Available online at www.elixirpublishers.com (Elixir International Journal)

Bio Technology

Elixir Bio Tech. 40 (2011) 5407-5414

ISSN: 2229-712X

Fundamentals of classification and regression trees

Qamruz Zaman[1], Muhammad Azam[2], Shah Khusro[3], Muhammad Iqbal[1] and Karl Peter Pfeiffer[4]

[1]Department of Statistics, University of Peshawar, Pakistan

[2]Department of Statistics, Forman Christian College, Lahore

[3]Department of Computer Science, University of Peshawar, Pakistan

[4]Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Austria.

ARTICLE INFO

Article history:

Received: 20 August 2011;

Received in revised form:

25 October 2011;

Accepted: 4 November 2011;

Keywords

Classification and regression trees, Misclassification, Gini Index, Software's, Resubstitution estimates.

ABSTRACT

This paper reviews the newly developed statistical analysis technique: classification and regression trees (CART), which is one of the most frequently used methods for complex and multi dimensional data sets. CART is a non parametric approach and can easily be used. A most popular and commonly used Gini index method for node impurity in classification trees and the concept of standard error for the measurement of accuracy in regression trees has been discussed. Important features and uses of CART in different fields especially in medical research have also been discussed with examples.

Introduction

During the last two decades, the use of classification and regression tree (CART) analysis has been increased. Classification and regression trees techniques are better choices for the analysis of complex and high dimensionality data sets. CART procedures are used in different fields such as medical, engineering, marketing, sociology and economics research. Examples of the use of techniques in these fields are segmentation, credit risk assessment, quality control, degree of spread of cancer, classification of blood cells, infant mortality, wildlife management, air pollution alerts, speech recognition, identification of high risk investment strategies and classification of radar images for the military (CART, 1995). Classification and regression trees (CART) are comparatively new statistical technique which is used to predict the ordered/continuous and unordered/categorical variables.

Its algorithm generates simple binary nodes from root node and finally one gets a binary tree structure and uses it to classify the objects with respect to their classes.

It was introduced by group of well known researchers Breiman, Friedman, Olshen and Stone (1984).

This technique divides the set of individuals or objects into finite number of classes on the basis of collected characteristics and predictors.

It uses binary trees method which was introduced by Morgan and Sonquist in 1960s at the university of Michigan, furthermore Morgan and Messenger made developments in the pioneer classification technique (Feldman et al., 2005).

The general procedure can be briefly explained by constructing a binary tree. Let the set of binary questions b_Q generates a set S of splits s of every node t. Those cases in t answering 'Yes' go to the left descendant node t_L and those answering 'No' go to the right descendant node t_R. Figure1.

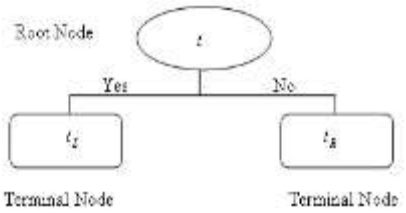

Figure 1Tree construction

Mathematically, we can write as $t_L = t \cap E$ and $t_R = t \cap E^C$. Where E is any element of t and E^C is its compliment. At each intermediate node t, the split selected is that split s^* which maximizes $\delta(s, t)$.

CART is one of the most popular statistical techniques used in medical research. Where one wishes to classify the newly admitted patient to hospital at the high/low risk of certain disease on the basis of information obtained. The concept can also be explained by the help of daily life example.

Consider the famous heart attack data set (Gilpin et al., 1983 and Breiman et al., 1984). Data set consisted of 215 heart attack patients. When a heart attack patient was admitted to the hospital, 19 ordered/ categorical variables were measured during the first 24 hours. These included patient's blood pressure (BP), age, presence of sinus tachycardia (ST), enzyme concentrations based on blood work etc. The main objective of this study was to develop a method which identifies high/low risk patients (those who will not/will survive at least 30 days) on the basis of 24 hours data. Within 30 days following the admission, 37 patients died (falling in high risk group) while 178 patients survived (falling in low risk group). Based on the information from all 215 heart attack patients a classification tree has been constructed, which is shown in figure 2. Summary statistics is given in table 1.

Tele:

E-mail addresses: ayanqamar@gmail.com

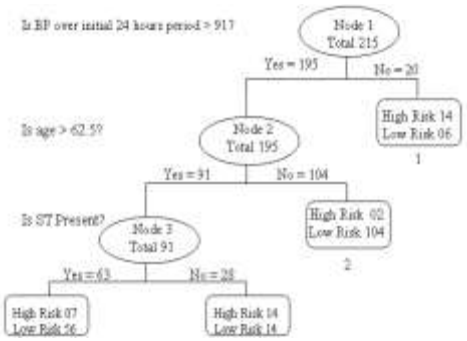

Figure 2: Binary Decision Tree of 215 Heart Patients

Figure 2 illustrated the procedure of classifying the patients into two classes. A terminal node containing higher number of high/ low risk patients is declared as 1 or 2 respectively. A classification rule has been applied using only three variables to find the particular class of the patient. It was revealed that the accuracy of CART classification was higher as compared to conventional/classical methods like multiple regression analysis, ANOVA, Logistic Regression, Log Linear Models and Linear Discriminant Analysis (LDA) etc.

