Global induction of oblique survival trees

Malgorzata Kretowska[0000−0003−4078−7832] and
Marek Kretowski[0000−0001−9175−2678]

Faculty of Computer Science, Bialystok University of Technology,
Wiejska 45a, 15-351 Bialystok, Poland
m.kretowska, m.kretowski@pb.edu.pl

Abstract. Survival analysis focuses on the prediction of failure time and serves as an important prognostic tool, not solely confined to medicine but also across diverse fields. Machine learning methods, especially decision trees, are increasingly replacing traditional statistical methods which are based on assumptions that are often difficult to meet. The paper presents a new global method for inducing survival trees containing Kaplan–Mayer estimators in leaves. Using a specialized evolutionary algorithm, the method searches for oblique trees in which multivariate tests in internal nodes divide the feature space using hyperplanes. Specific variants of mutation and crossover operators have been developed, making evolution effective and efficient. The fitness function is based on the integrated Brier score and prevents overfitting taking into account the size of the tree. A preliminary experimental verification and comparison with classical univariate trees was carried out on real medical datasets. The evaluation results are promising.

Keywords: survival tree · oblique splits · evolutionary computation

1 Introduction

Is it possible to predict the risk of death after a cancer diagnosis? Can we classify individuals into risk groups for disease relapse? These are some of the questions that survival analysis attempts to answer.

What exactly is survival analysis? It is a set of tools, often statistical, which are able to cope with survival data, in which time of a certain event occurrence is investigated. A characteristic element of this type of data is censoring, which means that for some observations the precise time of the event of interest, called failure, is unknown. Statistical methods often rely on various assumptions that must be met for the results to be accurate [5]. Machine learning methods, on the other hand, are not subject to such limitations. They are constantly developed attempting to successfully address the aforementioned questions.

Tree-based models are among the most commonly used machine learning methods for analyzing censored data. We can differentiate between individual trees and ensembles. In both cases, univariate (with axis-parallel tests in internal nodes) and oblique solutions are available, with the former being predominant.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33

The construction process for typical survival tree models follows a greedy, top-down approach. This involves two main phases: induction and pruning. During induction, the focus is on recursively minimizing a specified impurity measure [20] or maximizing between-node separation [17]. The pruning step typically involves cost-complexity pruning [3] or its survival extension, split-complexity pruning [17]. Despite pruning efforts, the resulting trees often remain overgrown [13], and adopting more global approaches may be beneficial.

A different approach, called a conditional inference framework, was introduced by Hothorn et al. [8], and next by Kundu and Ghosh [14], where the split importance is assessed during node creation, obviating the need for additional pruning phases. To other univariate survival tree models belong also median regression trees [4] or a non-greedy induction method introduced in [2]. Oblique trees, in which multivariate tests in internal nodes are in a form of a hyperplane, belong to less common solutions. Kretowska [12] proposed dipolar survival tree while oblique random survival forest was introduced by Jaeger et al. [10].

The paper presents a novel method for global inducing survival trees that include Kaplan–Meier estimators in their leaves. Using a specialized evolutionary algorithm, the method searches for a whole oblique tree, simultaneously tree structure and all tests. Specific variants of mutation and crossover operators have been developed to ensure effective and efficient evolution. The fitness function, based on the integrated Brier score, mitigates overfitting by considering the model complexity. A preliminary experimental verification was conducted using five medical datasets. The prediction ability was compared with that of two state-of-the-art survival trees. Additionally, the interpretable model obtained for the follicular cell lymphoma dataset was discussed in detail.

2 Preliminaries

Survival data We assume having a learning set, L, which consists of M observations. In survival analysis, the ith observation is described by a set of three values $(\mathbf{x}_i, t_i, \delta_i)$, where $\mathbf{x}_i$ is the N-dimensional feature vector, t_i is the observed time, which for uncensored subjects is equal to its failure time, for censored - it takes values of the follow-up time, δ_i is the failure indicator, which takes one of two values: 0 for censored observations or 1 otherwise.

