Statistics and Applications
Volume 12, Nos. 1&2, 2014 (New Series), pp. 21-33

Decision Tree Based Models for Classification in Agricultural Ergonomics

Sandip Kumar Sadhu[1], Ramasubramanian V.[2], Anil Rai[3] and Adarsh Kumar[4]

[1] *National Academy of Statistical Administration, MoSPI, New Delhi*
[2] *Central Institute of Fisheries Education, Mumbai*
[3] *Indian Agricultural Statistics Research Institute, New Delhi*
[4] *Division of Agricultural Engineering, Indian Agricultural Research Institute, New Delhi*

Abstract

Classification and prediction in agricultural systems are quite useful for planning purposes. In this study, tree based modeling has been employed for classification purposes in the field of agricultural ergonomics by using data on presence or absence of discomfort for the farm labourers in operating farm machineries along with associated quantitative and qualitative predictor variables. As the tree based classification method works better for larger datasets, such a dataset was generated by simulation procedure from the available dataset. To start with, the conventional Classification and Regression Tree (CART) models have been fitted on the different possible predictor variable subsets of this larger dataset. The fitted models have been compared by cross-validation error measures and misclassification rates. All these models were employed for classification of hold-out observations not used in models fitting. In addition, an alternative tree-based modeling procedure for classification has been proposed within the CART framework to overcome some of its inherent drawbacks. In this, the variable selection and split point selection has been done by separately employing sound statistical testing and optimal stratification procedures respectively to circumvent the selection bias which arise in the usual CART methodology due to simultaneous selection of the split variable and the split points. Thereafter, an independent test set has been used for pruning of the resulting maximal tree. The final fitted tree has been validated upon the same hold-out dataset employed earlier for classification using the conventional CART models fitted. Moreover, classifications using the fitted CART models were found to be better when compared to those obtained for corresponding logistic regression and discriminant function analysis methods. When the classificatory performances in terms of accuracy of the CART models fitted using both the conventional and proposed approaches are compared, the results came out to be at par under the two approaches. Also the proposed CART method has advantage over the conventional method in terms of computational speed and flexibility while handling large data sets.

Keywords: Classification subtree; Gini diversity index; Pruning, Simulation; Split point selection

1 Introduction

Classification and Regression Trees (CART) was introduced by Breiman *et al.* (1984). They produce a classification or prediction rule from a set of observations described in terms of a vector of a set of explanatory and a response variable. The CART methodology have found favour among researchers for classificatory application in several areas such as agriculture, medicine, etc. as alternatives to the conventional approaches such as discriminant function method, logistic regression etc. Tree based methods are not based on any stringent assumptions. These methods can handle large number of variables, are resistant to outliers, non-parametric, more versatile, can handle qualitative variables, though computationally more intensive.

Breiman *et al.*(1984) developed CART methodology which is a sophisticated program for fitting trees to data. Loh and Shih (1997) developed the QUEST (Quick Unbiased Efficient Statistical Tree) method to take care of the selection bias towards the variables with more possible splits. Shih (2001) worked out the variable selection bias towards the variables providing more number of splits in the exhaustive search algorithm in two-class case with numerical predictors and suggested a solution based on p-values. Cappeli *et al.* (2002) suggested the use of statistical significance in the pruning procedure of both classification and regression trees to obtain a statistically reliable tree. Izenman (2008) has given a detailed account on CART methodology in his book.

The rest of the paper is organized in the following manner. Section 2 gives some of the drawbacks that exist in the CART methodology. Section 3 gives the description about the data utilised and simulation procedure for getting larger dataset. Section 4 is devoted to the detailed discussion about CART methodology. It gives details about the splitting and pruning strategies. Section 5 discusses some of the drawbacks of the existing CART procedure and in view of these shortcomings, attempts to overcome them. A new method of growing a classification tree has been described in this paper. Section 6 consists of the results of the fitted tree based classifiers and the conventional classification methods upon the simulated dataset by taking different subsets of variables and comparison of different methods of classification including the proposed method. The comparisons have been made between the tree-based methods with that of some conventional methods.

