THE CRITERIA FOR CLASSIFICATION TREE METHODS IN CLINICAL RESEARCHES*

*Zeki Akkus[1**], S.Yavuz Sanisoglu[2], Mehmet Ugurlu[3], M. Yusuf Celik[4]*

1. Assoc. Prof, Department of Biostatistics, Medical Informatics Dicle University, Diyarbakir / TURKEY.
2. Assoc. Prof, Consultant, Turkish Ministry of Health, Ankara / TURKEY.
3. MD, Field Coordinator, Turkish Ministry of Health, Ankara / TURKEY.
4. Prof. Dr. Department of Biostatistics, Medical Informatics Dicle University, Diyarbakir / TURKEY.

Abstract

This study aimed at evaluating a statistical method, classification tree, which are recently developed parallel to the improvements in computer technology. The advantages over other methods and the criterions developed for classification tree are reported in this study. Classification tree (CT) is a non-parametric statistical method using a tree algorithm for reaching diagnosis by utilizing one or more risk factors.

Classifications (discriminative, logistic regression and cluster analysis etc) and regression methods are frequently employed in analysing data acquired from scientific studies. However, hypothesis in these models makes the statistical analysis limited to be performed in wide range of disciplines. As there is no need for hypothesis in analysing these data sets, classification trees are serious alternative for other statistical classification and regression techniques.

Classification tree, also known as Decision tree, is a good choice for data mining classifications in respect to both understanding and explaning the some particular rules about estimating the results. These methods are evolved following the improvements in computer technology.

Classification tree is becoming more important in practice as it provides reliable measures in building accurate classifications. The advantages of the method over others are the following: simplification of the results, provision of non-parametric and lineer solutions, generalization of the conclusions optained by inductive reasoning. More over the technique can utilize mixed data types and the same variable can be employed in different parts of the tree.

The determination of choices, which is crucially important in accurate interpretation of the results, needs time and effort in practicing the method. In field of medicine, classification tree is one of the favorable methods particulary utilized in clinical studies.

(*J Int Dent Med Res 2010; 3: (2), pp. 88-92*)

Keywords: Clinical research and classification tree method.

Received date: 22 March 2010 **Accept date: 15 july 2010**

Introduction

This study aimed at evaluating a statistical method, classification tree, which are recently developed parallel to the improvements in computer technology. Classification Trees were introduced during the early 90s by Grimm and Grochtmann for the structured representation of test cases[1,2].

Classisfication Trees (CT) is a nonparametric technique that can select from among a large number of variables those and their interactions that are most important in determining the outcome variable to be explained[3].

Classification tree is an observational method used in order to classify explanatory variables. Common features of the classification tree methods can be listed as follows:

1- Merging: In this method, relative to the target variable, non-significant predictor categories are

*****A part of the study was presented as a poster in the IX. National Biostatistics Congress 2006, Zonguldak/Turkey and its abstract was published in the congress booklet.

*****Corresponding author:***
Assoc. Prof. Zeki AKKUŞ,
Dicle University Medical School, Department of Biostatistics , Medical Informatics ,21280 Diyarbakir / TURKEY.

E-mail: zakkus@dicle.edu.tr

grouped with the significant categories.
2- Splitting: In this method, variable that distinguishes the split point population is chosen to be compared with all others.
3-Stopping: This method determines how far to extend the splitting of nodes.
4- Pruning: Determines the removed branches.

Trees are a completely different way of partitioning. All we require is that the partition can be achieved by successive binary partitions based on the different predictors. The main questions related to creating the classification tree are: How to (1) select the splits, (2) determine the terminal nodes, and (3) assign the terminal node a class?.[4]

Classification tree methods known also as decision trees can be accepted as a good method in the classification of data mining, estimation of the results and easy understanding and explanation of some rules. These methods were emerged in recent years by the developments in the computer technology.