The distribution of the survival time may be represented by a survival function $S(t) = P(T > t)$, which gives the probability of surviving beyond the time t. Kaplan–Meier (KM) method [11] is one on the most common nonparametric estimators of the survival function. If we assume that the events of interest occur at D distinct times $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$, it is calculated as follows:

$$\hat{S}(t) = \prod_{j|t_{(j)} \leq t} \left(\frac{m_j - d_j}{m_j} \right), \tag{1}$$

where d_j is the number of events at time $t_{(j)}$ and m_j is the number of patients at risk at $t_{(j)}$ (i.e., who are alive or experience the event of interest at $t_{(j)}$).

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33

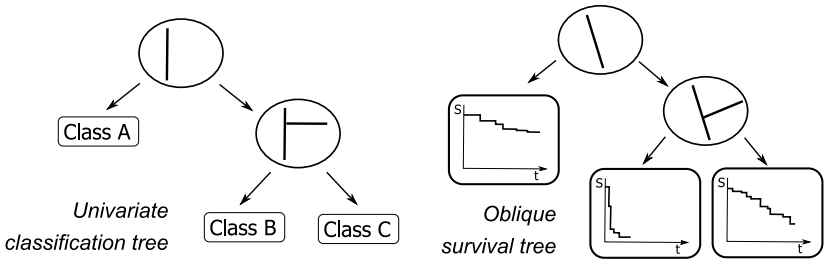

Fig. 1: Univariate decision tree vs. oblique survival tree.

Oblique survival trees A tree structure consists of internal and terminal nodes called leaves. The purpose of trees is to divide a feature space (internal nodes) into homogeneous data regions (leaves) for a given task. In case of survival tree, each terminal node, in our approach, is characterised by the Kaplan–Meier estimator and we aimed at obtaining areas with homogeneous survival experience.

Proposed here, oblique survival trees, divide the feature space by the test of the form of hyperplane $H(\mathbf{w}, \theta)$, where $\mathbf{w} = [w_1, \ldots, w_N]^T$, instead of one single variable x_i. Terminal nodes contain the feature vectors corresponding to distinct regions of the feature space (Figure 1).

Integrated Brier score Taking into account censored survival data, the exact failure time is unknown for some of the observations. Therefore, a direct comparison of real and predicted survival times is impossible. One of the most common performance measures is the integrated Brier score [7]. The first step in calculating the IBS is to compute the Brier score as:

$$BS(t) = \frac{1}{M}\sum_{i=1}^{M}(\hat{S}(t|\mathbf{x}_i)^2 I(t_i \leq t \wedge \delta_i = 1)\hat{G}(t_i)^{-1} + (1-\hat{S}(t|\mathbf{x}_i))^2 I(t_i > t)\hat{G}(t)^{-1}), \quad (2)$$

here $\hat{S}(t|\mathbf{x}_i)$ is the KM estimator, $\hat{G}(t)$ denotes the KM estimator of the censoring distribution, $I(condition)$ is equal to 1 if the condition is fulfilled and zero otherwise. Integrated Brier score (IBS) is obtained by:

$$IBS = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} BS(t)dt, \quad (3)$$

3 Evolutionary Induction

The proposed evolutionary algorithm extends the global induction of standard decision trees [13]. In this short paper, we have to concentrate on the elements crucial to survival analysis and omit more general elements.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33

3.1 Representation, Initialization, and Termination Condition

In non-terminal nodes, only oblique tests based on hyperplanes are allowed. It means that if nominal features are part of the analyzed datasets they need to be first converted (typically into a group of binary features). In every leaf of survival trees, a KM estimator is situated based on the training objects that reached that leaf. As the structure of the tree evolves during induction, the locations of all training data need to be constantly known. From the computational point of view, it is convenient to store this information smartly (once allocated table) and, as the modifications are often local, update only the corresponding areas/subtrees. As a consequence, individuals are not encoded but represented as standard binary trees with additional data structures.