2 Limitations of Conventional CART Method

The exhaustive search (and more popular) approach of CART method examines all possible binary splits of the data based on each predictor variable to select the split that produces maximal reduction in some measure of what is called ‘node impurity’(which will be discussed in detail in subsequent sections) . There are two problems with exhaustive search approach:

a) Computational complexity: A quantitative variable with n distinct values at a node induces (n-1) splits. Therefore the order of computations at each node is linear in the number of distinct data values. In case of a qualitative variable, the order of computations increases exponentially with the number of categories, the possible number of splits being (2^{M-1}-1) for a variable with M categories.

b) Bias in variable selection: A more serious problem from the standpoint of tree interpretation is that unrestrained search tends to the selection of variables that have more possible splits. This makes it hard to draw reliable conclusions from the tree structures.
The above limitations of the exhaustive search algorithms leave much scope to find some alternative tree growing strategies to improve the existing methodology.

3 Data Description and Simulation of Larger Dataset

The data for the present study has been taken in the area of agricultural ergonomics obtained from Division of Agricultural Engineering, IARI, New Delhi collected during 2007-08. The variable considered as the dependent variable (Y) is dichotomous i.e. ‘presence’ or ‘absence’ of discomfort for the farm labourers during agricultural field operation. The set of qualitative explanatory variables are: modes of operation (X_1) and percent aerobic capacity of the farm labourers (X_2), each having two levels. The variable X_2 has two levels labelled as low and high viz., less than or equal to 35% and greater than 35% of aerobic capacity of the farm labourers respectively. The dataset consists of 405 observations available for the study. In this dataset, broadly two levels of modes of operation viz. predominantly foot operated (e.g. Bicycle, Stepper, Pedal etc.) and other mode of operation (e.g. Flywheel, Rocking etc.) are considered. The quantitative explanatory variables are: load given to farm machinery (X_3), difference between working and resting heart rates (X_4) and oxygen consumption at the time of farm operation (X_5).
The tree based classification method works better for larger datasets. For this reason, the 405 data points were blown up to 10000 data points through simulation assuming multivariate normal setup for the quantitative variables preserving the structure of qualitative variables in the original set up. To generate the qualitative variables, the different combinations of the levels of these qualitative variables are found out and a probability rule is worked out to generate them with same number of possible combinations. These probabilities were nothing but the proportions of these combinations in the original dataset. Since each of the two qualitative variables have two levels (denoted by 0 and 1), they can have four possible structural relations, viz., (0,0), (0,1), (1,0) and (1,1), where the first digit denotes the value taken by X_1 and the second is the value taken by X_2. These four combinations occur for each of the two categories of the dependent variable. Hence a total of eight combinations of the qualitative variables exist. In this way, while simulating the larger dataset, the number of cases having these combinations was kept proportional to their numbers in the original dataset.

Regarding the simulation of quantitative variables, for each of the eight sets of combinations of the qualitative variables, a set of observations on each of the three quantitative variables can be obtained separately from the available dataset with the corresponding mean vectors and covariance matrices. For generating such data points, use of what is called the Cholesky’s algorithm given in Scheuer and Stroller (1962) has been employed. To retain the original structure for the quantitative variables, we take only those values lying between the maximum and minimum values of the variable under consideration in the real dataset.

Generated datasets for each of the eight combinations of levels of the qualitative variables are then merged together to obtain the final dataset containing 10000 data points to work with. Out of these 10000 data points, 8000 data points has been randomly selected for building the tree based model and henceforth referred to as learning dataset (L). The remaining dataset of 2000 data points is set aside for the purpose of evaluating the efficiency of the classification rule, i.e. how good is the classification method in classifying new observations that has not been used in building it. This dataset will be called the “hold out” dataset to make a distinction from what is called a ‘test set’ which will be discussed elsewhere.