Unlike other methods, CART does not require any additional information and can be used easily by any body. Once the classification and regression tree has been constructed on the basis of categorical/continuous variables, it becomes easier to classify the new objects like patients, into their proper class on the basis of information obtained.

CART is a non parametric technique which is used to represent the decision rules in the form of binary trees. These trees split the learning sample data (data under study) in such a way that

First/root node consists of learning sample.

Last/terminal nodes of a tree can be obtain by continuously splitting parent nodes until the end nodes contain only one or any specified number of units.

The analysis of CART deals with the categorical as well as continuous data. If the dependent variable is categorical, the resultant tree is called classification tree. While, the trees constructed with continuous dependent variables are called regression trees, which are representing the non parametric regression model describing possible relationship among different variables under study.

Actually, CART analysis is tree constructing technique unlike the other statistical techniques like linear discriminant analysis cluster analysis etc.

In the start, researchers were hesitant while using it, due to its complexity. But with the passage of time and development of CART software (CART, 1995), makes it more popular. By using the CART software one can easily get more accurate results as compared to other conventional techniques.

We divided the whole paper in eight sections. Major components of CART are described in section 2, merits of CART in section 3, tree growing procedure and application of CART in different fields of life specially in medical science has been considered in section in section 4, regression trees and tree growing procedure for regression with an example have been discussed in section 5, Software's used for the construction of classification and regression trees in section 6 and conclusion in section 7.

Major components of classification and regression trees

Following are the main components of classification and regression trees.

Categorical/continuous dependent variable: A variable that we want to predict based on number of predictors (categorical and continuous) is called the dependent/ response variable. The dependent variables which are often considered include presence/ absence of some kind of disease, survival, weather forecast etc. If the dependent variable is continuous, we get regression trees otherwise we obtain classification trees.

Predictors/input variables: These are the variables which are used to predict the dependent variable. Predictor/ Input variables may be continuous or categorical or mix.

Learning dataset: A dataset which is used in the analysis of classification and regression trees is known as learning dataset. It consists of values of both (output and input) variables. Further it is divided into training and test dataset.

Training sample is the subset of learning sample which is used directly to find the desired results. While a subset of learning sample which is used to validate the findings obtained by using the training sample.

Merits of cart

When we compare classification and regression trees analysis technique with other traditional multivariate data analysis techniques. Then CART holds number of advantages over other techniques.

- CART is an efficient non parametric technique (Feldman et al., 2005). Therefore no assumptions are made regarding the underlying distribution of values of the predictor variables.
- CART has the ability to handle ordered/continuous and unordered/categorical variables.
- As for as computational efficiency is concerned. CART performs better than any other classification or multivariate data analysis techniques like logistic regression etc.
- CART is capable to identify most significant predictors during a comprehensive search. It searches all the predictors, finally chooses most effective splits/predictors.
- CART can tackle any number of predictors, provides an opportunity to use maximum number of predictors which present more realistic prediction and give more accurate results.
- CART has the facility to handle missing observations in both dependent and independent variables. For these CART provides substitute values (known as surrogate) containing information similar to that contained in the primary splitter.
- CART works in such a way that if a dataset contains outliers, it neutralizes the effects of these by data splitting property. Therefore outliers cannot affect the results.
- CART deals with machine learning/computer based algorithms; saves a lot of time during the different steps involve in construction, stopping, pruning etc. process.
- Due to its straight forward and simple decision rules, CART seems be more popular among the non-statisticians/ clinicians.

Number of empirical as well as experimental comparisons have been conducted by different authors including Rousu et al., (2003), Moisen and Frescino (2002) etc., for the purpose to check the efficiency of this approach.

Tree growing procedure

There are four elements which are required to construct a tree.

A set b_Q of binary questions? Questions may be about sex (male/ female), marital status (married/ unmarried), a disease (high risk/ low risk) etc.

The second most important step in the tree construction is the application of goodness of split criteria and its numerical

measurement. A goodness of split criterion $\delta(s, t)$ can be computed for any split s of any node t.

i. A stop-splitting rule. In this step one has to decide when and where node splitting should be stopped.

ii. A particular class is assigned to each terminal node on the basis of majority vote.

A brief explanation of all four steps is given in the following section.

A standard set of questions

If the data have standard structure, the class Q of questions can be standardized (Breiman et al., 1984). Suppose we have a data set of form

$$X = (x_1, \text{L}, x_P),$$

where the variables/ predictors x_1, L, x_P can be mixture of ordered/ continuous and unordered/ categorical variables of the fixed dimensionality *P*. The standardized set of questions Q can be defined as follows:

Each split depends on the value of only a single variable.

If x_P is ordered/ continuous variable then it includes all questions of the form $\{\text{Is } x_P \le x?\}$, where x is any specific value ranges from $(-\infty, \infty)$. All the values that satisfy the above condition go to the left descendent node t_L and the values do not satisfy the condition go to the right descendent node t_R. Collectively, both these nodes are known as the child/ descendent nodes of root node t.