Appropriate population initialization significantly shortens evolution and allows for more efficient use of resources. The initial trees should be diverse and preferably similar in size to the target survival trees, which are usually quite compact. A simple top-down algorithm with a tree height limit is used, which is activated on random small subsets of the training set. Tests in internal nodes are created based on randomly selected dipoles - pairs of observations. Dipoles can be created between uncensored observations and between earlier uncensored observations and later censored observations. When generating tests, longer dipoles are preferred (taking into account the difference between failure times), because their intersection means that these observations will be in two disjoint subtrees (and ultimately leaves). The same mechanism of selecting dipoles and creating tests based on them is also used in the mutation operator.

Evolutionary induction ends when, after a given number of iterations, no individual with a better fitness value is found (default 1000 iterations) or when the limit of the number of iterations is reached (default 5000).

3.2 Genetic Operators

As in a typical evolutionary algorithm [13], two genetic operators are used. Crossover allows the exchange of genetic material between two individuals. In the most typical variant, two nodes (including subtrees) are randomly selected in two trees and the entire subtrees are replaced. It is also possible to: exchange the tests themselves or conduct a crossover with the best individual so far. Since crossing in relation to tree structures can be destructive [6], this operator is applied to trees with a rather low probability (default 0.2). Mutation is the main mechanism for differentiating individuals and is performed on the tree with a high probability (default 0.8). The tree structure can be modified directly, by pruning a randomly chosen subtree to a leaf, or by replacing the leaf with an internal node with a new test. The structure may change indirectly when an existing test is modified (e.g., by randomly changing or resetting one weight in the hyperplane) and the corresponding subtree is changed as a result. The key operation ensuring the efficiency of exploration of the search space is generating a new test. The dipole mechanism described earlier is used here, thanks to which it is possible to direct the search sensibly (by avoiding ineffective tests).

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33

3.3 Fitness Function

The fitness function is the most crucial component of any evolutionary algorithm. In the context of evolutionary machine learning, defining functions directly is not feasible, as the objective of the algorithms is to perform (predict) as effectively as possible not on the training data, but on data that is unavailable during induction. A common approach is to optimize a measure of solution quality on the training data, coupled with an additional factor reflecting model size to prevent overfitting (regularization). In the case of survival trees, we calculate the integrated Brier score (IBS) for the training data and determine the tree size (number of leaves). The fitness function is then defined as follows:

$$Fitness(T) = IBS(T) + \alpha(Size(T) - 1), \tag{4}$$

Adjusting the α value allows for the control of the expected complexity of the resulting tree. In this formulation, all tests (irrespective of the number of features used) hold equal importance. If a preference for simpler tests is desired, the *Size* term could be made dependent on the number of features used.

4 Preliminary Experimental Validation

The first part of the experiments aims to compare the predictive ability of the evolutionary induced oblique survival tree (EIOST) with two state-of-the-art univariate survival trees: the conditional inference tree (CItree) [8] and the recursive partitioning for survival trees (RPtree) [3]. Both of these solutions are publicly available in R packages `party` and `rpart`, respectively. The parameters of the EIOST (see Section 3) were kept constant during all experiments.

The experiments were conducted using five publicly available medical datasets with the percentage of censored cases from 30.4 to 87.3, the number of observations from 418 to 2231, and 4 to 39 attributes. In Table 1, the IBS calculated for RPART, CItree, and EIOST are presented. The EIOSTs were induced with a default value of $\alpha = 0.001$. This indicates that the reported values of IBS may not represent the optimal performance for the specific dataset, leaving potential for further enhancement. For this fixed α value, in three for five datasets the proposed method gives the best IBS values. The number of leaves is similar to other solutions.

The follicular cell lymphoma study (follic) dataset contains information about 541 patients described by four attributes: age, hemoglobin(hgb), clinical stage (stg), and chemotheraphy (ch). In Figure 2, we can observe the impact of the α value on the IBS and the number of leaves obtained for the follic dataset. It is evident that the tree complexity decreases with increasing values of α. We start with approximately 33 leaves for $\alpha = 0.0001$ and gradually decrease to only one leaf at $\alpha = 0.1$. The best IBS value of 19.9 was achieved for $\alpha = 0.005$, with a corresponding number of leaves equals 3. This result clearly outperforms the one reported in Table 1 and is better than the IBS calculated for the univariate trees.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33

Table 1: Integrated Brier score of three types of survival trees: RPtree, CItree, and EIOST inducted with $\alpha = 0.001$; Size denotes the mean number of leaves and IBS denotes the mean ± standard deviation of IBS calculated over 5 runs of 10-fold cross-validation multiplied by 100.