4 Conventional CART Methodology

The CART procedure for growing a decision tree is known as binary recursive partitioning. The CART split criterion is to split the node using the predictor variable and split point which results in the highest reduction in impurity measured by Gini diversity index in the present case among many such measures. Let, Π_1 and Π_2 be the two classes. For node τ, the node impurity function is defined as $i(\tau)=\emptyset\left(p(1|\tau), p(2|\tau)\right)$ where $p(k|t)$ (with k = 1 or 2) is an estimate of $P(\mathbf{X}\in\Pi_k|\tau)$. The Gini diversity index for the two class case is given by $i(\tau)=2p\left(1\text{-}p\right)$ where, $p=P(\mathbf{X}\in\Pi_1|\tau)$.

A classification tree with the best possible classification performance is called a right-sized tree. Breiman *et al.* (1984) proposed that the tree be allowed to be grown to its maximal size first and then gradually shrunk by pruning away branches that lead to the smallest decrease in accuracy compared to pruning other branches. The CART method uses minimal cost complexity for pruning the maximal tree in order to get a series of subtrees. The RPART package of the freely available R software has been used for doing this task with a standard ten-fold cross-validation technique and thus the conventional CART trees have been obtained considering various subsets of the original dataset.

5 Proposed Modifications in Conventional CART Methodology

As has already been discussed in Section 2, simultaneous selection of the split variable as well as the split points in the CART methodology introduces some selection bias for the variables with greater number of possible splits. To reduce this bias, a better way is to separate these two procedures by adopting different measures for the two selections. In the present study, the QUEST algorithm given by Loh and Shih (1997) has been adopted for variable selection. It is based on the testing of significance for each variable and selecting the variable giving minimum probability value (*p*) for the test statistic. As regards to the selection of split point, the stratification methodology in the field of sampling theory as proposed by Dalenius and Hodges (1957) has been followed, since this method can be used to split (divide) the observations in the parent node into two more pure (homogeneous) daughter nodes. In this way, a fully grown tree classifier can be obtained. At the final step, the pruning of the fully expanded classifier can be done based on a test set error estimate as is usually done in the conventional CART procedure.

The following general set up and notations will be used subsequently to explain the proposed method:

5.1 Variable Selection

For each split, the association between each explanatory variable and the dependent variable is computed using the ANOVA F-test (for quantitative predictors) or Pearson’s contingency table χ^2-test of independence (for qualitative predictors). The explanatory variable having the highest association with the dependent variable is selected for splitting. The variable with the smallest p-value is selected at each node whose split point is determined by the procedure discussed subsequently.

5.2 Split Point Selection

One major aim of splitting a parent node into two daughter nodes is to make each of the set of observations in the daughter nodes more homogeneous with respect to the class of the observations, i.e. the class the observations belong to. It can be looked upon as stratifying the parent node observations into two strata which are more homogeneous within themselves but more heterogeneous between them. Dalenius and Hodges (1957) proposed such a procedure of construction of stratification so that the sampling variance of the estimates for population parameters is minimized. The number of strata or groups to be made for the present study is two (since the split is binary).

If the variable under study (the split variable) X is quantitative, its class frequency distribution is tabulated. The point of division forming two strata is found in such a way that the variance of the stratified mean estimate thus obtained is the least. Since for the present study each of the qualitative variables has only two values, the split point can be obtained directly without applying Dalenius method. However, if a qualitative variable has more than two categories, then the split point can be obtained through this method by forming a discrete frequency distribution rather than quantitative frequency distribution, the subsequent steps being same for the qualitative variables.

5.3 Pruning and Selection of Right Sized Sub-tree from Candidate Sub-trees

The fully grown tree has been pruned using the test sample approach. One third of the learning data set (8000 data points) has been used as test data set (approximately 2700 data points). To create a sequence of different-sized sub-trees, the weakest-link cutting (pruning) method has been followed, wherein all of the nodes that arise from a specific non-terminal node are pruned off (leaving that specific node itself as terminal node), and the specific node selected is the one for which the corresponding pruned nodes provide the smallest per node decrease in the resubstitution misclassification rate. The selection of the best tree amongst the entire candidate subtrees is based on estimated misclassification rates, obtained using the test sample. The independent test set is used to estimate the error rates of the various trees in the nested sequence of subtrees, and the tree with minimum estimated misclassification rate is selected to be used as the tree-structured classifier.