Similarly, if x_P is unordered/ categorical variable then it include all the questions of the form $\{\text{Is } x_P \in C?\}$, where *C* is any subset from the set of categorical values $\{c_1, c_2, \text{L}, c_L\}$. The splits 2 and 3 for all *P* variables make the standard set of questions.

If we have *n* number of distinct values for an ordered/ continuous variable x_P, then there are at most '*n*' possible splits (Breiman et al., 1984) for the variable x_P of the form $\{\text{Is } x_P \le x?\}$.

Similarly if x_P is categorical, there are $(2^{P-1} - 1)$ possible splits. Node impurity measure criterion is used to find the best split. The most common impurity measure criterion includes Gini index and Twoing method. For finding the best split, CART developers preferred to use these techniques (Breiman et al., 1984). Both methods are mathematically equivalent for the two class problems (CART, 1995). Few other methods which are used in literature for the measurement of purity of nodes include the chi-square measure (Hart, 1984 and Mingers, 1987), entropy or variants (Quinlan, 1986) and the G statistic (Mingers, 1987). Lot of efforts have been made for the improvement of impurity measure, still there appears to be no single method which gives best results in every situations (Kothari et al, 2000).

At each node the tree algorithm searches through all variables from x_1 to x_P. During the searching process it finds the best split for each predictor. After obtaining the best split for each predictor, it compares the *P* best single variable splits and selects the best among them for that particular node. This process is repeated for all the nodes.

Node Splitting and Stopping Rule

The goodness of split criterion was originally derived from an impurity function. The function decides when and where to split and to stop splitting a node.

Node impurity

The concept of node impurity has been explained by number of authors like Breiman et al. (1984), Berzal et al. (2003) etc. If the observations of a dataset falls into a particular node belongs to the same class and none of the observation belongs to the other class, we say that there is no impurity or the node is pure. The impurity of a node increases and reaches/ achieves its maximum, as the proportion of observations belong to different classes becomes closer and closer to each other. Suppose we have a node t, impurity measure (binary or two class problems) of that node is denoted by $i(t)$ and is given by

$$i(t) = \sum_{1 \ne 2} P(t_{c_1}) P(t_{c_2}), \quad (1)$$

$$i(t) = 1 - \sum_{h=1}^{2} \{P(t_{c_h})\}^2 \quad (2)$$

such that

$$P(t_{c_1}) + P(t_{c_2}) = 1, \quad (3)$$

where $P(t_{c_1})$ and $P(t_{c_2})$ are the probabilities of observations fall into the terminal nodes belonging to classes 1 and 2 respectively. Similarly impurity measure for the whole tree T containing number of terminal nodes (say t^*) can be measured as follows

$$i(T) = \sum_{t \in t^*} i(t) P(t). \quad (4)$$

Similarly, if there are more than two classes in a particular node then the measure of node impurity will become

$$i(t) = \sum_{h \ne j}^{H} P(t_{c_h}) P(t_{c_j}), \quad (5)$$

$$i(t) = 1 - \sum_{h=1}^{H} \{P(t_{c_h})\}^2 \quad (6)$$

such that

$$P(t_{c_1}) + P(t_{c_2}) + \text{K} + P(t_{c_H}) = 1. \quad (7)$$

Hence, the result $i(t) = 1 - \sum_{h=1}^{H} \{P(t_{c_h})\}^2$ is called the Gini index. The index is simple to understand and fast in computation. The measure of goodness of split or measure of decrease in impurity of a node t is given by

$$g(t) = i(t) - i(t_L) P(t_L) - i(t_R) P(t_R). \quad (8)$$

Any function $i(t)$ discussed above is said to be an impurity function if it possesses the following properties.

(i) $i(t)$ is maximum only at the point $\left(\frac{1}{H}, \frac{1}{H}, \text{L}, \frac{1}{H}\right)$. It means that the node impurity reaches its maximum, when there is same number of observations fall in a node belongs to H number of classes.

(ii) $i(t)$ achieves its minimum only at the point

$$(1, 0, \text{L}, 0), (0, 1, \text{L}, 0), \text{L}, (0, 0, \text{L}, 1).$$

It means that the node impurity is minimum, when all the observations of a node belong to a single class.

(iii) $i(t)$ is a symmetric function of $\{P(t_{c_1}), P(t_{c_2}), \text{K}, P(t_{c_H})\}$.

Suppose we have done some splitting and arrived at a current set of terminal nodes. The set of splits used, together with the order in which they were used, determines what we call a binary tree T. Denote the current set of terminal nodes (final nodes) by t^*; set $I(t) = i(t)\,p(t)$, and define the overall tree impurity $I(T)$ by

$$I(T) = \sum_{t \in t^*} I(t) = \sum_{t \in t^*} i(t) p(t).$$

Selecting the splits that maximize $g(t)$ is equivalent to selecting those splits that minimize the overall tree impurity $I(T)$. Take any node $t \in t^*$ and using a split s, split the node t into t_L and t_R. The new tree T^* has impurity

$$I(T) = \sum_{t \in t^*} I(t) = \sum_{t \in t^*} i(t) p(t).$$

The decrease in tree impurity is

$$I(T) - I(T^*) = I(t) - I(t_L) + I(t_R).$$

This depends only on the node t and split s.

