MA691 : COBRA-18 Project Report

Anmol Choudhary (180123004)
Ashish Kumar Barnawal (180123006)
Jay Vikas Sable (180123019)
Shivam Kumaar Arya (180123044)

20 November 2021

Abstract

In this project we tried to predict the survival function of patient diagnosed with Primary biliarycholangitis (PBC) using COBRA with Survival Trees as weak learners.

Theory

We denote T to be the time at which an event occurs (in our case a patient dies). T is also called the response variable.
The survival function at time t is defined to be the probability of survival at that time. In other words,

$$S(t) = Pr\{t > T\} = 1–F(t)$$

where $F(t)$ is the Cumulative Distribution Function of variable T.

From lifetime data, we can estimate the survival function using the Kaplan-Meier estimator:

$$\widehat{S}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

with t_i a time when at least one event happened, d_i the number of events that happened at time t_i, and n_i the individuals known to have survived (have not yet had an event or been censored) up to time t_i

In this project we use Survival Trees [4] which use the Kaplan Meier estimator at its leaf node for estimating the survival function. Then we use COBRA [1] as Ensembling algorithm over the Survival Trees to get the survival time and the survival function.

Implementation

For the implementation we made use of scikit-survival [5] library. The library contains the class `SurvivalTree` which uses Survival Tree [4] for predicting the survival function after training it with training data.

We create a class `CobraSurvivalTree` which inherits from `SurvivalTree` , the difference is that the predict function returns a single value: the expected survival time rather than returning the survival function. This is useful for the COBRA implementation.

The expected survival time is calculated as follows:

$$
\begin{aligned}
E[T] &= \int_0^\infty t\, dF(t) = -\int_0^\infty t\, dS(t) \\
&\approx -\int_0^{T_{\max}} t\, dS(t) + T_{\max} S(T_{\max}) \\
&\approx \sum_{i=1}^{n-1} t_i\, \Delta S_i \; + \; t_n S_n
\end{aligned}
$$

Here n is the number of elements in `sksurv.SurvivalTree.event_times_` array.

Now, we apply COBRA on the survival time of the dataset with $\epsilon = 3$ and using `CobraSurvivalTree` the weak learner. This model will give predictions for the survival time.

To predict survival function, we will take the average of survival function over the set of selected observations $\{(x_i, y_i)\}_{i \in \mathcal{D}}$ and over the set of selected machines α_m (A datapoint (x_i, y_i) is selected if it in within ϵ range of all selected machines i.e. $i \in \mathcal{D}$ if $|r_j(\mathbf{x}_i) - r_j(\mathbf{x})| < \epsilon \; \forall \; r_j \in \alpha_m$). Here $\mathcal{D}$ denotes the set of indices of selected datapoints and α_m is the set of selected machines in the COBRA algorithm. The survival function is then calculated as follows

$$S(t) = \frac{1}{|\alpha_m| N} \sum_{r_j \in \alpha_m} \sum_{i \in \mathcal{D}} S_{r_j}(t \,|\, \mathbf{x}_i)$$

Here $S_{r_j}(t \,|\, \mathbf{x}_i)$ is the predicted survival probability of $\mathbf{x}_i$ at time t by machine r_j

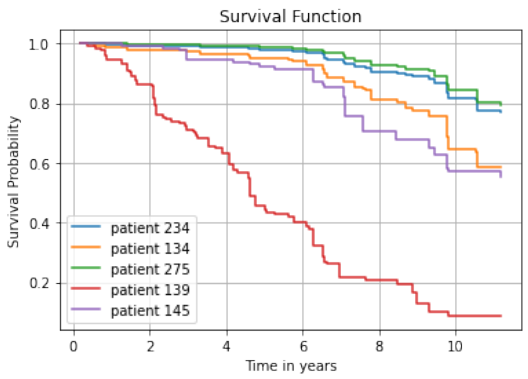
Results

After running our implementation of COBRA with survival tree we get the following result

```
Mean Absolute Error : 2.31 years

Concordance index: 0.68
```

And the following survival function prediction:



References

[1] G´erard Biau, Aur´elie Fischer, Benjamin Guedj, and James Malley. COBRA: A Combined Regression Strategy.Journal of Multivariate Analysis, 2016.

[2] Fleming and Harrington. Primary Biliary Cirrhosis (PBC) Data, Appendix D, 1991.

[3] Benjamin Guedj and Bhargav Srinivasa Desikan. Pycobra: A python toolbox for ensemblelearning and visualisation, 2019.

[4] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Randomsurvival forests.The Annals of Applied Statistics, 2(3):841 – 860, 2008

[5] Sebastian P¨olsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn.Journal of Machine Learning Research, 21(212):1–6, 2020