6 Comparison with Traditional Statistical Methods

In order to compare the efficiency of the tree based methods, both conventional and proposed CART methods, the traditional classification methods viz., logistic regression and discriminant function analysis methods have been employed on the same set of data. In the present study, for different combination of five explanatory variables, (2^5-1) =31 possible logistic regression models have been fitted for comparison. For the discriminant function, only the quantitative variables X_3, X_4 and X_5 are considered. All possible combinations of these three variables have been used to obtain the classification rules. Thus, for the comparison of discriminant analysis with other methods, out of 31 models mentioned previously, only 2^3-1=7 models will be comparable.

In the present study, the classifying ability has been measured with the help of (2 x 2) classification tables. The columns are the two predicted classes of the dependent variable, while the rows are the two observed classes of the dependent variable. In a perfect model, all cases will be on the diagonal and the overall correct percentage will be 100%. Critical terms associated with classification table are as follows: (Higher the CCR, sensitivity and specificity, lower the false positive rate and false negative rate, better the classification ability)

1) *Correct classification rate (CCR):* Number of correct predictions divided by sample size. The correct classification rate for the model should be compared to the correct classification rate for the classification table for the constant-only model.

2) *Sensitivity:* Percent of correct predictions in the reference category (usually 1 i.e. “present”) of the dependent. It also refers to the ability of the model to classify an event correctly.

3) *Specificity:* Percent of correct predictions in the given category (usually 0 i.e. “absent”) of the dependent variable. It also refers to ability of the model to classify a non event correctly.

4) *False positive rate (FPR):* It is the proportion of predicted event responses that were observed as nonevents.

5) *False negative rate (FNR):* It is the proportion of predicted nonevent responses that were observed as events.

7 Results and Discussion

For the simulation study, the probabilities with which each of the combinations of the qualitative variable occur in the real dataset, has been given in Table 7.1. The first three columns give the values of the qualitative variables and the last column gives the probability (as proportion of such cases to the total dataset) with which they are present in the real dataset. A ‘0’ value of Y represents the absence of discomfort while that of ‘1’ represents the presence of the same. The ‘0’ and ‘1’ for X_1 represents foot operated and hand driven modes of operation respectively, while the same for X_2 denotes low and high percentage values respectively. For the combination (1, 1, 0), the probability comes out to be zero since no observations for this combination was found in the real data set.

Table 7.1. Probabilities of occurrence of different combinations of values of qualitative variables

Y	X_1	X_2	Probability
0	0	0	0.123
0	0	1	0.173
0	1	0	0.030
0	1	1	0.160
1	0	0	0.005
1	0	1	0.254
1	1	0	0.000
1	1	1	0.254
Total			1.000

Regarding the quantitative variables, the mean vector and the covariance matrix of X_3, X_4 and X_5 for the real dataset ($\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$ respectively) and the simulated dataset ($\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$ respectively) are as follows,

$$\boldsymbol{\mu}_r' = [1.7 \quad 51.68 \quad 1.04]\ ; \qquad \boldsymbol{\mu}_s' = [1.64 \quad 51.42 \quad 1.14]$$

$$\boldsymbol{\Sigma}_r = \begin{bmatrix} 1.57 & 19.61 & 0.30 \\ 19.61 & 383.25 & 4.99 \\ 0.30 & 4.99 & 0.26 \end{bmatrix}\ ; \quad \boldsymbol{\Sigma}_s = \begin{bmatrix} 1.08 & 14.77 & 0.29 \\ 14.77 & 327.20 & 4.71 \\ 0.29 & 4.71 & 0.22 \end{bmatrix}$$

These depicts that the simulated dataset preserves the properties of the original dataset with the values simulated restricted to lie between the range of those values corresponding to the variable concerned in the original dataset.

7.1 Results of CART Procedure

With the five explanatory variables, conventional CART models have been fitted with each possible subset of the explanatory variables to see the effect of different variables on tree building procedure and the accuracy of classification. Thus, there are 2^5-1=31 CART models in all. For each of these models, different properties are given in Table 7.2. Each row of the table represents a tree built with the subset variables given in second column.

