Dipolar regression trees in survival analysis

Małgorzata Krętowska
Białystok Technical University, Faculty of Computer Science,
Wiejska 45a, Białystok, Poland
e-mail: *mmac@ii.pb.bialystok.pl*

Summary: In this paper a new method for induction of multivariate regression trees is presented. The technique is designed for the survival time prediction and based on given data. The proposed method aims at identification of subgroups of patients with homogenous survival experience i.e. homogeneous response for a given treatment. The method allows using information from censored cases for which the exact failure time is unknown. An appropriate degree of generalization is obtained by using a pruning algorithm, which is based on rank correlation coefficient D.

Keywords: survival analysis, dipolar criterion function, regression tree.

1. Introduction

Construction of the prediction tools can have two purposes: to predict the response variable corresponding to the future measurement vectors as accurately as possible and to understand the structural relationships between the response and the measured variables [1]. In case of the survival analysis, the respond variable is related to the time of the failure occurrence. Since the structure and the quality of the prognostic tools is closely connected with the quality of the analyzed data and the form in which the data is presented, special attention should be paid to the nature of the survival data. The main problem that appears while analyzing such kind of data is censoring. Censored observations include incomplete information about the failure time. We only know that the failure time for censored cases is not less than their follow-up time. Ignoring such cases during development of the prognostic tool is equivalent to rejection of some knowledge of the analyzed phenomenon and may bias the outcome.

In the paper, tree-based models are used as the prediction tool. Such models are widely used in medical applications: in classification and regression [2]. Their usage in the survival analysis is aimed at identification of subgroups that are homogeneous in their survival experience. Marubini *et. al.* [3] proposed an approach based on the Cox's proportional hazards model and partial likelihood approach. A similar method was developed by Ciampi *et. al.* [4]. Davis and Anderson [5] assumed an exponential model for the survival distribution and exploited the exponential log-likelihood as a goodness-of–split criterion. Other techniques were based on the Cox's model [6] or on tests for difference between survival curves [7]. In general the approaches were limited to the univariate trees that used only one feature in each internal node and partition feature space with the axis-parallel hyper-planes.

In the paper, a new algorithm for induction of the multivariate regression tree for the survival time prediction is presented. The proposed method aims at dividing the feature space into such areas which would include the patients with similar survival time. The splits connected with the internal nodes of multivariate tree take on a form of a hyper-plane. Searching for optimal hyper-plane is based on minimization of dipolar criterion function that is closely related to the form of the analyzed data set. The information included in the data is shown in the form of dipoles – pairs of feature vectors. Such a representation allows using the information from both uncensored and censored cases. An appropriate generalization degree can be received by using the build-in feature selection method and the pruning algorithm. The feature selection technique, which is a part of the dipolar criterion function minimization method, allows omitting the attributes that are unimportant from the point of view of the survival time and, on the other hand, distinguishes the attributes that are significant for the analyzed problem. The proposed pruning algorithm uses the rank correlation coefficient for assessing the quality of sub-trees.

The experimental evaluation of the method was carried out on the base of two real data sets. The first data set contained the feature vectors describing the patients with primary biliary cirrhosis (PBC) of the liver [8], the other included the information on the patients with malignant melanoma [9].

The paper is arranged as follows. Section 2 introduces the concept of dipole and describes how to create the dipolar criterion function. The induction of the regression tree together with the pruning algorithm is presented in Section 3. Section 4 includes experimental results for two data sets, and, finally, Section 5 summarises the results.

2. Dipolar Criterion Function

The *i*-th observation (patient) in the survival data is described by set ($\boldsymbol{x}_i$, t_i, δ_i) $i=1,2,\ldots,M$, where $\boldsymbol{x}_i$ is *N*-dimensional covariate vector, t_i – survival time, and δ_i – failure indicator. Thew failure indicator is equal to 1 for patients for whom the event of interest is noted (uncensored cases) and 0 otherwise (censored cases).