Classification models can be created using various statistical approaches, including generalized linear models (GLM) such as logistic regression, generalized additive models which are semi-parametric extensions of GLMs, and fully nonparametric methods such as classification trees[5].

Classification tree gains more importance since it yields confidence measurements in accurate classification. Advantages of the methods against the other ones are listed as follows: Simplifies the results after the analysis, ensures non-parametric and non-linear results, gets results that may be generalized by means of induction, may use mixed data types, same variable can be reused in the different parts of a tree.

Two algorithms are used in the classification tree: First one is classification and regression trees (CART) and the second one is the QUEST algorithms formed of quick, unbiased, efficient, statistical trees.[6] Since the variables used in CART analysis are independent from the distributions, it ensures a great convenience to the users.

Classification Trees

In CT analysis there are four basic steps. First step is the structuring of the tree. Second step is stopping the tree structuring method. Third step is called as pruning and fourth step is called as optimal tree selection. Tree construction begins on the main node by handling all observations. Method tries to create a separate node by checking the possible sub-variables and all variable values in order to find out the best variable. Possible sub groups in the categorical variables are divided into the number of categorical variables quickly. Therefore, it is beneficial to determine the maximum class numbers in each categorical variable in the program.

Determining the node classes is carried out as follows: Including root node, each node is determinant in class formation. Each node is dependent on three factors in class formation.
1- Priority possibility of each class in data set,
2- Decision or cost matrix,
3- Distinction of the observations stopped in each node.

In a classification model, error rate is calculated as the proportion of mis-classified events to the entire events and accuracy rate is calculated by dividing the number of the accurate events to the number of the entire events (Accuracy Rate = 1- Error Rate).

Risk matrix is used in order to decide on the error rates of the models established to classify the data. In the result of the distinction, the most appropriate class to be assigned for any node is estimated as follows:
Where;
C (j / i) : cost of classifying i class as j class (coefficients of risk matrix),
π_i : former probability of class i,
N_i : number of trial units found in dataset
$N_i^{(t)}$: number of the trial units found in class i of node t,
If the inequation

$$\frac{C(j/i)\pi_i N_i^{(t)}}{C(i/j)\pi_j N_j^{(t)}} > \frac{N_i}{N_j}$$

is ensured for all values of j (j = 1, 2, …. k and j≠ i), class i is assigned to node t as the most appropriate class.[7]

Classification tree is a non-parametric statistical method developed to estimate the values of a dependent variable in a categorical structure.[7,8]

In the classification trees, there are three alternative accuracy estimation methods. These are replacement estimation, test sample estimation and cross validity test.[7-9]

Even in the situations where dataset is very complex, CART may display the variables that affect the dependent variable and the significance of these variables in the model within a simple tree structure. If the handled dependent variable is in a categorical structure, method is called as Classification Tree, CT and if they are continuous it is called as Regression Tree, RT.[7-9]

Geometrical Structure of the Classification Tree

For a better understanding of the geometrical structure of the classification tree, it should be examined visually how the observations are classified on a geometrical plane.

Such examination will be displayed by means of a graphical distinction that is realized by the multiple linear discriminating analysis that has a strong mathematical structure. Cutoff planes on the classification of data for discriminant model are shown in the following model.