From the results given in Table 7.2, the effect of number of variables as well as that of different variables can be worked out. Firstly considering the number of splits (equivalently, the number of terminal nodes) it can be seen that whenever there is a quantitative variable present, in most cases, the tree has more number of splits than in the absence of the quantitative variable. When three or more variables are used to build the tree, the number of splits is higher in case all the variables used are quantitative as compared to those built with a mix of quantitative and qualitative variables. From these it is clear that conventional CART favours the quantitative variables for splitting the nodes as compared to the qualitative variables since the quantitative variables can afford more number of split. From the columns of the resubstitution error estimate and 10-fold cross validation errors ($R^{CV/10}$) of the trees, it is evident that more the number of continuous variables involved tree building more is its classification accuracy. In this respect the trees M29 and M31 performs almost equivalently. But since the SE of $R^{CV/10}$ for the tree M31 is less than that of M29, hence it is the best performing among all possible 31 CART models.

The different accuracy measures of the 31 CART models in terms of the classification of the 2000 new observations from the hold-out data set are displayed in Table 7.3. All the above quantities can be calculated from the 2×2 classification tables obtained for each of the 31 CART models.

Table 7.2. Properties of different trees built with the subset variables

Model no.	Variables used	Complexity Parameter (α)	Number of splits	Resubstitution error estimate	$R^{CV/10}$	$SE(R^{CV/10})$
M1	x_1	0.0000000	1	0.9397	0.9397	0.0116
M2	x_2	0.0000000	1	0.6942	0.6942	0.0110
M3	x_3	0.0004000	1	0.1072	0.1096	0.0035
M4	x_4	0.0001000	1	0.1592	0.1613	0.0411
M5	x_5	0.0005000	2	0.1576	0.1584	0.0041
M6	x_1,x_2	0.0000000	1	0.6942	0.6942	0.0110
M7	x_1,x_3	0.0005636	1	0.1737	0.1792	0.0065
M8	x_1,x_4	0.0001000	12	0.2055	0.2099	0.0070
M9	x_1,x_5	0.0010403	13	0.2255	0.2429	0.0075
M10	x_2,x_3	0.0005202	11	0.1654	0.1828	0.0066
M11	x_2,x_4	0.0000000	1	0.3046	0.3134	0.0083
M12	x_2,x_5	0.0006936	1	0.2531	0.2546	0.0076
M13	x_3,x_4	0.0009103	13	0.1373	0.1493	0.0060
M14	x_3,x_5	0.0005202	19	0.0572	0.0674	0.0041
M15	x_4,x_5	0.0005852	21	0.1360	0.1535	0.0061
M16	x_1,x_2,x_3	0.0004942	1	0.1737	0.1808	0.0066
M17	x_1,x_2,x_4	0.0000000	15	0.2000	0.2068	0.0070
M18	x_1,x_2,x_5	0.0006502	15	0.2239	0.2398	0.0074
M19	x_1,x_3,x_4	0.0010403	14	0.1064	0.1204	0.0054
M20	x_1,x_3,x_5	0.0005202	19	0.0572	0.0674	0.0041
M21	x_1,x_4,x_5	0.0003901	36	0.1066	0.1319	0.0057
M22	x_2,x_3,x_4	0.0009103	16	0.1319	0.1516	0.0060
M23	x_2,x_3,x_5	0.0005202	19	0.0572	0.0674	0.0041
M24	x_2,x_4,x_5	0.0005852	21	0.1360	0.1535	0.0061
M25	x_3,x_4,x_5	0.0005202	18	0.0554	0.0658	0.0041
M26	x_1,x_2,x_3,x_4	0.0006069	26	0.0926	0.1087	0.0052
M27	x_1,x_2,x_3,x_5	0.0005202	19	0.0572	0.0674	0.0041
M28	x_1,x_2,x_4,x_5	0.0005852	21	0.1360	0.1545	0.0061
M29	x_1,x_3,x_4,x_5	0.0005202	23	0.0477	0.0622	0.0041
M30	x_2,x_3,x_4,x_5	0.0005202	18	0.0554	0.0661	0.0041
M31	x_1,x_2,x_3,x_4,x_5	0.0005202	23	0.0476	0.0622	0.0040