The aim of the analysis is to divide the patients into subgroups homogeneous from the point of view of the survival experience. Each subgroup should be connected with a distinct terminal node. The main issue in tree induction algorithms is finding an appropriate split in each internal node. In the presented approach minimization of the dipolar criterion function is used for this purpose.

The dipolar criterion function is based on the concept of dipoles. The dipole [10] is a pair of different covariate vectors ($\boldsymbol{x}_i$, $\boldsymbol{x}_j$) from the learning set. Mixed and pure dipoles are distinguished. Mixed dipoles are formed between objects that should be separated, while pure dipoles between objects that are similar from the point of view of the analyzed criterion. In case of the survival data, pure dipoles are created between pairs of the feature vectors for which the difference of the failure times is small and mixed dipoles, between pairs with distant failure times. Taking into account the censored cases, the following rules of the dipole construction can be formulated:

a) a pair of feature vectors ($\boldsymbol{x}_i$, $\boldsymbol{x}_j$) forms the pure dipole, if
   - $\delta_i = \delta_j = 1$ and $| t_i - t_j | < \eta$;

b) a pair of feature vectors ($\boldsymbol{x}_i$, $\boldsymbol{x}_j$) forms the mixed dipole, if
   - $\delta_i = \delta_j = 1$ and $| t_i - t_j | > \zeta$;
   - ($\delta_i = 0$, $\delta_j = 1$ and $t_i - t_j > \zeta$) or ($\delta_i = 1$, $\delta_j = 0$ and $t_j - t_i > \zeta$).

Parameters η and ζ are equal to quartiles of the absolute values of differences between the uncensored survival times.

The aim is to find such a hyper-plane *H(v)* that would divide a possibly high number of mixed dipoles and possibly low number of pure ones. It is done by minimization of the dipolar criterion function. Two types of piece-wise linear and convex (CPL) penalty functions $\varphi_i^+(\boldsymbol{v})$ and $\varphi_j^-(\boldsymbol{v})$ are considered:

$$\varphi_j^+(\mathbf{v}) = \begin{cases} \delta^j - < \mathbf{v},\mathbf{y_j} > & if < \mathbf{v},\mathbf{y_j} >< \delta^j \\ 0 & if < \mathbf{v},\mathbf{y_j} > \geq \delta^j \end{cases} \tag{1}$$

$$\varphi_j^-(\mathbf{v}) = \begin{cases} \delta^j + < \mathbf{v},\mathbf{y_j} > & if < \mathbf{v},\mathbf{y_j} >> -\delta^j \\ 0 & if < \mathbf{v},\mathbf{y_j} > \leq -\delta^j \end{cases} \tag{2}$$

where $\boldsymbol{y} = [1, \boldsymbol{x}^{\mathrm{T}}]^{\mathrm{T}}$ is an augmented covariate vector, $\boldsymbol{v} = [-\theta, w_1, w_2, ..., w_N]^{\mathrm{T}}$ is an augmented weight vector and δ^{j} is a margin. Each mixed dipole ($\boldsymbol{y}_i$, $\boldsymbol{y}_j$), which should be divided, is associated with a function being a sum of two function with opposite signs

$$\varphi_{ij}^{m}(\boldsymbol{v}) = \varphi_i^{-}(\boldsymbol{v}) + \varphi_j^{+}(\boldsymbol{v}) \text{ or } \varphi_{ij}^{m}(\boldsymbol{v}) = \varphi_i^{+}(\boldsymbol{v}) + \varphi_j^{-}(\boldsymbol{v}) \tag{3}$$

With pure dipoles that should remain undivided we associate function:

$$\varphi_{ij}^{p}(\boldsymbol{v}) = \varphi_i^{-}(\boldsymbol{v}) + \varphi_j^{-}(\boldsymbol{v}) \text{ or } \varphi_{ij}^{p}(\boldsymbol{v}) = \varphi_i^{+}(\boldsymbol{v}) + \varphi_j^{+}(\boldsymbol{v}) \tag{4}$$