The Class Assignment Rule

The final step in the tree growing process is the assignment of a particular class to each terminal node t. Suppose there are t^* total number of terminal nodes and the values within each terminal node are classified into h; $(h = 1, 2, \text{K}, H)$ classes. Each terminal node t contains $N_h(t)$ number of values for each class. Therefore a terminal node containing $Max\{N_h(t)\}$ i.e. maximum number of values for any particular class h; $(h = 1, 2, \text{K}, H)$ is declared as that class $h(t) = h$. In other words, if a particular class achieves maximum probability among all other classes in any node t is declared as class h; $(h = 1, 2, \text{K}, H)$.

Measure of accuracy

After lot of discussion on required elements for the initial tree growing process, the next question is about certain node accuracy. To compute the accuracy of a particular terminal node t, one wishes to check how much of the value have been misclassified/ resubstitution.

Resubstitution Estimate

The probability of misclassifying the values falling in a certain node is called the resubstitute estimate $r(t)$ and is defined as

$$r(t) = 1 - \max_h \{p(h|t)\},$$

$$r(t) = \min_h \{p(h|t)\}.$$

Similarly, the resubstitution estimate for overall misclassifying the values for a tree T is denoted by $R(T)$ and is given by

$$R(T) = \sum_{t \in t^*} R(t),$$

where

$$R(t) = r(t)\,p(t).$$

Use of CART in medical research

The beginning of the practical use of CART was very limited due to lack of availability of software's as well as expertise. But currently, it became one of the most applied statistical tools as compared to other classical tools for the classification of objects (e.g. logistic regression etc). The following section describes some of the real life examples.

Asthma: Asthma is the sever, fatal disease which affects the large number of people and Western Europe is one of the most affected region of the world by the disease (Rabe, et al., 2000). Although, it affects people of all ages, but more prevailed in children as compared to adults (among children the regional average is 13% and 8.4% in adults). Some of results based on different studies showed an annual increase of 2 to 4% in most of the European countries (Sears, 1997). Furthermore, there was a rapid increase in asthma patients from 1970 to 1980 in many industrial countries especially in USA, Japan, New Zealand, England and Wales (ISAAC, 1998).

Grassi et al (2001) took1103 asthma patients of three Italian centers of European Community Respiratory Health Survey (ECRHS) and applied different classification methods (Classification and Regression Trees (CART), Fisher's Linear Discriminant Function (LDF) and Neural Network Method (Multi-Layer Perceptron, MLP model))

$$I(T^*) = \sum_{t^* - \{t_L, t_R\}} I(t) + I(t_L) + I(t_R). \quad (11)$$

On the basis of decision rule sensitivity, specificity and accuracy, they concluded that the performance of classification trees is much higher as compared to other methods.

Cancer: For the next application, we considered cancer disease, a most commonly referenced in many theoretical as well as applied medical journals.

Camp et al. (2002) make use of classification tree analysis on a data set of 4403 patients to identify the pathways of causes as well as variables/ high risk factors of colon cancer.

There are number of reasons which cause cancer including smoking (Shopland et al., 1991), diet (Potter et al., 1993, WCRF, 1997 and Potter, 2002) and breast cancer (Kash et al., 1992). Also there are number of factors which reduce the chances of high risk of cancer including physical activity which is inversely associated with colon cancer (Giovannacci et al., 1995). Similarly, the use of non-steroidal anti inflammatory drugs (NSAID) reduces the colorectal cancer incidence (Collet et al., 1999).

Camp et al. (2002) concluded in his findings that the most important factor reduces high risk of colon cancer is use of NSAID, use of proper food, regular exercise etc. Also results showed that the family history plays an important role. This was the first study on multi level model interactions in colon cancer risk using classification tree analysis.

Similarly, Zhang et al. (2001) introduced a method based on classification trees and suggested that the results are more accurate as compared to other statistical approaches used for discriminating among distinct colon cancer tissues. A dataset containing 2000 genes using an affymetrix oligonuleotide array in 22 normal and 40 colon cancer tissues was used. The data was divided into four terminal nodes after applying recursive partitioning scheme. It is observed that only 1 out of 62 tissues has been misclassified (i.e. misclassification rate=1.61%), shows the results in favour of classification trees method. Others who did work on this disease includes Golube et al. (1999) used supervised learning and derived discriminant decision rules,

(12)

(13)

Brown et al. (2000) classified the genes using the DNA microarray hybridization experiments data with the help of support vector machines, Xion et al. (2000) applied LDA (Fisher's Linear Discriminant Analysis) on the data analyzed by Zhang et al. (2001) for the classification of tumor.

Dose absorption in humans: Jane et al. (2004) applied the classification and regression tree technique to predict oral absorption of dose independent range in human beings. A dataset consists of 1260 structures and their human oral pharmacokinetic data (containing 28 predictors) with 899 compounds as training and 361 test set was used. Dose absorption range was taken from 0 to 1 with 6 classes (class interval was approximately 0.16). It is finally concluded that CART is useful in silico prediction of oral absorption and it may not be less accurate than absorption derived from those in vitro laboratories artificial membrane studies.

