International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:2, No:2, 2008

Decision Tree for Competing Risks Survival Probability in Breast Cancer Study

N. A. Ibrahim, A. Kudus, I. Daud, and M. R. Abu Bakar

***Abstract*—**Competing risks survival data that comprises of more than one type of event has been used in many applications, and one of these is in clinical study (e.g. in breast cancer study). The decision tree method can be extended to competing risks survival data by modifying the split function so as to accommodate two or more risks which might be dependent on each other. Recently, researchers have constructed some decision trees for recurrent survival time data using frailty and marginal modelling. We further extended the method for the case of competing risks. In this paper, we developed the decision tree method for competing risks survival time data based on proportional hazards for subdistribution of competing risks. In particular, we grow a tree by using deviance statistic. The application of breast cancer data is presented. Finally, to investigate the performance of the proposed method, simulation studies on identification of true group of observations were executed.

***Keywords*—**Competing risks, Decision tree, Simulation, Subdistribution Proportional Hazard.

I. Introduction

A huge amount of data has been rapidly accumulated, due to the fast development of computer technology. A new data analysis problem has arisen in such situation. Data mining is used to find important "knowledge" from large databases. Decision tree as one of many data mining techniques has become a popular approach for segmentation, classification and prediction by applying a series of simple rules. The advantage that researchers have is that the results can be understood and explained easily, since it is expressed by a tree structured diagram as a final output. Decision tree automatically constructed from data have been used successfully in many real-world situations. Their effectiveness has been compared widely to other automated data exploration methods and to human experts. Decision tree can be an one of the important tools in data mining and can provide useful insight of the data being analysed.

The landmark work of a decision tree is the Classification and Regression Trees (CART) methodology of [1], who introduced a tree methodology for univariate discrete or continuous response. A different approach was C4.5 proposed by [2].

There is now quite a lot of work dealing with decision tree especially in survival analysis ([3] – [13]), but there are no decision tree methods for competing risks survival data. Since analysis of competing risks survival data is complex due to the presence of more than one cause of failure, then it should be a useful method to develop.

In this paper, we extended the decision tree for competing risks survival time data analysis by utilizing the proportional hazards model for subdistribution. In the application part, breast cancer data was studied where there were several events that might occur following the treatment after breast-conserving surgery [14].

II. Regression Tree for Competing Risks

CART is not directly applicable to survival and competing risks data because of the censored observations. Additionally, the major focus in survival analysis is on the survival (distribution) or hazard function rather than the mean function. In extending tree based techniques to cope with univariate or independent failure times, some approaches allow the direct use of the CART procedure by defining appropriate prediction error terms, while the others have made modifications to CART in an effort to overcome the difficulties naturally associated with censored failure times.

In the development of tree for competing risks, we used deviance for the between-node difference. Deviance is derived from likelihood ratio test statistic of proportional hazards model for subdistribution developed by [15]. Therefore, only internal nodes have associated deviance statistics. The tree structure is different from CART because, for original tree, each node, either terminal or internal, has an associated impurity measure. This is why the CART pruning procedure is not directly applicable to such type of tree. However, Segal's pruning algorithm [4] which exerts little computational burden, has resulted in tree that have become a well-developed tool.

Manuscript received January 19, 2008. This work was supported by Malaysian Ministry of Science, Technology and Innovation (MOSTI) UNDER grant IRPA.

N. A. Ibrahim is with the Institute for Mathematical Research and Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia (phone: 603-89466873; fax: 603-89423789; e-mail: nakma@putra.upm.edu.my).

A. Kudus is a post-graduate student pursuing his Ph.d Degree at the Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia (e-mail: akudus@yahoo.com).

I. Daud is with the Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia (e-mail: drisa@science.upm.edu.my).

M. R. Abu Bakar is with the Department of Mathematics, Faculty of Science and Associate Researcher at the Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia (e-mail: mrizam@science.upm.edu.my).

