*Communications in Applied Analysis* **xx** (200x), no. N, xxx–xxx

# THEORY AND APPLICATIONS OF DECISION TREE WITH STATISTICAL SOFTWARE

CHUNLING CONG[1] AND CHRIS TSOKOS[2]

[1]University of South Florida, Department of Mathematics and Statistics
Tampa, FL 33613 USA.
*E-mail:* ccong@mail.usf.edu

[2]University of South Florida, Department of Mathematic and Statistics
Tampa, FL 33613 USA.
*E-mail:* profcpt@cas.usf.edu

**ABSTRACT.** Recently, the decision tree analysis plays a very significant role in the analysis and modeling of various types of medical data, especially in cancer research. In addition, decision tree analysis has been extensively used in areas in the financial world, for example, loan approval, portfolio management, health & risk assessment, insurance claim evaluation, supply chain management, etc. It is also widely applied in fields such as engineering, forensic examination and biotechnology. The objective of present study is to review the theory behind decision tree analysis and to illustrate its usefulness by applying the subject area to various applications. Furthermore, statistical software information is given to assist scientists in applying decision tree analysis.

**AMS (MOS) Subject Classification.** 91B06.

## 1. INTRODUCTION

A decision tree as a visual and analytical decision support tool is a hierarchical tree structure. Inductive machine learning algorithms are used to learn the decision function stored in the data of the form $(X, Y) = (X_1, X_2, X_3, \dots X_k, Y)$ that maps some sets of attributes $(X_1, X_2, X_3, \dots X_k, Y)$ to the conclusion about some target variable $Y$, and then the target variable $Y$ can be classified or predicted as necessary. The attributes could be any type of variables and based on the type of the outcomes that we are interested in, a decision tree can be called classification tree in descriptive manner if the outcome is discrete or regression tree in a predictive manner if there are continuous outcomes.

The theory of a decision tree has the following main parts: a "root" node is the starting point of the tree; branches connect nodes showing the flow from question to answer. Nodes that have child nodes are called "interior" nodes. "leaf" or "terminal" nodes are nodes that do not have child nodes and represent a possible value of target

Received June 30, 2009

1083-2564 $15.00 ©Dynamic Publishers, Inc.

variable given the variables represented by the path from the root. The following graphs are two examples of decision trees of 320 breast cancer patients who received the medical treatment of tamoxifen and radiation and 321 patients who received tamoxifen alone respectively. The target variable is relapse time, and the attributes are age, hgb, hist, nodediss, hrlevel, and pathsize (will be explained in detail in section 3). As can be seen Figure 1.1 and 1.2, not all attributes are used to split the nodes. The next section explains the mathematical algorithms of how to construct a decision tree including how an attribute and the value of attribute are chosen to split a given node.

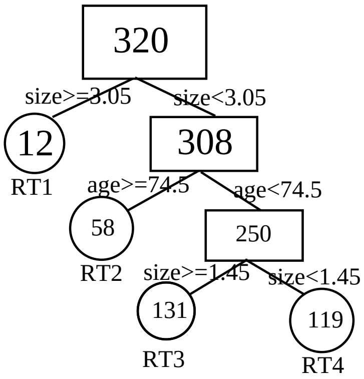

FIGURE 1.1. Radiation+ Tamoxifen

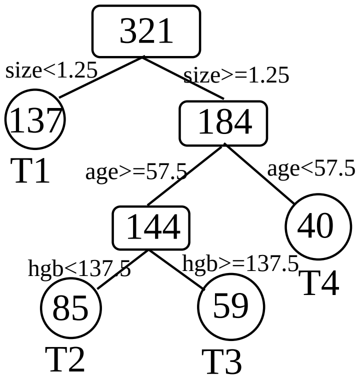

FIGURE 1.2. Tamoxifen

There are several advantages of decision tree over other classification theory tools that make decision tree popular besides its simplicity and interpretability. The approach is supervised learning that given a training data that consist of input and output, we can induce a decision tree even with little hard data; it performs well with large data in a short time, and other statistical or mathematical techniques can be easily incorporated in it.

## 2. THEORY BEHIND DECISION TREE ANALYSIS

The basic idea of decision tree analysis is to spit the given source data set into subsets by recursive portioning of the parent node into child nodes based on the homogeneity of within-node instances or separation of between-node instances with respect to target variables. For each node, attributes are examined and the splitter is chosen to be the attribute such that after dividing the nodes into two child nodes according to the value of the attribute variable, the target variable is differentiated to the best using algorithm. Because of this, we need to be able to distinguish between important attributes, and attributes which contribute little to overall decision process. This process is repeated on each child node in a recursive manner until splitting is either non-feasible or all certain pre-specified stopping rules are satisfied.