Regression Trees

Suppose we have n observations on a response variable Y and a set of K predictors. It is assumed that the variable Y depends upon the K predictors and there exist a relationship given by

$Y = h(x) + e$, where $x \varepsilon X$ and e is the random/ error term. The purpose of regression analysis is to predict/ estimate $h(x)$ which will minimize the loss function (Chaudhuri et al., 1994).

Let

$$h(x) = \beta_0 + \beta_1 h_1(x_1) + \beta_2 h_2(x_2) + \text{K} + \beta_K h_K(x_K), \quad (15)$$

where β_k, $(k = 0, 1, \text{K}, K)$ are unknown parameters, while $h_k(x_k)$ are known functions. This problem comes under the umbrella of parametric regression. Nonparametric regression techniques are also available, which includes B- splines and smoothing splines (de Boor, 1978), Eubank (1988), Ramsay(1988), kernal smoothers (Gausser and Muller, 1979, 1984), project pursuit regression (Friedman and Stuetzle, 1981) etc. Another nonparametric technique which is based on the recursive partitioning of sample space is known as regression tree analysis. The regression trees approach to predict a continuous dependent variable is much simple as compared to classification trees approach. Regression trees procedure is one of the powerful statistical tools for describing and organizing knowledge of any discipline (Ciampi, 1991). The use of tree structure idea in least squares regression goes back to the development of Automatic Interaction Detection (AID) algorithm by Morgan and Sonquist (1963). Further developments in AID are due to Sonquist and Morgan (1964), Fielding (1977), Van Eck (1980) and many others.

The classification and regression trees (CART) by Breiman et al. (1984) was the next development in tree structure approaches. Later on smoothed and unsmoothed piecewise polynomial regression trees (SUPPORT) was introduced by Chaudhuri et al. (1994). The FIRM (Kass,1975 and Hawkins, 1997) method discussed the bias problem by using Bonferroni adjusted significance tests to select predictors for splitting. Loh(2002) developed a new algorithm called GUIDE (Generalized Unbiased Interaction Detection and Estimation) to build piecewise constant and piecewise linear regression models with univariate splits. The AID algorithm works in such a way that at each stage, the binary partitioning is made which minimizes the total sum of squares of residuals (SSR). Splitting process stops, when the fractional decrease in total SSR is less than α, where α is some pre specified/ threshold value for stopping the splitting process. The pre specification of α is one of the drawback in AID, because too small or too large value of α may cause over or under fitting (Loh, 2002).

The CART (Breiman et al., 1984) algorithm avoids the use of pre specified value α by applying backward elimination strategy to get the tree. Generally, it grows with maximum number of terminal nodes and prunes away some of its nodes using a certain criteria (e.g. cross validation). Both AID and CART are not free from selection bias because of greed (G) search approach. The approach is computationally expansive, so one can use an alternate SUPPORT algorithm, which uses sample mean as splitting point. The resultant value is as good as any other split value. The procedure gives better results for normal data as compared to the skewed data and this problem can be overcome by using sample median.

The concept of regression tree can be explained by considering the famous data set (Harrison and Robinfeld, 1978) which was also used by (Breiman et al. 1984). Data set consists of 506 cases. For each case, 14 variables have been measured as

y : Median value of homes in thousands of dollars

x_1 : Crime rate

x_2 : Percent land zoned for lots

x_3 : Percent non retail business

x_4 : 1 represents Charles River and 0 otherwise

x_5 : Nitrogen oxide concentration

x_6 : Average number of rooms

x_7 : Percent built before 1940

x_8 : Weighted distance to employment centers

x_9 : Accessibility to radial highways

x_{10} : Tax rate

x_{11} : Pupil/ teacher ratio

x_{12} : Percent black

x_{13} : Percent lower status population

Harrison and Rubinfeld (1978) fitted the least squares regression equation to the data after number of transformations.

$$\begin{aligned}\log(y) = {} & \alpha_1 + \alpha_2 (x_6)^2 + \alpha_3 (x_7) + \alpha_4 \log(x_8) + \alpha_5 \log(x_9) + \alpha_6 (x_{10}) + \alpha_7 (x_{11}) \\ & + \alpha_8 (x_{12} - 63)^2 + \alpha_9 \log(x_{13}) + \alpha_{10} (x_{10}) + \alpha_{11} (x_2) + \alpha_{12} (x_3) + \alpha_{13} (x_4) \\ & + \alpha_{14} (x_5)^b + e \end{aligned} \quad (16)$$

where b is a parameter to be estimated.

The regression tree for the data set is shown in figure 3.

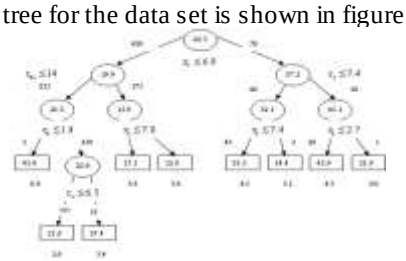

Figure 3. Boston housing values data

The value within each node is the average of the response variable. Each node has been splitted according to condition given exactly below each node.

Number of cases going left or right are mentioned along with the arrows, an easy way to understand and interpret the results as compared to equation (16). One can easily predict the class of a new case by using the tree in the above figure as a predictor. The above tree has been grown using the procedure discussed in the next section.