We consider a typical setting for competing risks survival data. Suppose that there are n individuals and each subjected to J ($J \geq 2$) types of event. Let T_i^* be the time when ith unit experiences one of the jth type of event, and let C_i be the corresponding censoring time, where $j = 1, 2, \ldots, J$; $i = 1, 2, \ldots, n$. The sample consists of the set of vectors $\{(T_i,\delta_i,Z_i) : i = 1, 2, ..., n\}$. Here, $T_i = \min(T_i^*, C_i)$ are the observed failure times; $\delta_i = I(T_i^* < C_i)$, where $I(\bullet)$ is an indicator function; $Z_i \in \Re^p$ denotes the covariate vector for the ith unit. Since recursive partitioning handles covariates one by one, we assume $p = 1$ for the ease of illustration. In order to ensure identifiability, we also assume that the failure time T_i^* is independent of the censoring time C_i conditional on the covariate Z_i, for any $i = 1, \ldots, n$.

In the subdistribution approach by [15], the subdistribution hazard for each type of failure is formulated with the proportional hazards model. Since we only consider splitting on a single covariate, the subdistribution hazard function $\tilde{\lambda}(t)$ is assumed to take the following form:

$$\tilde{\lambda}_j(t; Z_i) = \tilde{\lambda}_{j0}(t)\exp\left[\beta_\gamma^j I(Z_i < \gamma)\right], \; j = 1,..., J \tag{1}$$

where $\tilde{\lambda}_{j0}(t)$ is an unspecified baseline subdistribution hazard function and β_γ^j is an unknown regression parameter corresponding to cutpoint γ and type of event j. We assume that there is a change point effect of Z_i on the subdistribution hazard function with cutpoint γ.

When there is no censoring, β_γ^j can be estimated in exactly the same way as in the Cox model for right-censored data using a modified risk set. Here the risk set, $R(t)$, at time t is all individuals yet to experience any event plus all those individuals who experienced types other than jth event at a time prior to t. The risk set leads to a partial likelihood:

$$L(\beta_\gamma^j) = \prod_{i=1}^{n} \left(\frac{\exp\left(\beta_\gamma^j I(Z_i < \gamma)\right)}{\sum_{k \in R(t_i)} \exp\left(\beta_\gamma^j I(Z_k < \gamma)\right)} \right)^{I(\delta_i = j)} \tag{2}$$

The log partial likelihood is

$$l(\beta_\gamma^j) = \sum_{i=1}^{n} I(\delta_i = j)\left(\beta_\gamma^j I(Z_i < \gamma) - \log \sum_{k \in R(t_i)} \exp\left(\beta_\gamma^j I(Z_i < \gamma)\right)\right) \tag{3}$$

For the case where competing risks data contained censored observations, the score function is constructed by using weight developed by inverse probability of censoring weighting technique. Value of β_γ^j that solve the score function is the desired estimators.

Given the estimated β_γ^j, the deviance is defined as $-2l\left(\hat{\beta}_\gamma^j\right)$. This is the summary measure of agreement between model and the data, where the smaller value corresponds to better goodness of fit [16].

The splitting function is defined in term of deviance as $R(\gamma, h) = -2l\left(\hat{\beta}_\gamma^j\right)$. This statistic can be derived from likelihood ratio for testing the significance of β_γ^j in which $\hat{\beta}_\gamma^j$ is its maximum likelihood estimator.

In summary, when a tree is constructed, a proportional subdistribution hazard structure is assumed within each node. The splitting function $R(\gamma, h)$ is evaluated at each allowable split, and the best cutpoint γ^* is chosen such that $R(\gamma^*, h) = \min_{\gamma \in h} R(\gamma, h)$. This process is applied recursively until all the nodes cannot be split further.

Algorithm to Grow Tree

To grow a tree, the deviance statistic is evaluated for every possible binary split of the predictor space Z. The split, s, could be of several forms: splits on a single covariate, splits on linear combinations of predictors, and boolean combination of splits. The simplest form, in which each split relates to only one covariate, can be described as follows:

1. If Z_k is ordered, then the data will be split into two groups specified by $\{Z_k < \gamma\}$ and $\{Z_k \geq \gamma\}$ respectively;
2. If Z_k is nominal, then any subset of the possible nominal values could induce a split.

The "best split" is defined to be the one corresponding to the minimum deviance statistic. Subsequently the data are divided into two groups according to the best split.

Apply this splitting scheme recursively to the sample until the predictor space is partitioned into many regions. There will be no further partition to a node when any of the following occurs:

1. The node contains less than, say 10 or 20, observations.
2. All the observed times in the subset are censored.
3. All the observations have identical covariate vectors or the node has only complete observations with identical survival times.