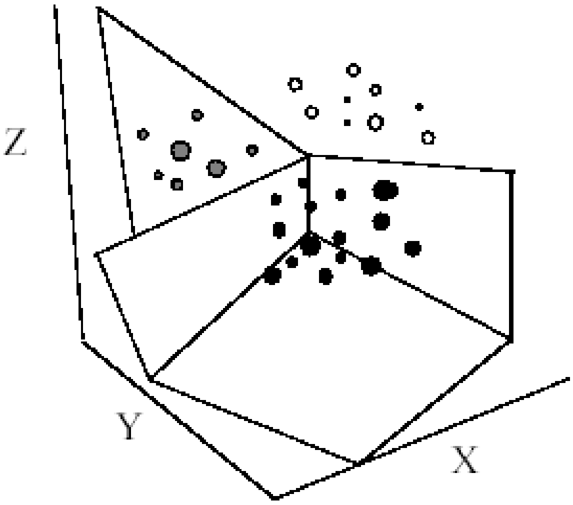

Figure 1. Cutoff planes for discriminant model

When we examine the graphic above, we can see the explanatory variables of X, Y, Z and four subgroups of data. Four subgroups are determined as black, shadowed, white and hidden. Fourth group is not seen on the graphic. Because it lays under the plane. Although there are three explanatory and six cutoff planes for four groups, only four planes are seen in Graphic 1. Generally, if there is group k, k.(k - 1)/2 plane is formed in linear discriminant model.

Cut off planes on the classification of data for classification tree are shown in the following graphic.

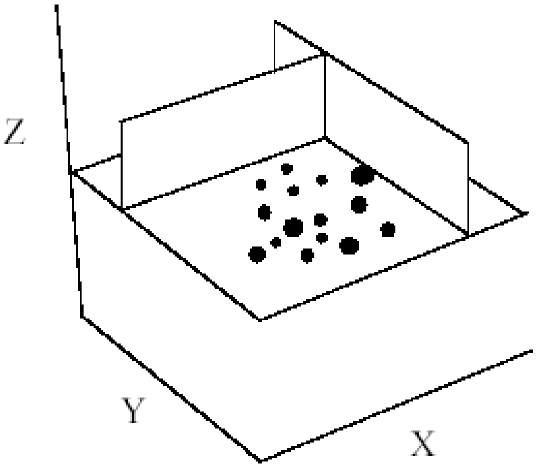

Figure 2. Cutoff planes for tree model.

When we examine the graphic above, the most significant difference here is that the cutoff planes are parallel with the axis. In Graphic 2, we can only see the black dotted group that is the closest subgroup. Others cannot be seen since they lay under the planes.

Although such situation restricts the flexibility of the planes, tree model allows for the interaction between the variables that are not seen in the sequenced linear discriminant model.[10]

When the study results are examined, it is seen that the classification tree is able to distinct the complex problems into simple and comprehensible sub problems.[11]

Advantages and disadvantages of the Method

Classification trees are computationally efficient, can handle mixed variables (continuous and discrete) easily and the rules generated by them are relatively easy to interpret and understand[12].

Increase in the usage rate of classification tree models is connected with the following reasons:

- Since CT is a non-parametric model, its assumptions are limited.
- In the model there is not any assumption and limitation regarding the types of the variables (continuous, categorical, sequential or mixed).
- Since the relationship between the

dependent and independent variables have a visual presentations, model results in form of a tree can be interpreted easily without having the necessity for a lot of statistical information.

- For the defined dependent variable, CT includes all possible independent variables and combinations of the model and performs the most accurate classification possible.
- It is easily applied on the complex datasets.
- It is a method that remains unaffected from the lost or missing values and also from the extreme values for both dependent and independent variables.
- It is an alternative for many traditional statistic techniques (multiple regression, variance analysis, logistic regression, discriminant analysis, grouping analysis).
- It considers the tree methods that are not certain but based on solid grounds.
- It is a practical method in ensuring objective results in complex and broad datasets.

On the other hand, CT method has some disadvantages as listed below:

- May have unstable decision trees.
- Splits only by one variable [13]
- The tree-space is huge, so it may need a lot of data.
- It can be hard to assess uncertainty in inference about trees.
- Actual additivity becomes a mess in a binary tree.
- Simple trees usually do not have a lot of predictive power.
- There is a selection bias for the splits[14].

Conclusions

CART (classification and regression trees), has been used extensively as a means for clinical risk assessments[15,16]. In comparison with other statistical methods, CART analysis has been shown to perform equally or better than logistic regressions[17-20], discriminant function analysis[21,22], and neural networks[23]. In this study, we aimed to give some useful information about classification trees.