Dataset	#obs	#at	%cen	RPtree		CItree		EIOST	
				IBS	Size	IBS	Size	IBS	Size
pbc [5]	418	8	61.5	15.42 ± 0.7	12.5	14.65 ± 0.5	6.9	**14.50** ± 0.9	8.2
follic [18]	541	4	49.7	20.88 ± 0.4	3.9	**20.63** ± 0.5	4	21.43 ± 1.1	9.6
nwtco [19]	668	5	87.3	10.43 ± 0.2	4.1	10.24 ± 0.2	3.8	**10.10** ± 0.3	4.1
mgus2 [15]	1384	6	30.4	14.76 ± 0.3	2	14.14 ± 0.3	12.6	**13.16** ± 0.3	4.7
peakv02 [9]	2231	39	67.5	**16.37** ± 0.1	3.9	16.63 ± 0.1	7.3	16.46 ± 0.6	3.2

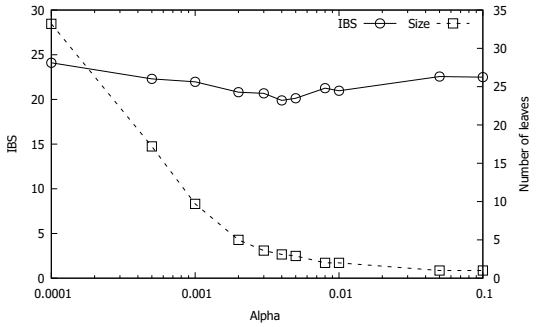

Fig. 2: Accuracy versus interpretability trade-off for follic dataset.

In Figure 3a, we can see the EIOST induced for the follic dataset with $\alpha = 0.005$. The tree divides the feature space into three distinct regions represented by leaves. L3 represents the region with the worst prognosis, having a median survival time of 6.17 years, while the best prognosis is for L1, where the median survival time cannot be calculated. This is evident in Figure 3b, where three KM survival functions are depicted. The disparities between estimators are statistically significant (log-rank test, $p < 0.0001$), indicating that the hyperplanes in the internal nodes have split the feature space into areas with varying survival experiences.

5 Conclusions

In this paper, we propose a novel method for global induction of oblique survival trees tailored for analyzing censored data, which includes observations with unknown exact failure times. An essential aspect of survival analysis tools is the ability to leverage this incomplete information during the induction process. We achieve this through specialized evolutionary algorithm with specific initializa-

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33

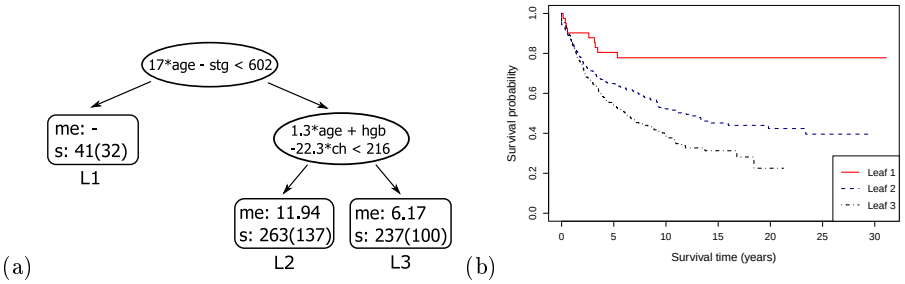

Fig. 3: EIOST inducted for follic dataset (a) with the corresponding KM survival functions (b); *me* denotes the median survival time in the leaf, *s* indicates the number of observations (number of censored cases).

tion, variants of genetic operators, as well as the fitness function combining the integrated Brier score and tree complexity.

