Response to the review of opponent – doctoral dissertation

Reviewer: Roland Molontay, PhD
Title: Machine learning supported survival analysis
Author: Róbert Csalódi
Supervisors: János Abonyi, DSc
Tamás Ruppert, PhD

Dear Dr. Roland Molontay,

I am writing to express my sincere gratitude for your invaluable feedback and insightful comments on my dissertation. I have thoroughly reviewed your questions and you can find my detailed responses below this letter.

Sincerely
Róbert Csalódi

1. In the third chapter, in the student dropout case study, how does treating the problem as a single outcome problem limit the usefulness of the model considering that "Failing to account for competing risks leads to biased estimates"?

Thank you for this great question. The principles for handling competing risks are outlined in Section 3.3.1 following the list of variables. In this scenario, for simplicity, I censored those students who successfully graduated. While this is a common practice, it comes with the cost of violating the condition of independent censoring, introducing a slight inaccuracy. However, a detailed analysis of the case study is provided in Chapter 4. Fundamentally, the thesis aims to present novel algorithms rather than identify dropout characteristics. Therefore, I made the decision to overlook this aspect in this chapter. Otherwise, applying the subdistribution hazard model would have been the simplest approach in this case, as the explanatory variables remain constant over time here.

2. In the Covid case study, how was it taken into account that different countries were affected by the pandemic at different times?

Thank you very much for your important question. I downloaded the dataset on 27.09.2021, during the descending trend of the third local peak observed worldwide in weekly cases. Since the declaration of COVID-19 as a pandemic happened on 14.03.2020, approximately one and a half years had elapsed by the time of the download. We operated under the assumption that the virus had sufficiently impacted every country by this time to allow for representative analysis. No other adjustments were made to compensate for any delay in its spread.

3. What are the advantages of the heterogeneous survival model over the homogeneous one? To what extent can the model obtained with the novel procedure introduced by the author be considered homogeneous?

Thank you for addressing this important question. The model is designed to provide as homogeneous results as possible. To achieve this, I utilized information criteria to determine the optimal number of components. When the optimal number of components is reached, the likelihood of fitting is maximized, suggesting the best possible fit. Moreover, the information criterion ensures that the number of components is not excessively high, thus mitigating overfitting issues. In fact, selecting a larger number of clusters than suggested by the information criteria often results in merged clusters, where there are no statistical differences in survival time and explanatory variables. Therefore, it is straightforward to manually detect if the number of components is too large. For those seeking more in-depth results, the goodness of fit of the data in different clusters can be compared.

4. In Chapter 3, the author emphasizes that the introduced method's important contribution is providing interpretable local models. Does the author plan to compare his solution with other interpretable machine learning methods (any supervised learning method, e.g., neural networks + interpretable layers such as SHAP)?

Thank you for your question. Throughout the case study, I conducted several comparisons, all of which have been included in the dissertation. However, a comparison with other machine learning models has not been made for this algorithm.

5. Could the method presented in the fourth chapter be extended to continuous time?

Thank you very much for your significant question. Extending the analysis to significantly more time instances increases computational complexity, as frequent itemset mining must be performed at each time instance. Furthermore, it is crucial that each time instance contains a sufficient number of events that are worth executing the method on it. Without events, frequent itemsets cannot be obtained, and this issue must be addressed. One approach could be to increase the sampling frequency, or alternatively, consider treating previous time instances as equivalent to empty time instances. In conclusion, the method could be extended if the dataset is sufficient for this purpose.

6. In the fourth and fifth chapters, association rules are evaluated using the confidence metric. However, the confidence metric has several limitations, and often the lift metric is preferable. Why was the lift metric not considered in these studies?

Thank you for raising such an important question. Indeed, mining frequent itemsets, sequences, and association rules involves various metrics beyond confidence, one of which is the lift parameter. This metric represents the ratio of an association rule's confidence to the probability of its consequent part. Its purpose is to assess whether a given antecedent boosts or hinders the occurrence of the consequence. In Chapter Five, the presented case study aims to predict future diseases based on existing ones, rather than solely identifying diseases to avoid to prevent future ailments. While confidence served as an intuitive metric in this context, I acknowledge the potential significance of Lift and intend to explore its implications in future research or applications.