As a result, use of Classification Tree method is more convenient than the other methods since it is a computer based analysis method and has a non-parametric feature.

Declaration of Interest

The author report no conflict of interest and the article is not funded or supported by any research grant.

References

1.Grimm, K., “Systematisches Testen von Software - Eine neue Methode und eine effektive Teststrategie (Systematic Software Testing–A new method and an effective test strategy), GMD-Report-251, GMD, Oldenbourg, 1995.
2.Grochtmann, M,Grimm K. Classification Trees for Partition Testing, Software Testing, Verification and Reliability, 1993,3(2), 63–82.
3. Yohannes Y, Hoddinott J. Classification and regression trees-An Introduction, International Food Policy Research Institute, Technical Guide #3, USA,1999.
4. Creating Imputation Classes Using Classification Tree Methodology, Creel1 D V, Krotki K,1RTI International, ASA Section on Survey Research Methods,2884-2887.
5.Edwards T C, Cutler D R, Zimmermann N E, Moisen L G and G. Effects of sample survey design on the accuracy of classification tree models in species distribution odels, Ecological Modelling, 2006, 199(2), 132-141.
6. Lewis R. An introduction to classification and regression tree (CART) analysis. Academic Emergency Medicine.California, 2004; 1-14.
7. Fu CY. Combining loglinear model with classification and regression tree (CART): An application to birth data . Computational Statistics&Data Analysis. 2004;45(4):865-874.
8. Breiman L, Friedman JH, Stone CJ, Olshen RA. Classification and Regression Trees, Boca Raton, Florida: Chapman&Hall . 2003;18-23.
9. Temel GO, Çamdeviren H, Akkuş Z. Sınıflama Ağaçları Yardımıyla Restless Legs Syndrome (RLS) Hastalarına Tanı Koyma. Journal Of İnönü Üniversity School Of Medicine 2005;12(2):111-117.
10. Wilkinson L. Graphical displays.Statistical Methods in Medical Research.1992;(1):3-25.
11. Bittencourt H R, Clarke RT. Feature selection by using classification and regression trees (CART) International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences. 2004;35(7):66-70.
12.Ding Y, Simonoff J S. An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data, Journal of Machine Learning Research,2010,11: 131-170
13.Timofeev R. Classification and Regression trees Theory and applicances(Master thesis).2004;20-25
14.Breiman L. Friedman J H, Olshen R A, Stone C J.*Classification and Regression Trees*, Wadsworth Inc, 1984.
15. Steadman H J, Silver E, Monahan J et al. A classification tree approach to the development of actuarial violence risk assessment tools. Law Human Behav., 2000, 24:83-100.
16. Huland H. Radical prostatectomy: Options and issues,Eur. Urol.,2001, 39: 3 -9.
17. Germanson, T P, Lanzino G, Kongable G L et al.Risk classification after aneurysmal sub-arachnoid hemorrhage. Surg. Neurol.1998, 49: 155-163.
18. Rudolfer S M, Paliouras G,Peers I S. A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. Comput Biomed Res,1999, 32: 391-414.
19. Wietlisbach V, Vader J P, Porchet F et al. Statistical approaches in the development of clinical practice guidelines from expert panels: the case of laminectomy in sciatica patients. Med Care, 1999.37: 785-797.
20.Vayssie`res M P, Plant R E,Allen-Diaz B H,Classification trees: An alternative non-parametric approach for predicting species distributions, Veg Sci, 2000, 11: 679-694.

21. Smith S J, Iverson S J, Bowen W D,Fatty acid signatures and classification trees: new tools for investigating the foraging ecology of seals, Can J Fish Aquat Sci, 1997,54:1377-1386.

22. Kirkwood C A, Andrews B J,Mowforth P, Automatic detection of gait events: a case study using inductive learning techniques, J Biomed Eng, 1989. 11: 511-516.

23. Selker H P, Griffith J L, Patil S et al., A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients, J Invest Med, 1995. 43: 468-476..