Least Square Regression

Suppose there are N elements/ cases and each case contains the data on x and y variables. Where x belongs to X and y is a real valued number. Here x represents the independent/ predictor variables and y the dependent/ response variable. Following are different steps involved in the least square regression process

Prediction Rule

It is a function $c(x)$ depending on the values of X, listed in the learning sample L. A prediction rule mainly based on two purposes.

(i) Estimation of dependent variable on the basis of future values of X with maximum achievable accuracy.

(ii) To find and understand the existing relationship between response and predictor variable.

Measure of accuracy

Suppose a learning sample L contains n elements/ cases $(x_i, y_i), i = 1, 2, \mathrm{K}, n$. A predictor $c(x)$ has been constructed and for testing the accuracy of predictor $c(x)$, we took a very large sample contains n_L elements/ cases $(x_j, y_j), j = 1, 2, \mathrm{K}, n_L$. The closeness/ accuracy of predictor $c(x)$ in this case can be achieved by averaging the absolute error terms i.e.

$$AE_{ave} = \frac{1}{n_L}\sum_{j=1}^{n_L}\left|y_j - c(x_j)\right|. \tag{16}$$

AE_{ave} is known as least absolute deviation regression. As compared to least absolute deviation regression method of accuracy, the method based on squared error is more popular and commonly used. The main attraction of this method is that it minimizes the error sum of squares.

$$R_{ESS} = \frac{1}{n_L}\sum_{j=1}^{n_L}\left[y_j - c(x_j)\right]^2.$$

Theoretically mean squared error $R^*(c)$ is

$$R^*(c) = E\left[Y - c(x)\right]^2$$

One can get the estimate of (1) as

$$\hat{R}^*(c) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - c(x_i)\right]^2.$$

Test Sample Estimate

Suppose we have a large data set (learning sample) L contains n elements/ cases and we divide the data set into two parts L_1 and L_2 (equally or unequally), traditionally with the ratio 1:2. L_1 consists of 1/3 elements and is used to construct c, where L_2 used 2/3 elements to construct $R^{(TS)}(c)$. Where.

$$R^{(TS)}(c) = \frac{1}{n_2}\sum_{i\varepsilon L_2}\left[y_i - c(x_i)\right]^2.$$

V-fold Cross Validation Estimates

Divide the learning sample L into $v\ (v = 1, 2, \mathrm{K}, V)$ parts. And each L_v, $(v = 1, 2, \mathrm{K}, V)$ contains almost equal number of elements/ cases. Obtain the predictor $c^{(v)}(x)$ for each of the learning sample $(L - L_v)$. We compute $R^{(CV)}(c)$ as

$$R^{(CV)}(c) = \frac{1}{n_v}\sum_{v=1}^{V}\sum_{(x_i, y_i)\varepsilon L_v}\left[y_i - c^{(v)}(x_i)\right]^2.$$

Equation (1) can also be written as in terms of constant mean value μ

$$R^*(\mu) = E\left[Y - \mu\right]^2.$$

Which is equal to variance of variable Y. For getting a unit free value for the comparison purposes, the concept of relative mean squared error can be used

$$RE^*(c) = \frac{R^*(c)}{R^*(\mu)}.$$

In case of test sample

$$RE^{(TS)}(c) = \frac{R^{(TS)}(c)}{R^{(TS)}(\bar{y})},$$

where

$$R^{(TS)}(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Similarly, these results can be obtained for the cross validated estimates as under

$$RE^{(CV)}(c) = \frac{R^{(CV)}(c)}{R(\bar{y})} \tag{17}$$

Tree Based Regression Analysis

A tree based predictor in regression analysis is same as tree based classifier in classification analysis. A tree T is grown using the information/ data (x_i, y_i), starting from root node to terminal nodes. Each node is split according to certain threshold/ condition on variable x_i. When a tree T has been grown in full length, the predicted values of $y(t)$ are computed for each terminal node. This is the average of response variable for any particular node (Fig. 2).

Tree Growing Procedure (18)

Following are the elements which are required to grow/ construct a regression tree.

(a) Splitting Criteria at every intermediate node.

(b) Stopping Rule which declare a node as terminal. (19)

(c) A rule of assigning a value to response variable at each terminal node.

A brief introduction of these elements is already given in section 4.

Resubstitution Estimate

The resubstitution estimate of $R^*(c)$ is given by

$$R(c) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - c(x_i)\right]^2,$$

where $c(x_i) = y(t)$ and $y(t)$ is the average of y_i for all cases given in the learning sample (x_i, y_i) falling in that particular node t. In terms of resubstitution estimate for a single node t in regression trees, we can write the equation (2) as

$$R(t) = \frac{1}{n}\sum_{x_i \varepsilon t}\left[y_i - \hat{y}(t)\right]^2 \quad (28)$$

where,

$$\hat{y}(t) = \frac{1}{n(t)}\sum_{x_i \varepsilon t}^{n(t)} \hat{y}_i(t) = \frac{1}{n(t)}\sum_{x_i \varepsilon t} y_i \,.$$