This procedure results in a large tree Γ_0, which could be used for the purpose of data structure exploration.

Algorithm to Prune Tree

The idea of pruning is to iteratively cut off branches of the initial tree, Γ_0, in order to locate a limited number of candidate subtrees from which an optimally sized tree is selected. For the proposed method, we adopt Segal's pruning algorithm (Segal, 1988) which exerts little computational burden. The steps for adopting this algorithm are as follows:

1. Initially growing a large tree.

2. To each of the internal nodes in the original tree, assign the maximal splitting statistics contained in the corresponding branch. This statistic reflects strength of linking for the branch to the tree.
3. Among all these internal nodes, finds the one with the smallest statistic. That is, find the branch that has the weakest link and then prune off this branch from the tree.
4. The second pruned tree can be obtained in a similar manner by applying the above two steps to the first pruned tree.
5. Repeating this process until the pruned tree contains only the root node, finally a sequence of nested trees is obtained.

The desired tree can be obtained by plotting the size of these trees against their weakest linking statistics. The tree corresponding to the "kink" point in the curve is chosen as the best one.

III. Illustration

As an application of competing risks tree, we used breast cancer data from Fyles *et al*. (2004). Between December 1992 and June 2000, 639 women 50 years of age or older who had undergone breast-conserving surgery for an invasive adenocarcinoma 5 cm or less in diameter (pathological stage T_1 or T_2) were randomly assigned to receive breast irradiation plus tamoxifen, RT+Tam, (319 women) or tamoxifen alone, Tam, (320 women). Participating centers included the Princess Margaret Hospital, the Women's College Campus of the Sunnybrook and Women's College Health Science Centre in Toronto, and the British Columbia Cancer Agency centers in Vancouver and Victoria. Table I contains the list of variables and their description.

TABLE I
Description of Variables in the Breast Cancer Study

Variable name	Description
tx	Randomized treatment: 0=tamoxifen, 1=radiation+tamoxifen
Variable assessed at the time of randomization	
pathsize	Size of the tumour (cm)
hist	Histology: 1=ductal, 2=lobular,3=medullary, 4=mixed, 5=other
hrlevel	Hormone receptor level: 0=negative, 1=positive
hgb	Haemoglobin (g/l)
nodedis	Whether axillary node dissection was done, 0=Yes, 1=No
age	Age (years)
Outcome variables	
time	Time from randomization to event (relapse, second malignancy or death) or last follow up (years)
d	Status at last follow-up: 0=censored, 1=relapse, 2=malignancy, 3=death

The events that might occur in breast cancer study were relapse, second malignancy and death. The patient's survival time was the time length between the date of randomization and the occurrence of one event or last follow-up date.

Since the goal of regression tree is to partition patients into groups on the basis of similarity of their responses to treatment, we constructed a separated regression tree for each treatments (tamoxifen alone and tamoxifen plus radiation). The partitioning was based on baseline characteristics such as patient demographics and clinical measurements. The final tree structure provides treatment effect within each group of patients. The question to be answered by this type of analysis is – for whom does the treatment work best?

The cumulative probability for relapse by time t is shown in Fig. 1. Here we compared the probability for two types of treatment. The patients with tamoxifen plus radiation have less probability to relapse compared to those with tamoxifen alone as expected. It showed the advantage of radiation in reducing the occurrence of relapse.

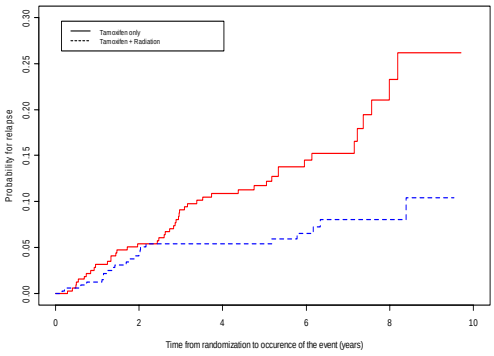

Fig. 1 Cumulative Probability for relapse for two types of treatment

The exploration was further executed to find group of patients for each treatment by using decision tree. With respect to probability for relapse, we obtained four groups of patients which were treated by tamoxifen alone, and three groups of patients which were treated by tamoxifen plus radiation (Fig. 2).