Based on five real datasets, two characteristics of the survival trees, predictive ability and model complexity, were compared with two existing univariate tree models. The preliminary results are encouraging. In three datasets, the predictive ability of EIOST is better than the results obtained for the competitors, while the number of nodes is small facilitating the interpretability of the model. The experiments were conducted with default values of α and the quality of the oblique survival tree may be improved by adjusting it to a given problem.

One of the major tasks of the resulting tree is the ability to distinguish areas in the feature space that would contain patients with varying survival experiences. The example of the tree model induced for the follicular cell lymphoma dataset points out that this objective was achieved. The Kaplan-Meier survival functions calculated for leaves differ significantly.

The proposed solution requires further investigation. A possible path is to replace the integrated Brier score in the fitness function with other measures, such as the likelihood method [16], as the IBS can favor tests with higher specificity [1]. Additionally, we aim to extend the algorithm to accommodate other types of survival data, such as discrete survival data or data with competing risks.

Acknowledgments. This work was supported by Bialystok University of Technology under the grant WZ/WI-IIT/4/2023 founded by Ministry of Science and Higher Education.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Assel, M., Sjoberg, D.D., Vickers, A.J.: The brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic and prognostic

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33

research **1**(1), 1–7 (2017)
2. Bertsimas, D., Dunn, J., Gibson, E., Orfanoudaki, A.: Optimal survival trees. Machine Learning **111**(8), 2951–3023 (2022)
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth, Belmont, CA (1984)
4. Cho, H.J., Hong, S.M.: Median regression tree for analysis of censored survival data. IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans **38**(3) (2008)
5. Fleming, T.R., Harrington, D.P.: Counting Processes and Survival Analysis. John Wiley & Sons (1991)
6. Freitas, A.A.: Data mining and knowledge discovery with evolutionary algorithms. Springer Science & Business Media (2002)
7. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine **18**, 2529–2545 (1999)
8. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics **15**(3), 651–674 (2006)
9. Hsich, E., Gorodeski, E.Z., Blackstone, E.H., Ishwaran, H., Lauer, M.S.: Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circulation: Cardiovascular Quality and Outcomes **4**(1), 39–45 (2011)
10. Jaeger, B.C., Long, D.L., Long, D.M., Sims, M., Szychowski, J.M., Min, Y.I., Mcclure, L.A., Howard, G., Simon, N.: Oblique random survival forests. The Annals of Applied Statistics **13**(3), 1847–1883 (2019)
11. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association **53**, 457–481 (1958)
12. Kretowska, M.: Piecewise-linear criterion functions in oblique survival trees induction. Arificial Intelligence in Medicine **75**, 32–39 (2017)
13. Kretowski, M.: Evolutionary Decision Trees in Large-scale Data Mining. Springer (2019)
14. Kundu, M.G., Ghosh, S.: Survival trees based on heterogeneity in time-to-event and censoring distributions using parameter instability test. Statistical Analysis and Data Mining: The ASA Data Science Journal **14**(5), 466–483 (2021)
15. Kyle, R.A., Therneau, T.M., Rajkumar, S.V., Offord, J.R., Larson, D.R., Plevak, M.F., Melton III, L.J.: A long-term study of prognosis in monoclonal gammopathy of undetermined significance. New England Journal of Medicine **346**(8), 564–569 (2002)
16. LeBlanc, M., Crowley, J.: Relative risk trees for censored survival data. Biometrics **48**, 411–425 (1992)
17. LeBlanc, M., Crowley, J.: Survival trees by goodness of split. Journal of the American Statistical Association **88**(422), 457–467 (1993)
18. Pintilie, M.: Competing Risks: A Practical Perspective, vol. 58. John Wiley & Sons (2006)
19. Therneau, T.M.: survival: Survival Analysis (2016), http://CRAN.R-project.org/package=survival, R package version 2.39
20. Therneau, T.M., Grambsch, P.M., Fleming, T.R.: Martingale-based residuals for survival models. Biometrika **77**(1), 147–160 (1990)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_33