Also, the resubstitution estimate for the whole tree T is given by

$$R(T) = \frac{1}{n}\sum_{t \varepsilon T}\sum_{i=1}^{n}\left[y_i - \hat{y}(t)\right]^2 \quad (29)$$

or

$$R(T) = \sum_{t \varepsilon T} R(t) \quad (30)$$

A split s of t considered best among the set of splits S which decreases $R(T)$. For any split s of t into t_L and t_R. we have

$$\delta R(s, t) = R(t) - R(t_L) - R(t_R) \,. \quad (31)$$

Choose the best split $s^{(b)}$ among the set of splits S in such a way.

$$\delta R\left(s^{(b)}, t\right) = \max \sum_{s \varepsilon S} \delta R(s, t) \,. \quad (32)$$

Software's Used for the Construction of Classification and Regression Trees

A large number of classification and regression tree analysis programs as well as software's have been developed. Due to the development of these software's it becomes easy to handle complex as well as datasets with high dimensionality. Few of these programs/ software's are FACT (Loh and Vanichsetakul, 1988), THAID (Morgan and Messenger, 1973), AID (Morgan and Sonquist, 1963), CHAID (Kass, 1980), QUEST (Loh and Shih, 1997), CART (Breiman et al., 1984), C4.5 (Quinlan, 1993).

Conclusion

In this paper, we reviewed a newly developed statistical analysis technique classification and regression trees, applicable to problems related to medical diagnosis, engineering, economics, market research etc. By using the classification and regression trees technique one can obtain results in a short time and meaningful form. One of the main reasons of the growing use of the technique is its nonparametric behaviour and is therefore free of any parametric assumptions. The developing era of machine learning and computer software's makes it possible to handle huge data in a sophisticated manner. Further more, the method can easily handle the small as well as very large data sets without losing the efficiency.

Another important factor which causes its popularity is that of handling categorical, continuous and mix data. Classification and regression trees have the ability to handle missing values and surrogates are the better substitute for these values. To handle the problem of large number of terminal nodes the concept of pruning is to be used, but this may create the problem of increasing the misclassification rate. Much work has to be done to decrease the misclassification rate, so that the researchers can easily apply the classification and regression tree technique in their projects.

There is much space for the improvement of the technique and the Random Forests technique (Breiman, 2001) which combines large number of trees is one of the major steps towards the new development.

Acknowledgement:

This work is supported by the Higher Education Commission of Pakistan.

References

F. Berzal, J.C. Cubero, F. Cuenca, and M. J. Martin-Bautista. On the quest of easy to understand splitting rules. *Data & Knowledge Engineering*, 44, 31-44, 2003.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees*. Wadsworth International Group, Belmont, CA, 1984.

L. Breiman. Random Forests. *Machine Learning*, 45, 5-32, 2001.

M.P.S. Brown, W.N. Grundi, D. Lin, N. Christianini, C.W.Sugnet, T.S. Furey, M. Jr. Ares and D. Huassler. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 97, 262-267, 2000.

N.J. Camp and M.L. Slattery. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes and Control*, 13, 813-823, 2002.

CART. *Salford Systems*, San Diego, California, USA, 1995.

P. Chaudhuri, M.C. Haung, Y.W. Loh, and R. Yao. Piecewise polynomial regression trees. *Statistica Sinica*, 4, 143-167, 1994.

A. Ciampi. Generalized regression trees. *Computational Statistics and Data Analysis*. 12, 57-78, 1991.

J.P. Collet1, C. Sharpe, E. Belzile1, J.F. Boivin, J. Hanley and L. Abenhaim. Colorectal cancer prevention by non-steroidal anti-inflammatory drugs: effects of dosage and timing. *British Journal of Cancer*, 81, 62-8, 1999.

C. de Boor. *A practical guide to splines*. Springer, New York, 1978.

R.L. Eubank. *Spline smoothing and nonparametric regression*. Dekker, New York, 1988.

D. Feldman, and S. Gross. Mortgage default: Classification Tree Analysis. *The Journal of Real Estate Finance and Economics*, 30, 369-396, 2005.

A. Fielding. Binary segmentation: The automatic interaction detector and related techniques for exploring data structure. *In the analysis of survey data*, 1, ed. C. A. O'Muircheartaigh and C. Payne. Chichester: Wiley, 1977.

J.H. Freidman, and W. Stuetzle. Projection pursuit regression. *Journal of American Statistical Association*. 76, 817-823, 1981.

T. Gausser, and H.-G. Muller. Kernel estimation of regression functions. *Lecture notes in Math*. 757, 23-68, Springer-Verlag, New York, 1979.

T. Gausser, and H. -G. Muller. Estimation regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*. 11, 171-185, 1984.

E. Gilpin, R.A. Olshen, H. Henning and J.Jr. Ross. Risk prediction after myocardial infarction, comparison of three multivariate methodologies. *Cardiology*, 70, 73-84, 1983.

E. Giovannucci, A. Ascherio, E.B. Rimm, G.A. Colditz, M.J. Stampfer and W.C. Willett. Physical activity, Obesity and Risk for Colon Cancer and Adenoma in Men. *Annals of Internal Medicine*, 122, 327-334, 1995.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, et el. *Science*, 286, 531-537, 1999.

M. Grassi, S. Villana, and A. Marinoni. Classification methods for the identification of case in epidemiological diagnosis of asthma. *European Journal of Epidemiology*, 17, 19-29, 2001.