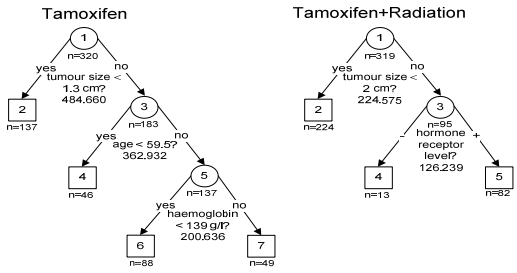

Fig. 2 Decision tree for probability for relapse

The description of four groups of patient which was treated by tamoxifen alone is:

1. Node 2: tumour size < 1.3 cm
2. Node 4: tumour size ≥ 1.3 cm and age < 59.5 years
3. Node 6: tumour size ≥ 1.3 cm and age ≥ 59.5 years and haemoglobin < 139 g/l
4. Node 7: tumour size ≥ 1.3 cm and age ≥ 59.5 years and

haemoglobin ≥ 139 g/l

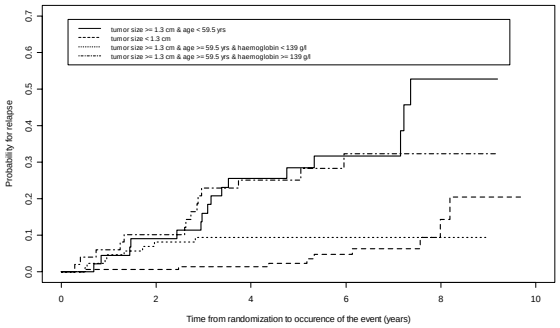

Fig. 3 Cumulative Probability for relapse for groups resulted by decision tree for patient with tamoxifen alone

For patients with tamoxifen plus radiation, there were three groups resulted by decision tree, namely:

1. Node 2: tumour size < 2 cm
2. Node 4: tumour size ≥ 2 cm and hormone receptor level negative
3. Node 5: tumour size ≥ 2 cm and hormone receptor level positive

Women with tamoxifen plus radiation whose tumour size less than 2 cm have the lowest probability to relapse, whereas the highest probability is for those whose tumour size ≥ 2cm and negative hormone receptor level (Fig. 4).

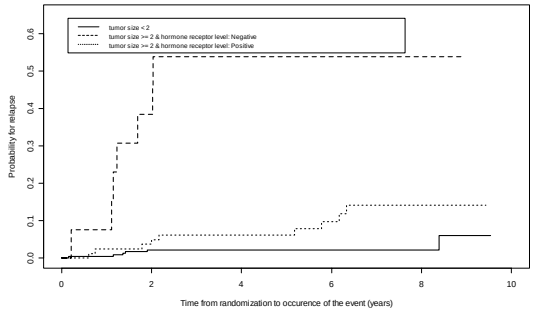

Fig. 4 Cumulative Probability for relapse for groups resulted by decision tree for patient with tamoxifen plus radiation

Overall comparison for both treatments reveals that the poorest and best prognosis are from tamoxifen plus radiation treatment group. We found that tamoxifen plus radiation is not effective for those women whose tumour size greater than 2 cm and negative hormone receptor level. This group is more likely to relapse compared to the others. On the other hand, patients with tamoxifen plus radiation and tumor size less than 2 cm have the best prognosis, because they are less likely to relapse (Fig. 5).

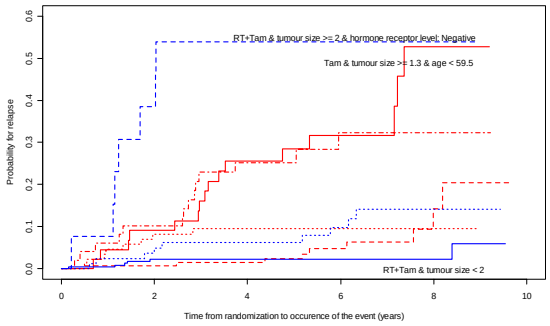

Fig. 5 Cumulative Probability for relapse for all groups resulted by decision tree

The analysis for two other events (second malignancy and death) showed different results. This showed that patients give different responses to the treatment.

IV. Simulation Studies

Several simulation studies were conducted to investigate the performance of the proposed method, since it is difficult to assess regression method analytically. For simplicity, we considered competing risks data with two types of event (j=1,2) and their survival times might be censored.