## MATHEMATICAL ALGORITHMS

Classification & Regression Tree is a decision tree algorithm (L. Breiman, 1984) [1] is a non-parametric probability distribution free technique to construct binary classification or regression trees as shown in Figure 2.1. Splitting points - attribute variables and values of chosen variables - are chosen based on Gini impurity and Gini gain are given by:

$$i(t) = 1 - \sum_{i=1}^{m} f(t,i)^2 = \sum_{i \neq j} f(t,i) f(t,j)$$

$$\Delta i(s,t) = i(t) - P_L \cdot i(t_L) - P_R \cdot i(t_R),$$

where $f(t,i)$ is the probability of getting $i$ in node $t$ , and the target variable takes values in {1, 2, 3, . . . , m}. $P_L$ is the proportion of cases in node $t$ divided to the left child node and $P_R$ is the proportion of cases in t sent to the right child node. If the target variable is continuous, the split criterion is used with the Least Squares Deviation (LSD) as impurity measure. If there is no Gini gain or the preset stopping rule are satisfied, the splitting process stops.

CHAID (Chi-Squared Automatic Interaction Detection) classification technique introduced by Kass (1980) [2] for nominal predictors and extended by Magidson (1993) [3] to ordinal predictors is another effective approach for nominal or ordinal target variable. CHAID exhausts all possible pairs of categories of the target variable and merge each pair until there is no statistically significant difference within the pair using Chi-square test.

ID.3 (Iterative Dichotomiser 3) developed by Ross Quinlan [4] in 1986 is a classification tree used the concept of information entropy first brought in a publication by Claude Shannon and Warren Weaver (1949) [5]. This provides a method to measure the number of bits each attribute can provide, and the attribute that yields the most information gain becomes the most important attribute and it should go at the top of the tree. Repeat this procedure until all instances in the node are in the same category. As shown in Figure 2.2, it works in the following manner. Suppose there are only two outcomes "Yes" and "No" in the root node $T$ of target variable. Let

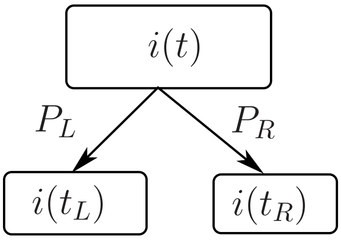

Figure 2.1. CART

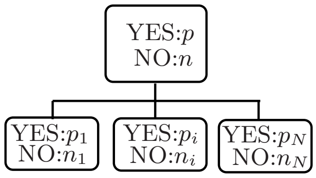

Figure 2.2. ID.3

$p$ and $n$ denotes the number of "positive records and negative records, respectively. The initial information entropy is given by:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n},$$

If attribute $X$ with values $\{x_1,x_2, \ldots,x_N\}$ is chosen to be the split predictor and partition the initial node into $\{T_1,T_2, \ldots,T_N\ \}$, and $p_i$ and $n_i$ denotes the number of positive records and negative records in the child node $i$. Then the expected information $EI(X)$and information gain $G(X)$are given by,

$$EI(X) = \sum_{i=1}^{N}\frac{p_i+n_i}{p+n}\cdot I(p_i,n_i),$$

And information gain $G(X) = I(p,n) - EI(X)$.

In 1993, Ross Quinlan made several improvements to ID.3 and extended it to C4.5 [6]. Unlike ID.3 which deals with discrete attributes, C4.5 handles both continuous and discrete attributes by creating a threshold to split the attribute into two groups, those above the threshold and those that are up to and including the threshold. C4.5 also deals with records that have unknown attribute values. C4.5 algorithm used normalized information gain or gain ratio as a modified splitting criterion of information gain which is the ratio of information gain divided by the information due to the split of a node on the basis of the value of a specific attribute. The reason of this modification is that the information gain tends to favor attributes that have a large number of values.

The best approach in selecting the attribute for a specific node is to choose the one that maximize the given ratio. Stopping rule of C4.5 needs to be pre-specified and it initiated a pruning procedures by replacing branches that do not help with leaf nodes after they are created to decrease overall tree size and the estimated error of the tree. A rule set can be derived from the decision tree constructed by writing a rule for each path from the root node to the leaf node. After C4.5, Quinlan created C5.0 [7] as an extended commercial version of C4.5 featuring a number of improvements including smaller decision trees, weighting different attributes and misclassification types, reducing noise, speed and memory efficiency, support for boosting which gives the trees more accuracy.