Table 7.3. Different accuracy measures for the 31 CART models

Model no.	Variables used	CCR	Sensitivity	Specificity	FPR	FNR
M1	x_1	0.4960	0.4293	0.5605	0.5144	0.5707
M2	x_2	0.4905	0.8433	0.1495	0.5106	0.1567
M3	x_3	0.5015	0.4431	0.5633	0.2120	0.2865
M4	x_4	0.5150	0.5014	0.5294	0.2285	0.2565
M5	x_5	0.5130	0.4752	0.5530	0.2170	0.2700
M6	x_1,x_2	0.6440	0.9888	0.3107	0.4190	0.0112
M7	x_1,x_3	0.9075	0.8413	0.9715	0.0339	0.1587
M8	x_1,x_4	0.8945	0.8891	0.8997	0.1045	0.1109
M9	x_1,x_5	0.8830	0.8260	0.9381	0.0720	0.1740
M10	x_2,x_3	0.9110	0.8444	0.9754	0.0292	0.1556
M11	x_2,x_4	0.8575	0.8586	0.8564	0.1475	0.1414
M12	x_2,x_5	0.8710	0.8301	0.9105	0.1003	0.1699
M13	x_3,x_4	0.9215	0.8800	0.9617	0.0431	0.1200
M14	x_3,x_5	0.9595	0.9410	0.9774	0.0243	0.0590
M15	x_4,x_5	0.9305	0.9166	0.9440	0.0595	0.0834
M16	x_1,x_2,x_3	0.9050	0.8566	0.9518	0.0550	0.1434
M17	x_1,x_2,x_4	0.8985	0.8993	0.8977	0.1053	0.1007
M18	x_1,x_2,x_5	0.8830	0.8271	0.9371	0.0730	0.1729
M19	x_1,x_3,x_4	0.9400	0.9064	0.9725	0.0305	0.0936
M20	x_1,x_3,x_5	0.9595	0.9410	0.9774	0.0243	0.0590
M21	x_1,x_4,x_5	0.9405	0.9379	0.9430	0.0592	0.0621
M22	x_2,x_3,x_4	0.9245	0.8881	0.9597	0.0449	0.1119
M23	x_2,x_3,x_5	0.9595	0.9410	0.9774	0.0243	0.0590
M24	x_2,x_4,x_5	0.9305	0.9166	0.9440	0.0595	0.0834
M25	x_3,x_4,x_5	0.9660	0.9451	0.9862	0.0148	0.0549
M26	x_1,x_2,x_3,x_4	0.9420	0.9247	0.9587	0.0442	0.0753
M27	x_1,x_2,x_3,x_5	0.9595	0.9410	0.9774	0.0243	0.0590
M28	x_1,x_2,x_4,x_5	0.9305	0.9166	0.9440	0.0595	0.0834
M29	x_1,x_3,x_4,x_5	0.9725	0.9573	0.9872	0.0136	0.0427
M30	x_2,x_3,x_4,x_5	0.9660	0.9451	0.9862	0.0148	0.0549
M31	x_1,x_2,x_3,x_4,x_5	0.9725	0.9573	0.9872	0.0136	0.0427

The efficiency of different tree based methods can be judged from Table 7.3. From the table, it is obvious that the tree built with X_1, X_3, X_4 and X_5 has the maximum correct classification rate of about 97.25%, which is the same with that of the tree built from all the five explanatory variables. So, inclusion of X_2 in the model does not improve the classificatory ability much. But the latter has a lower resubstitution error estimate of the misclassification rate than the former. So, although the model M29 performs as good as the M31, but the tree built with all the five explanatory variables, i.e. M31 is the best performing if all the aspects of efficiency of the models are considered.

7.2 Results of Logistic Regression Models

In all, 31 logistic regression models were fitted using all the combination of the five explanatory variables and their classification accuracies are given in Table 7.4.