The true population model consists of four groups of observations determined by their covariate values. Those four groups are node 4, node 5, node 6 and node 7 of tree Γ_1 or Γ_2 (Fig. 6).

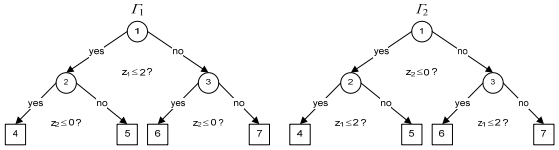

Fig. 6 True tree for simulation

In addition to the observed data, censoring times were also introduced independently from a uniform distribution to obtain approximately 23%, 47% and 71% censoring [16]. Each simulation had 1000 repetitions and each repetition consist of 400 competing risks survival time data.

Simulations were summarized in terms of capability in identifying prognostic groups. A tree will be classified as "Identified" if all 4 prognostic groups or parts of true tree are correctly identified. Otherwise, the tree will be classified as "Not identified". Table II summarizes the simulations. The results show that the proposed method performs quite well in the identification of correct data structure. However, the performance decreases as the censoring percentage increases.

TABLE II
Evaluation of Identifying Data Structures

Censoring (%)	Identified	Not identified
0	99.4%	0.6%
23	98.3%	1.7%
47	94.1%	5.9%
71	62.4%	37.6%

V. Conclusion

In this article, we have proposed a competing risks tree method based on proportional hazards of subdistribution model. The proposed method intends to provide an exploratory data analysis for competing risks survival data, and it is complimentary rather than competitive to those parametric or semi-parametric methods. The application on breast cancer data showed that the method could find groups of observations which had similar response to treatment. Simulation results showed that the proposed method performs well for prognostic classification. In all of the simulations, high portion of data structures can be correctly identified.

References

[1] L. Breiman, J. Friedman, R. Olshen and C. Stone, "Classification and regression trees", New York: Chapman and Hall, 1984.
[2] J. R. Quinlan, "C4.5: Program for Machine Learning", 1992, California: Morgan Kaufmann.
[3] L. Gordon, and R. Olshen, "Tree-structured survival analysis", 1985, Cancer Treatment Reports 69, pp. 1065-1069.
[4] M. R. Segal, "Regression trees for censored data", 1988, Biometrics 44, pp. 35-47.
[5] R. Davis and J. Anderson, "Exponential survival trees", 1989, Statistics in Medicine 8, pp. 947-962.
[6] M. LeBlanc, and J. Crowley, "Relative risk trees for censored survival data", 1992, Biometrics 48, pp. 411-425.
[7] M. LeBlanc, and J. Crowley, "Survival trees by goodness of split", 1993, Journal of the American Statistical Association 88, pp. 457-467.
[8] M. R. Segal, "Extending the elements of tree-structured regression", Statist. Methods Med. Res. 4, pp. 219–236.
[9] X. Huang, S. Chen, and S. Soong, "Piecewise exponential survival trees with time-dependent covariates", 1998, Biometrics 54, pp. 1420-1433.
[10] M. R. Segal, "Tree-structured method for longitudinal data", 1992, Journal of the American Statistical Association 87, pp. 407-418.
[11] H. P. Zhang, "Classification tree for multiple binary responses", 1998, Journal of the American Statistical Association 93, pp. 180-193.
[12] X. G. Su and J.J. Fan, "Multivariate survival trees: a maximum likelihood approach based on frailty models", Biometrics 60, pp. 93-99.
[13] F. Gao, A. K. Manatunga, and S. Chen, "Identification of prognostic factors with multivariate survival data", 2004, Computational Statistics and Data Analysis 45, pp. 813-824.
[14] A. W. Fyles, D. R. McCready, L. A Manchul., M. E. Trudeau, P. Merante, M. Pintilie, L. M. Weir, and I. A. Olivotto, "Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer", 2004, New England Journal of Medicine 351, pp. 963-970.
[15] J. P. Fine and R. J. Gray , "A proportional hazards model for the subdistribution of a competing risk", 1999, Journal of the American Statistical Association 94, pp. 496-509.
[16] D. Collett, "Modelling survival data in medical research", London: Chapman and Hall, 1994.