As a binary-split algorithm, like CART, QUEST (Quick, Unbiased, Efficient, Statistical Tee) [8] proposed by Loh and Shih in 1997 is a classification algorithm dealing with either categorical or continuous predictor $X$. Pearson's chi-square test is applied to target variable $Y$ and predictor $X$'s independence if $X$ is a categorical predictor. Otherwise, if $X$ is continuous, ANOVA $F$ test is performed to test if all the difference classes of $Y$ have the same mean of $X$. In both cases, $p$-values are calculated and compared to a Bonferroni adjusted threshold to determine if further

Levene's $F$-statistics test needs to be performed to determine if the predictor should be chosen as the split predictor for the node. Overfitting occurs in large tree models where the model fits noise in the data, such as including some attributes that are irrelevant to the decision-making process. If such a model is applied to data other than the training set, the model may not perform well. There are generally two ways to reduce overfitting: stop growing when data is split not statistically significant, or grow full tree, and then post prune. For example, if Gain of the best attribute at a node is below a threshold, stop and make this node a leaf rather than generating children nodes.

## SURVIVAL TREE

A decision tree is of great importance in classification and modeling of health-related data and in many situations the data is censored due to various reasons one of which is that some patients left before the end of the period of study. Due to the incompleteness of the data, a special portioning and pruning algorithm should be used to construct a survival tree. Gordon and Olshen (1985) [9] gave the first adaption of CART algorithm in censored data using Wasserstein metrics to measure distances between Kaplan-Meier [10] curves and certain point masses . After that, Segal (1988) [11] extended regression-tree methodology to right-censored target variables by replacing the splitting rules with between-node separation rules based on the Tarone-Ware [12] or Harrington-Fleming [13] classes of two-sample statistics and a new pruning algorithm was also devised, and truncation and time-dependent covariates were included in the method proposed by Bacchetti and Segal (1995) [14]. Davis and Anderson (1989) [15] used likelihood-ratio test to split nodes under parametric exponential distribution or within-node constant hazard assumptions. LeBlanc and Crowley [16] used martingale residuals for splitting rule assuming a proportional hazards model and also developed an corresponding efficient pruning algorithm, and the model was extended to time-dependent case assuming the survival times are piecewise exponential by Huang, Chen and Soong(1998) [17]. Both Davis and Leblanc algorithms are based on a definition of a within-node homogeneity measure, unlike Segal's algorithm which tried to maximize between-node separation. Su and Fan [18] extended the CART algorithm to multivariate survival data by introducing a gamma distributed frailty to account for the dependence among survival times based on likelihood ratio test as the splitting function. In addition, this method was extended to competing risks based on proportional hazards for subdistribution of competing risks and deviance was used to grow a tree proposed by Ibrahim, Kudus, Daud and Bakar [19].

## RANDOM FOREST

Random forest is an ensemble classifier first developed by Leo Breiman and Adele Cutler [20] in 2001. Random forest has more accuracy than the single-tree model, and handles a very large number of input variables. Besides, it provides an experimental way to detect variable interactions, etc. Instead of using all training data, a random sample of $N$ observations with replacement is chosen to build a tree. In the tree building process, for each node, a random subset of the predictor variables is considered as possible splitters for each node, a predictor excluded from one split is allowed to be used as splitters in the same tree. Repeat the above procedure until a large number of trees are constructed. The average of the predicted value in regression trees are computed as the predicted value and the most frequently predicted category in the classification trees are considered to be the predicted category.

## 3. DECISION TREE IN BREAST CANCER

The survival times or other health-related time measurements, such as relapse time, are major concerns of modeling of medical data, especially cancer data. Due to the fact that the clinicopathological characters of cancer patients are heterogeneous, the survival times are quite different between subgroups of patients based on those heterogeneous clinicopathological characters. Decision tree analysis is a useful tool to homogenize the data by separating the data into different subgroups based on those clinicopathological characters with respect to relapse time of patients in different treatments. In a study of 641 patients at Princess Margaret Hospital are included, of which 320 patients received radiation plus tamoxifen and the other 321 patients received tamoxifen alone as shown in Figure 3.1. The purpose of the study [21] is to compare the relapse time of the two groups to draw a conclusion about the effectiveness of the two different treatments. The attributes recorded are size of tumor (pathsize), histology (hist), hormone receptor level (hrlevel), hemoglobin (hgb), whether axillary node dissection was done (nodediss), and age of the patient