The dipolar criterion function is a sum of the penalty functions associated with each dipole:

$$\Psi(\mathbf{v}) = \sum_{(i,j)\in I_p} \alpha_{ij}\varphi_{ij}^{p}(\mathbf{v}) + \sum_{(i,j)\in I_m} \alpha_{ij}\varphi_{ij}^{m}(\mathbf{v}) \tag{5}$$

where α_{ij} determines the relative importance (price) of the dipole ($\boldsymbol{y}_i$, $\boldsymbol{y}_j$), I_p and I_m are sets of pure and mixed dipoles, respectively.
Because $\Psi(\boldsymbol{v})$ is a convex, piece-wise linear function, basis exchange algorithms [11], similar to linear programming, are used as the minimization method.

3. Induction of survival tree

Most of the existing tree induction systems proceed in a greedy, top-down fashion. At each non-terminal node, starting from the root, the best split is learned by using the criterion of optimality. The learned decision function divides the training subset into two (or more) subsets generating child nodes. The process is repeated at each newly created child node until a stopping condition is satisfied, and the node is declared the terminal node.

The binary tree is taken into consideration [12]. At the i-th non-terminal node, set S_i of the input vectors is divided into two subsets S_i^L and S_i^R by hyper-plane $H(\boldsymbol{v}_i)$:

$$S_i^L = \{\boldsymbol{y}: \boldsymbol{y}\in S_i \text{ and } <\boldsymbol{v}_i, \boldsymbol{y}> \ \geq 0\} \tag{6}$$

$$S_i^R = \{\boldsymbol{y}: \boldsymbol{y}\in S_i \text{ and } <\boldsymbol{v}_i, \boldsymbol{y}> < 0\} \tag{7}$$

All the feature vectors from set S_i^L constitute the left child of the i-th node, and S_i^R the right one. To find the optimal dividing hyper-plane $H(\boldsymbol{v}_i)$ dipolar criterion function $\Psi(\boldsymbol{v}_i)$ is used. The function is built from dipoles created from the feature vectors $\boldsymbol{y}\in S_i$. The survival tree, inducted in such a way, separates patients (feature vectors) with the distant failure times. Patients with similar failure times reach the same terminal node. Induction of the survival tree is stopped if one of these conditions is fulfilled: 1) all the mixed dipoles are divided; 2) set S_i consists of less than 5 uncensored cases. The tree, received in such a manner, may be overfitted. Improvement of the generalizing ability may be received by using a pruning technique, which decreases the tree structure by changing some of the internal nodes into leaves. The pruning algorithm proposed in the paper uses modified rank correlation coefficient D [13] to assess the quality of the tree:

$$D = \frac{no.of\ concordant\ pairs - no.of\ discordant\ pairs}{no.of\ pairs - no.of\ ties} \tag{8}$$

where:

$$no.of\ pairs = \sum_{i=1}^{M} \sum_{j=i+1}^{M} I(t_i < t_j)\delta_i + I(t_i > t_j)\delta_j$$

$$no.of\ concordant\ pairs = \sum_{i=1}^{M} \sum_{j=i+1}^{M} I(t_i < t_j)I(p_i < p_j)\delta_i + I(t_i > t_j)I(p_i > p_j)\delta_j$$

$$no.of\ discordant\ pairs = \sum_{i=1}^{M} \sum_{j=i+1}^{M} I(t_i < t_j)I(p_i > p_j)\delta_i + I(t_i > t_j)I(p_i < p_j)\delta_j$$

$$no.of\ ties = \sum_{i=1}^{M} \sum_{j=i+1}^{M} I(t_i < t_j)I(p_i = p_j)\delta_i + I(t_i > t_j)I(p_i = p_j)\delta_j$$

In the equations t_i is the observed survival time, δ_i – failure indicator and p_i is the predicted survival time. The indicator function *I*(*condition*) equals 1 if the *condition* is fulfilled, and 0 otherwise.