Table 7.4. Classification accuracies of different Logistic regression models

Model no.	Variables used	CCR	Sensitivity	Specificity	FPR	FNR
M1	x_1	0.4385	0.4700	0.4081	0.3010	0.2605
M2	x_2	0.3560	0.0112	0.6893	0.1580	0.4860
M3	x_3	0.9120	0.8942	0.9292	0.0360	0.0520
M4	x_4	0.8575	0.8586	0.8564	0.0730	0.0695
M5	x_5	0.8735	0.8688	0.8781	0.0620	0.0645
M6	x_1,x_2	0.3560	0.0112	0.6893	0.1580	0.4860
M7	x_1,x_3	0.9075	0.8718	0.9420	0.0295	0.0630
M8	x_1,x_4	0.7465	0.6948	0.7965	0.1035	0.1500
M9	x_1,x_5	0.8745	0.8698	0.8791	0.0615	0.0640
M10	x_2,x_3	0.9080	0.8454	0.9685	0.0160	0.0760
M11	x_2,x_4	0.8540	0.7986	0.9076	0.0470	0.0990
M12	x_2,x_5	0.8095	0.9268	0.6962	0.1545	0.0360
M13	x_3,x_4	0.9235	0.9135	0.9331	0.0340	0.0425
M14	x_3,x_5	0.9505	0.9410	0.9597	0.0205	0.0290
M15	x_4,x_5	0.9190	0.9095	0.9282	0.0365	0.0445
M16	x_1,x_2,x_3	0.9000	0.8067	0.9902	0.0050	0.0950
M17	x_1,x_2,x_4	0.7465	0.6897	0.8014	0.1010	0.1525
M18	x_1,x_2,x_5	0.8095	0.9268	0.6962	0.1545	0.0360
M19	x_1,x_3,x_4	0.8660	0.8128	0.9174	0.0420	0.0920
M20	x_1,x_3,x_5	0.9505	0.9247	0.9754	0.0125	0.0370
M21	x_1,x_4,x_5	0.8480	0.8321	0.8633	0.0695	0.0825
M22	x_2,x_3,x_4	0.9225	0.9461	0.8997	0.0510	0.0265
M23	x_2,x_3,x_5	0.8735	0.9797	0.7709	0.1165	0.0100
M24	x_2,x_4,x_5	0.8050	0.9746	0.6411	0.1825	0.0125
M25	x_3,x_4,x_5	0.9550	0.9451	0.9646	0.0180	0.0270
M26	x_1,x_2,x_3,x_4	0.8660	0.8576	0.8741	0.0640	0.0700
M27	x_1,x_2,x_3,x_5	0.8820	0.9705	0.7965	0.1035	0.0145
M28	x_1,x_2,x_4,x_5	0.8095	0.9410	0.6824	0.1615	0.0290
M29	x_1,x_3,x_4,x_5	0.9215	0.8942	0.9479	0.0265	0.0520
M30	x_2,x_3,x_4,x_5	0.8415	0.9827	0.7050	0.1500	0.0085
M31	x_1,x_2,x_3,x_4,x_5	0.8620	0.9746	0.7532	0.1255	0.0125

7.3 Results from Discriminant Analysis

Seven sub-sets could be obtained using the three quantitative variables and were used for classification purposes. An observation having the largest discriminant score for a class is assigned to that class. The different correct classification rates for the different linear discriminant analysis methods have been reported in Table 7.5.

Table 7.5. Correct classification rates of discriminant analysis using subset explanatory variables

Model No.	Variables used	Correct classification rate
M3	X_3	0.914
M4	X_4	0.857
M5	X_5	0.879
M13	X_3,X_4	0.918
M14	X_3,X_5	0.949
M15	X_4,X_5	0.911
M25	X_3,X_4,X_5	0.952

7.4 Result of Proposed CART Method

For the proposed modification in the CART model, the same learning and test sets have been employed. The proposed method has been employed using all the five explanatory variables since the performance of the CART was found to be best when all the explanatory were considered for building the tree classifier.

Using the training dataset, the fully expanded tree had 52 terminal nodes. At the root node the variable X_3 has been found to have maximum association with the dependent variable. Hence the root node was split based on X_3. The recursive partitioning gives a fully grown tree with 52 terminal nodes out of which 22 are labeled as “present” i.e. presence of discomfort, while 30 are labeled as “absent” i.e. absence of discomfort. The labeling of the terminal nodes has been done on plurality basis. While pruning the tree, ten subtrees were obtained by the weakest–link cutting method. From the test-sample error estimates, the final tree chosen consists of 38 terminal nodes. It has been observed that, pruning beyond 38 terminal nodes, although reduces the complexity of the tree (number of terminal nodes), but did not contribute much to the reduction in the misclassification error.