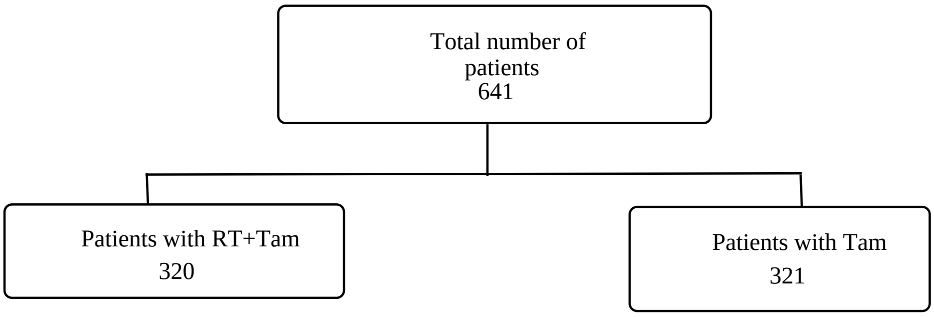

FIGURE 3.1. Breast cancer patients

at diagnosis (age). Although Kaplan-Meier and log rank test can be used to show and compare the overall effectiveness of the two treatment groups, they give no discretion to patients who have heterogeneous clinicopathlogical characteristics which could be of crucial importance on which treatment should be given to a specific breast cancer patient. After using the exponential decision tree algorithm, the two groups are divided into subgroups, as illustrated in Figures 1 and 2. Then it can be shown that for patients with different attributes, their relapse time survival probability could be much different.

## 4. DECISION TREE IN MANAGEMENT AND ENGINEERING

Decision trees are widely used in business management since decision trees give a visual tool of different possible scenarios of investment portfolio and their associated probability, identifying important attributes so that corresponding strategies can be taken to maximize profit. It is also a useful and efficient tool in supply chain management [22]. In this paper, decision tree analysis, C5.0 to be exact, is compared to other two types of model: logistic regression analysis and multivariate determinate analysis and to predict supply chain management (SCM) sustainable collaboration to determine if SCM sustainable collaboration is needed. There are 16 attributes (V1, V2,. . . , V16) from learning perspective, internal process perspective, customer perspective, and financial perspectives. Each of those 16 attributes is assigned values over a 7-point Likert scale. The final decision tree is constructed and the decision making rules are shown in the following table:

| Sustainable collaboration | Non-sustainable Collaboration |
|---|---|
| Rule 1: if V16>5, then 1 | Rule 3: If V16<=5 then 0 |
| Rule 2: If V1>4 and V4<=5 and V8>4, and V11>3 and V15<-5, then 1 | Rule 4: If V1<=4 and V8<=3 and V16<=4, then 0 |

Table 1. SCM sustainable collaboration decision tree

After running the three models, the decision tree model C5.0 proves to perform better than both the logistic model and the multivariate determinate model, which give the same prediction accuracy. It gives the higher prediction accuracy than logistic and multivariate models by 1.02%. Decision tree was also applied in the engineering field [23]. The objective is to appraise the capability of a power system to withstand major contingencies which is part of transient stability assessment (TSA). Decision tree is built and decision rule is deduced to apply them online. ID.3 algorithm is applied to determine the attributes upon which the stability behavior of the steady-state operation points of the training set depends. The stability behavior of a power

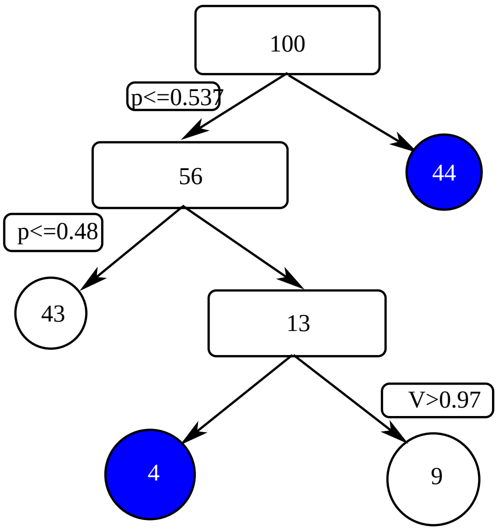

FIGURE 4.1. Decision tree of a one-machine-infinite-bus system

system as the target variable can be divided into classes in which one desires to classify based on the value of critical clearing time. For example, a 2-class partition may use a threshold value of CCT to partition the stability behavior into "stable", and "unstable". A simple "one-machine-busbar" system [24]is illustrated using three attributes Vm [0.9,...,1,1], V[0.9,...,1.1], and P[0.3,...,0.7] . The following decision tree is obtained and could be used to predict the stability of the power system where dark nodes are unstable and the other two terminal nodes are stable.