The pruning algorithm is conducted according to the following steps:

1. Calculate correlation coefficient *D* for the whole tree *T*.
2. Determine a group of sub-trees *T*(*j*), *j*=1,2,…,*K*, in which one of the internal nodes (with two descendants-leaves) was changed into a leave.
3. Calculate correlation coefficients *D*(*j*) for each subtree *T*(*j*) *j*=1,2,…,*K*
4. Determine index *j'*: *D*(*j'*) = max *D*(*j*), *j*=1,2,…, *K*
5. If *D*(*j'*) ≥ *D* than *T*=*T*(*j'*); go to step 1)
6. If *D*(*j'*) < *D* than STOP.

4. Experimental results

The experimental evaluation of the method was carried out on the basis of two real data sets. The first set contains the data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver. 312 patients participated in the randomized controlled trial of the drug D-penicillamine conducted between 1974 and 1984. Each patient is described by 17 features: drug (D-penicillamine, placebo), age, sex, presence of asictes, presence of hepatomegaly, presence of spiders, presence of edema, serum bilirubin in mg/dl, serum cholesterol in mg/dl, albumin in gm/dl, urine copper in ug/day, alkaline phosphatase in U/liter, SGOT in U/ml, triglicerides in mg/dl, platelets per cubic [ml/1000], prothrombin time in seconds, histologic stage of disease. For each patient the number of days between the registration and the event (death, liver transplantation, study analysis), which occurred earlier is also given.

The other data set contains the information on 205 patients (14 censored cases) with malignant melanoma following radical operation. The data was collected at Odense University Hospital in Denmark by K.T. Drzewiecki. Each patient is described by 4 features: sex, age, tumor thickness [cm] and ulceration.

During the induction of regression trees parameter η was fixed as 0.2 quartile and ζ as 0.6 quartile. The data sets were divided into two subsets which consisting of 2/3 and 1/3 elements. The bigger part was used for induction of the preliminary tree structure, the other for pruning.

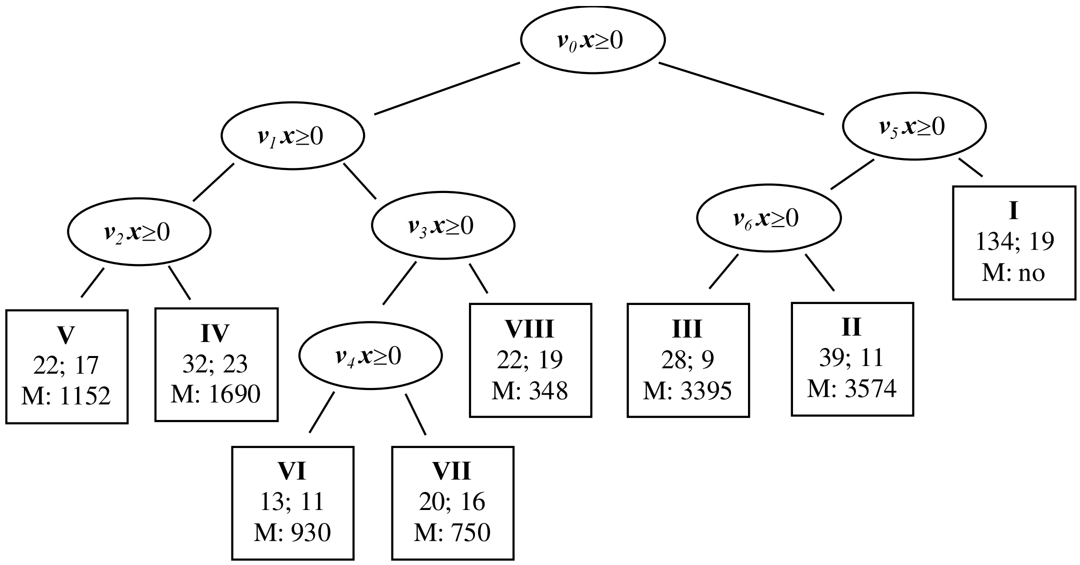

Fig. 1 Regression tree for *PBC* data set
($\boldsymbol{v}_{0}$,.., $\boldsymbol{v}_{6}$ – hyper-planes, $\boldsymbol{x}$ – feature vector, *M* – median survival time)

The regression tree obtained for PBC data is shown in Fig. 1. The tree consists of seven internal nodes. The hyper-planes v_0, v_1, ..., v_6 divide the feature space into eight areas represented by leaves. Each leaf contains the information about the number of all the cases that have reached the leaf, and the number of censored cases among them. *M* is the median censored time.