D. Harrison, and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5, 81-102, 1978.

A. Hart. Experience *in the use of inductive system in knowledge engineering*. In M. Bramer, editor, Research and Developments in Expert Systems, Cambridge. Cambridge University press, 1984.

D.M. Hawkins. FIRM (Formal inference based recursive modeling). PC version released 2.1. *Technical Report* 546, School of Statistics, University of Minnesota, 1997.

The International Study of Asthma and Allergies In Childhood (ISAAC Steering Committee). Worldwide variations in the prevalence of asthma symptoms: the International Study of Asthma and Allergies in Childhood (ISAAC). *European Respiratory Journal*, 12, 315-335, 1988.

P.F.B. Jane A. Utis, G. Crippen, H.D. He V. Fischer, R. Tullman, H.Q. Yin, C.P. Hsu, L. Jiang and K.K. Hwang. Use of Classification Regression Tree in Predicting Oral Absorption in Humans. *Journal of Chemical Information and Computer Sciences*, 44, 2061-2069, 2004.

K. M. Kash, J.C. Holland, M.S. Halper, and D.G. Miller. Psychological distress and surveillance behavior of women with a family history of breast cancer. *Journal of National Cancer Institute (JNCI)*, 84, 24-30, 1992.

G.V. Kass. Significance Testing in Automatic Interaction Detection (AID). *Applied Statistics*, 24, 178-189, 1975.

G. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 29, 119-127, 1980.

R. Kothari and M. Dong.(, Decision trees for classification: A review and some new results. (submitted to World Scientific on June 30, 2000).

citeseer.ifi.unizh.ch/479713.html.

W.Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361-386, 2002.

W.Y. Loh, and N. Vanichsetakul. Tree structured classification via generalized discriminant analysis (with discussion). *Journal of American Statistical Association*, 83, 715-728, 1988.

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

J. Mingers. Expert systems, experiments with rule induction. *Journal of the Operation Research Society*, 38, 39-47, 1987.

G.G. Moisen, and T.S. Frescino. Comparing five modeling techniques for predicting forest characteristics. *Ecological Modeling*, 30, 209-225, 2002.

J.N. Morgan and R.C. Messenger. THAID; A sequential analysis program for the analysis of nominal scale dependent variables, *Technical report, Institute for Social Research*, University of Michigan, Ann Arbor, 1973.

J.N. Morgan and J.A. Sonquist. Problems in the analysis of survey data and a proposal. *Journal of American Statistical Association*, 58, 415-434, 1963.

J.D. Potter, M.L. Slattery, R.M. Bostick and S.M. Gapstur. Colon cancer: a review of the epidemiology. *Epidemiologic Reviews*, 15, 499-545, 1993.

J.D. Potter. Methyl Supply, Methyl Metabolizing Enzymes and Colorectal Neoplasia. Trans-HHS Workshop: Diet, DNA Methylation Processes and Health. *The Journal of Nutrition*, The American Society for Nutritional Sciences, 132, 2410S-2412S, 2002.

J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1, 81-106, 1986.

J.R. Quinlan. *Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, California, 1993.

K.F. Rabe, P.A. Vermeire, J.B. Soriano and W.C. Maier. Clinical management of asthma in 1999: the Asthma Insights and Reality in Europe (AIRE) study. *European Respiratory Journal*, 16, 802-807, 2000.

J.O. Ramsay. Monotone regression splines in action (with discussion). *Statistical Science*, 3, 4, 425-441, 1988.

J. Rousu, L. Flander, M. Suutarinen, K. Autio, P. Kontkanen, and A. Rantanen. Nord Computational Tools in Backing Process Data Analysis. A comparative study. *Journal of Food Engineering*, 57, 45-56, 2003.

M.R. Sears. Descriptive epidemiology of asthma. *Lancet*, 350, 1-4, 1997.

D.R. Shopland, H.J. Eyre and T.F. Peachacek. Smoking-attributable cancer mortality in 1991: is lung cancer now the leading cause of death among smokers in the United States? Journal of National Cancer Institute, 83, 1842-1848, 1991.

J.A. Sonquist and J.N. Morgan. The detection of interaction effects. *Ann Arbor Institute for Social Research*, University of Michigan, 1964.

N.A. Van Eck. Statistical analysis and data management highlights of OSIRIS IV. *The American Statistician*, 34, 119-121, 1980.

World Cancer Research Fund. American Institute for Cancer Research Expert Panel (J.D. Potter chair). *Food, Nutrition and the Prevention of Cancer:* A global perspective. Washington, DC: American Institute for Cancer Research, 1997.

M.M. Xion, L. Jin, W. Li, and E. Boerwinkle. *Bio Techniques*, 29, 1264-1270, 2000.

H. Zhang, C.Y. Yu, B. Singer, and M. Xiong. Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 98, 6730-6735, 2001.

Table 1: Summary of 215 heart patients

	Is BP > 91?				
	Yes		No		
	Is Age > 62.5?				
	Yes		No		
	Is ST Present?				
	Yes	No			Total
Died	7	14	2	14	37
Alive	56	14	102	6	178
Total	63	28	104	20	215