## 5. DECISION TREE WITH STATISTICAL SOFTWARE

In R [25], package rpart and tree can be used to construct classification, regression and survival trees. Package rpart is recommended for computing CART-like trees. More partitioning algorithms are available in package RWeka which is an interface of implementation of Weka. Detailed information can be found in http://cran.r-project.org/web/packages/rpart/index.html. Package mvpart is an adaption of rpart for multivariate responses available in R. Random forest can be found under randomForest package in R. In addition, random forest variance for response variables measured at arbitrary scales based on conditional inference trees is implemented in package party. Package randomSurvivalForest offers a random forest algorithm for censored data.

SPSS [26] also offers some tree techniques such as CHAID, CART, and QUEST in its Decision Tree dialog box.

## 6. SUMMARY

In the present study, we have reviewed the theory over other machine learning methods of decision tree analysis and emphasized its usefulness in analyzing complicated data. There are several advantages of decision tree over other data mining or machine learning methods including: it performs well with large data in a short time, and it is a white-box model easy to interpret and other statistical or mathematical techniques can be incorporated into it. Real world examples of decision tree analysis in breast cancer, supply chain management and power system are given to illustrate the extensive applications of it. Finally, we provided references of how these analyses could be done via statistical software so individuals interested in performing decision tree analysis can use.

## REFERENCES

[1] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984): *Classification and Regression Trees*, New York; Chapman and Hall.

[2] Kass, G. (1980): An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, **29**, 119–127.

[3] Magidson, J. (1993): The use of the new ordinal algorithm in CHAID to targetprofitable segments, *The Journal of Database Marketing*, **1**, 29–48.

[4] Quinlan, J. R. (1986): Induction of Decision Trees. *Machine Learning* **1**, 1 (Mar. 1986), 81-106.

[5] Claude Shannon and Warren Weaver in their publication *model of communication* in 1949

[6] Quinlan, J. R. (1993): *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, California ISBN 1-55860-238-0.

[7] Quinlan, J. R.: C5.0: An Informal Tutorial, http://www.rulequest.com/see5-unix.html.

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

[9] Gordon, L. and Olshen, R. A. (1985): Tree-structured survival analysis. *Cancer Treatment Reports* **69**, 1065-1069.

[10] Kaplan, E.L.; Meier, Paul. (1958): Nonparametric estimation from incomplete observations, *J. Am. Stat. Assoc.* **53**, 457-481.

[11] Segal M. R. (1988): Regression trees for censored data, *Biometrics* **44**, pp.35-47.

[12] Tarone, R. E. and Ware, J. (1977): On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-160.

[13] Harrington, D. P. and Fleming, T. R. (1982): A class of rank test procedures for censored survival data. *Biometrika* **69**, 553-566.

[14] Bacchetti, P. and Segal M. R. (1995): Survival trees with time-dependent covariates: *application to estimating changes in the incubation period of AIDS Lifetime Data Analysis* Vol. **1**, number1.

[15] Davis, R. and Anderson, J. (1989): Exponential survival trees, *Statistics in Medicine* **8**, pp 947-962.

[16] Lebalanc, M.; Crowlry, L. (1992): Relative risk trees for censored survival data, *Biometrics.* v**48**. 411-425.

[17] Lebalanc, M.; Crowlry, L. (1993): Survival trees by goodness of split, *Journal of the American Statistical Association.* v**88**. 457-467.

[18] Su, X. G.; Fan, J. J. (2004): Multivariate survival trees: a maximum likelihood approach based on frailty models, *Biometrics* **60**, pp. 93-99.

[19] N.A Ibrahim, et al. (2008): Decision tree for competing risks survival probability in breast cancer study, *International Journal of Biomedical Sciences* Volume **3** Number 1.

[20] Breiman, Leo (2001): Random Forests, *Machine learning*, **45** 1: 5–32.

[21] Cong, Chunling; Tsokos, C. P.: parametric and nonparametric analysis of breast cancer treatments.(2009): Submited to *International Journal of Biomedical Sciences.*

[22] Lim, Se Hun (2006): The design of controls in supply chain management sustainable collaboration using decision tree algorithm, *International Journal of Computer Science and Network Security*, Vol. **6** No.5A.

[23] Wehenkel, L.; Cutsem, T.Van, and Ribbens-Pavella, M. (1989): An artificial intelligence framework for online transient stability assessment of powersystems, *IEEE Trans. Power Syst.*, vol. **4**, no. 2, pp. 789–800.

[24] Wehenkel, L.; Cutsem, T.Van, and Ribbens-Pavella, M. (1987): Artificial intelligence applied to on-line transient stability assessment of electric power systems, *Proc. of the 10th IFAC LVorlcl Congress*, pp. **308-313,** Munich.

[25] www.r-project.org.

[26] www.spss.com.