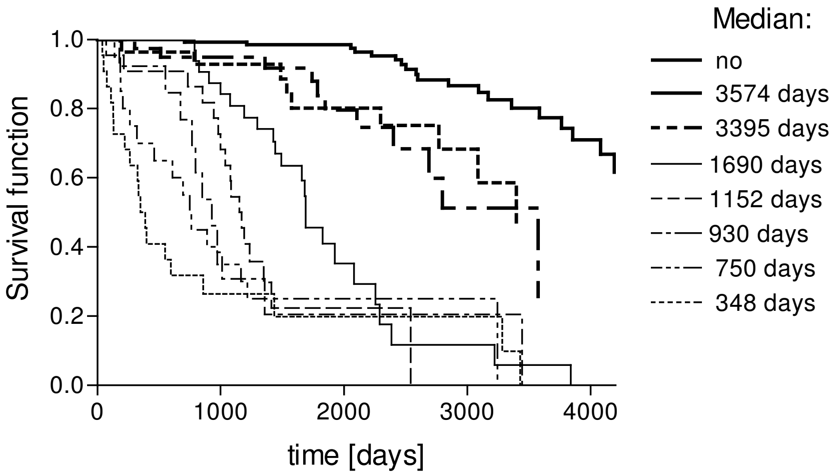

Fig. 2 Kaplan-Meier estimates of the survival functions for distinguished subgroups obtained for *PBC* data set

Kaplan-Meier estimates of the survival functions for each terminal node are presented in Fig. 2. The best prediction is for the patients belonging to the first area, in whom the median survival time cannot be calculated. We only know that it is greater than 4000 days. The worst survival prediction is for the patients who reach the eighth area. The median survival time for them is equal to 348 days.

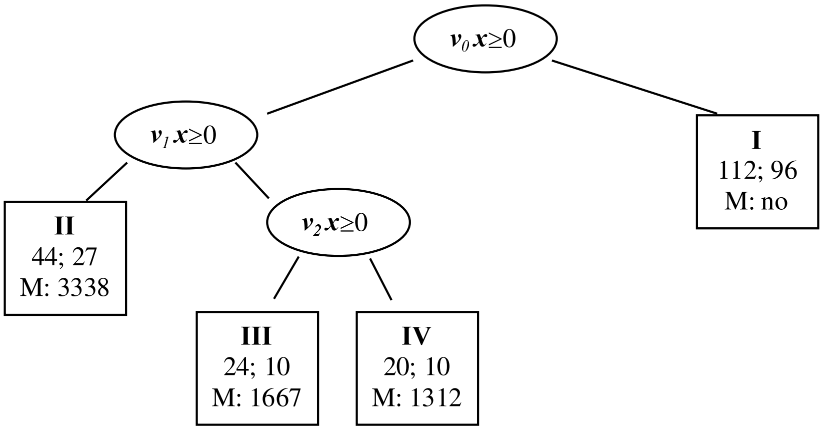

Fig. 3 Regression tree for *Malignant melanoma* data set
(v_0,.., v_2 – hyper-planes, x – feature vector, M – median survival time)

The regression tree received for the *Malignant melanoma* data is shown in Fig. 3. The tree consists of three internal nodes that divide the feature space into four areas represented by leaves (I, II, III, IV).