After obtaining the final tree classifier having 38 terminal nodes, the efficiency of the classifier was judged by an independent hold-out dataset of size 2000. Each of the observations in this hold out data set was dropped down the tree to reach at one of the 38 terminal nodes. Since for each of the 2000 observations in the hold-out dataset, the class is known, the accuracy of classification was evaluated by observing how many of the observations have been correctly classified. Table 7.6 gives the results of the classifications made by the proposed CART method on the 2000 hold-out dataset. The different measures of classification accuracies as mentioned earlier are given in Table 7.7.

Table 7.6. 2×2 classification table of the proposed CART method

		Predicted		
		Absent	**Present**	**Total**
Actual	**Absent**	1012	05	1017
	Present	72	911	983
	Total	1084	916	2000

Table 7.7. Different classification accuracy measures of the proposed CART method

Measures	**Values**
Correct classification rate	0.9615
Sensitivity	0.9268
Specificity	0.9951
False positive rate	0.0055
False negative rate	0.0732

8 Conclusions

After having obtained all the required measures of classification accuracies, it is clear that tree based methods of classification outperforms that of the conventional methods. Firstly, when the results of the discriminant analysis is taken into account, for the seven variable subsets used in the discriminant analysis, the classification accuracies of the conventional CART methodology is either at par or more than the accuracy of the discriminant procedure. The logistic regression models of classification gives a good basis of comparison since all the 31 models can be compared for both logistic and CART model. Comparison of Tables 7.3 and Table 7.4 reveals that the correct classification rate, specificity and sensitivity are very high and the values of false positive rate and false negative rate are quite low for CART models than those of the logistic regression models. Although, models built with one or two explanatory variables provide almost the same accuracy in both the methods, but as the number of variables involved in model building increases CART begins to outperform the logistic models. While the correct classification rates of the logistic models using four variables is around 90%, the same for the CART method is around 95%. A common phenomenon can be observed in both the methods that is, the effect of the variable X_2. Inclusion of this variable in the model does not bring much improvement in the classification accuracy of the model. Since the proposed method was developed by considering all the five variables in building the classifier, hence comparison was possible only with the conventional CART methodology and the logistic regression method of classification. From Tables 7.3 and Table 7.7, it can be concluded that although the proposed method of growing a tree classifier does not outperform the CART methodology in terms of classification accuracies, but it is as good as the existing CART procedure in that aspect. Comparing with the logistic regression model obtained from all the five variables, it can be observed that the proposed tree-based method obviously outperforms this method for classifying new observations in terms of correct classification rates. In all, it can be concluded that the tree-based method of classification which are a new approach towards such problems, are better performing than the conventional procedures, and so does the new proposed methodology. Also the proposed CART method has advantage over the conventional method in terms of computational speed and flexibility while handling large data sets.

References

Breiman, L., Freidman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth, Belmont CA.

Cappeli, C., Mola, F. and Siciliano, R. (2002). A statistical approach to growing a reliable honest tree. *Computational statistics and data analysis*, **38**, 285-299.

Dalenius, T. and Hodges, J. L. (1957). The choice of stratification points. *Skand. Akt.*, **34**, 198-203.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification and manifold learning*. Springer, New York.

Loh, W. Y. and Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, **7**, 815-840.

Scheuler, E. M. and Stoller, S. (1962). On the generation of normal random vectors. *Technometrics*, **4(2)**, 278-281.

Shih, Y.S. (2001). Selecting the best categorical split for classification trees. *Statisics and Probability Letters*, **54**, 341-345.

Author for correspondence

Sandip Kumar Sadhu
National Academy of Statistical Administration,
MoSPI, New Delhi

Received: 26 December, 2013 *Revised: 31 July, 2014* *Accepted: 10 November, 2014*