SurTree: Constructing Optimal Survival Trees using MurTree

Tim Huisman (T.J.Huisman-1@student.tudelft.nl)

(1) Problem

Survival analysis is a field focused on predicting the survival rates of patients. **Survival trees** can divide patients in groups to make predictions more accurate. Making good survival trees is hard, and finding the best tree **takes a lot of time**.

MurTree (Demirović et al., 2022) is an algorithm that finds the **best classification trees** for a dataset using dynamic programming. We could use its techniques to find the **best survival trees** instead.

(2) Research questions

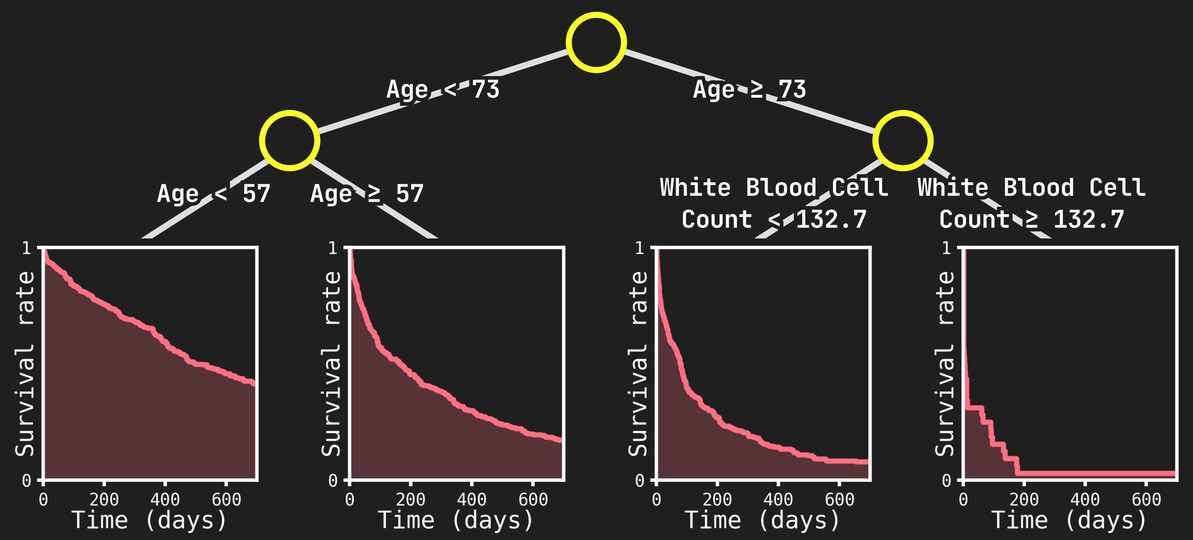
- **How can MurTree be adapted for survival analysis?**
- **How does the algorithm compare in runtime to the state-of-the-art?**
- **How do the generated survival trees compare to the state-of-the-art's trees with regard to other metrics?**

(3) Methodology

The aim is to minimize the objective function defined by LeBlanc & Crowley (1992). The **Optimal Survival Trees** (**OST**) algorithm (Bertsimas et al., 2022) also tries to minimize this function, using gradient descent.

SurTree, our new algorithm, manages to find survival trees that minimize this function as much as possible. It is based on **MurTree**, with the most relevant changes being:

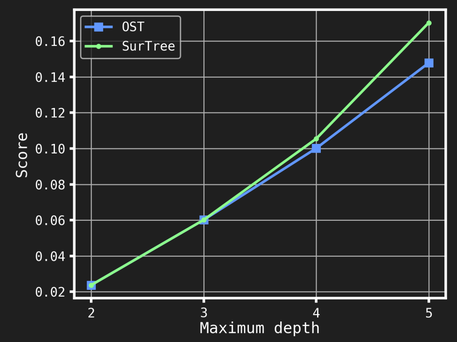
- The replacement of an **integer misclassification** with a **floating-point error**, calculated with the **objective function**.
- A new implementation of the **terminal solver**, which can calculate an error using **three precomputed values**.
- The removal of the **similarity lower bound**.



An example of an optimal survival tree, generated by SurTree for the LeukSurv-dataset (Henderson et al., 2002). The horizontal axis denotes the time since diagnosis, the vertical axis denotes the fraction of people still (possibly) alive. The red line is used to predict the survival rates of new instances.

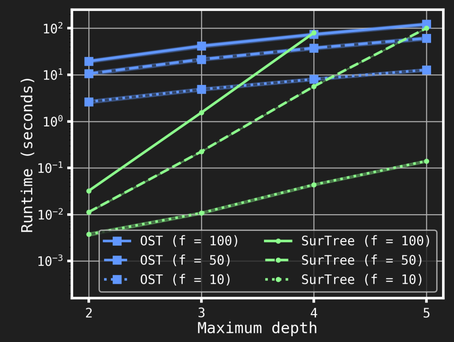
(4) Experiments & Conclusions

To evaluate **SurTree**, we compared it to **OST** in a number of aspects. Synthetically generated datasets were used for experimentation, allowing us to observe how both algorithms respond to different problem sizes.



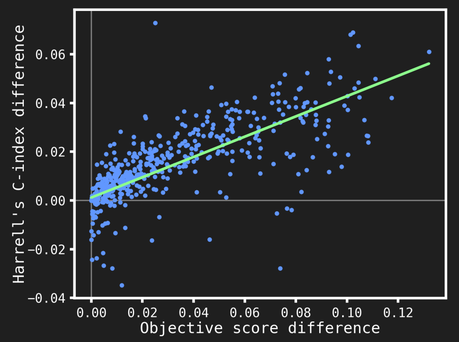
SurTree's and OST's objective score over the maximum tree depth. A higher objective score is considered better.

Since **SurTree** is designed to find **optimal** survival trees, it finds trees with a **higher score** than **OST** once the size constraints allow for more complex trees.



SurTree's and OST's average runtime over the maximum tree depth. The number of features differs per curve.

SurTree is **fast for small trees**, considerably faster than **OST**. However, due to its **exponential runtime**, it takes much longer to finish for higher maximum depths.



SurTree's improvement on OST in Harrell's C-index plotted over its improvement in the objective score.

Harrell's C-index (Harrell et al., 1982) appears to correlate positively with the objective score, suggesting that **SurTree** tends to find trees that are also **better by other metrics**.

References

Bertsimas, D., Dunn, J., Gibson, E., & Orfanoudaki, A. (2022). Optimal survival trees. Machine Learning, 111(8), 2951–3023.
Demirović, E., Lukina, A., Hebrard, E., Chan, J., Bailey, J., Leckie, C., Ramamohanarao, K., & Stuckey, P. J. (2022). MurTree: Optimal Decision Trees via Dynamic Programming and Search. HAL (Le Centrepour la Communication Scientifique Directe).
Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. JAMA, 247(18), 2543.
Henderson R., Shimakura S., Gorst D. (2002) "Modeling spatial variation in leukemia survival data". In: Journal of the American Statistical Association 97.460, pp. 965–972.
LeBlanc, M., & Crowley, J. (1992). Relative Risk Trees for Censored Survival Data. Biometrics, 48(2), 411.

Supervisors:
Dr. Emir Demirović
Jacobus G. M. van der Linden

TU Delft

CSE3000 Research Project
Final Presentation Poster