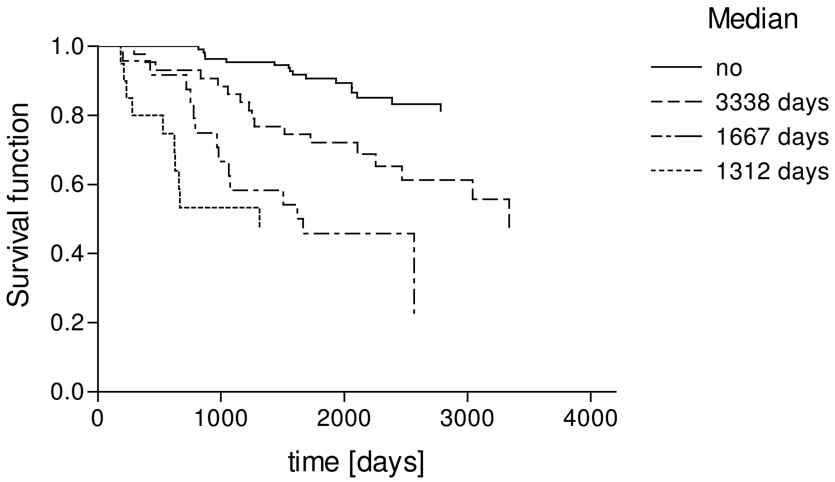

Fig. 4 Kaplan-Meier estimates of the survival functions for distinguished subgroups obtained for *Malignant melanoma* data set

Kaplan-Meier estimates of the survival functions for each terminal node are presented in Fig. 4. The best prediction is for patients belonging to the first area, in whom the median survival time can not be calculated. The worst survival prediction is for the patients who reach the forth area. The median survival time for them is equal to 1312 days.

5. Conclusions

The proposed method allows induction of multivariate regression tree for the survival time prediction. The approach is able to identify subgroups of patients with homogeneous survival experience. The dipolar criterion function, which is exploited in searching for the splits in internal nodes, enables to use information both from the uncensored and the censored cases. Appropriate degree of generalization is received by using the pruning algorithm that can cope with censoring. Application of the regression tree as a prediction tool facilitates interpretation of the results.

Acknowledgements

The work was supported by the grant W/WI/1/02 from Białystok Technical University.

Bibliography

[1] Breiman L., Friedman J. H., Olshen R. A., Stone C. J., *Classification and Regression Trees,* Wadsworth, 1984.-1

[2] Murthy S. K., Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining and Knowledge Discovery*, 2, 345-389, 1998-2

[3] Marubini E., Morabito A., Valsecchi G., Prognostic factors and risk groups: Some results given by using an algorithm suitable for censored survival data, *Statistics in Medicine* 2, 295-303, 1983.-3

[4] Ciampi A., Negassa A., Lou, Z., Tree-structured prediction for censored survival data and the Cox model, *Journal of Clinical Epidemiology* 48(5), 675-689, 1995.-4

[5] Davis R. B., Anderson J. R., Expotential survival trees, *Statistics in Medicine* 8, 947-961, 1989.-5

[6] LeBlanc M., Crowley J., Relative Risk Trees for Censored Survival Data, *Biometrics* 48, 411-425, 1992-6

[7] Segal M. R., Regression trees for censored data, *Biometrics*, 44, 35-47, 1988-7

[8] Fleming T.R., Harrington D.P., *Counting Processes and Survival Analysis*, John Wiley & Sons, Inc., 1991-8

[9] Andersen PK, Borgan O., Gill RD, et al., *Statistical Models based on Counting Processes*, New York: Springer, 1993-9

[10] Bobrowski L., Kretowska M., Kretowski M., Design of neural classifying networks by using dipolar criterions, Proceedings of the Third Conference ”Neural Networks and their Applications”, Częstochowa, Poland, 689-694, 1997-10

[11] Bobrowski L., Design of piecewise linear classifiers from formal neurons by some basis exchange technique, *Pattern Recognition*, 24(9), 863-870, 1991-11

[12] Bobrowski L., Krętowski M., Induction of multivariate decision trees by using dipolar criteria, Zighed D.A., Komorowski J., Żytkow J. (Eds.): PKDD 2000, LNAI 1910, Springer-Verlag, 331-336, 2000.-12

[13] Korn E. L., Simon R., Measures of explained variation for survival data, *Statistics in medicine*, 9, 1990, 487-503-13