Tree-structured non-parametric data analysis. San Diego, Calif., U.S.A.: Salford Systems. 1995.

[10] Steinberg D, Colla P, Martin K. CART—Classification and regression trees: Supplementary manual for Windows. San Diego, Calif., U.S.A.: Salford Systems.1998.

[11] Loh WY. Classification and regression trees. WIREs Data Mining Knowl Discov. 2011; 1:14–23.

[12] Loh WY, Shih Y. Split selection methods for classification trees. Stat Sin. 1997; 7:815–840.

[13] Kim H, Loh WY. Classification trees with unbiased multiway splits. J Am Stat Assoc. 2001: 96:589– 604.

[14] Kim H, Loh WY. Classification trees with bivariate linear discriminant node models. J Comput Graphical Stat.2003; 12:512–530.

[15] Hothorn T, Hornik K, Zeileis A.Unbiased recursive partitioning: a conditional inference framework. J Comput Graphical Stat.2006; 15:651–674.

[16] Loh WY. Improving the precision of classification trees. Ann Appl Stat. 2009; 3:1710–1737.

[17] Yohannes, Y, Webb P. Classification and regression trees: A user manual for identifying indicators of vulnerability to famine and chronic food insecurity. International Food Policy Research Institute, Washington, D.C. Mimeo: 1998.

[18] Tuberculosis Research Centre (Indian Council of Medical Research), Chennai, India. Split-drug regimens for the treatment of patients with sputum smear-positive pulmonary tuberculosis- a unique approach, Tropical Medicine and International Health. 2004; 9: 551-558.