Support Vector Machines for Survival Analysis with R

Aarjav Jain

2024-10-26

Every day, we live a life of unpredictable outcomes in a world full of uncertainties. Amidst this chaos, our survival itself can sometimes feel like a matter of fate or chance. Yet, behind every journey of survival - whether it is a patient undergoing treatment, a startup navigating its early years, or even a machine outlasting its warranty - lies a series of patterns that can be analyzed, leading to new insights. This statistical procedure, which is dedicated to examining and predicting the duration of time until a specific event (or multiple events) occurs, is known as survival analysis. Survival analysis is used with a binary or dichotomous outcome of interest[1]: success and failure, life and loss, managing to finish a Netflix series before all my free trials run out. It provides a framework to predict and understand such outcomes across fields like healthcare, engineering, and finance. As *The Hunger Games* stated, "May the odds be ever in your favor." Luckily for us, survival analysis can help us understand those odds, not merely hope for them.

This report examines the purpose behind the research, the statistical methods used, the results, its critiques, and the ethical considerations (especially in the medical sphere) of the research presented in "Support Vector Machines for Survival Analysis with R" by Césaire J. K. Fouodo, Inke R. König, Claus Weihs, Andreas Ziegler, and Marvin N. Wright. As outlined in the paper, traditional survival analysis techniques (such as the Cox Proportional Hazards model) face significant challenges in high-dimensional datasets. To put it briefly, the method relies on assumptions that may not hold true in all cases and demands very high computational power. To address these limitations, the researchers explore a different approach using support vector machines (SVMs).

The paper presents three distinct approaches to implementing SVMs for survival analysis. As we know, SVMs, at their core, work by finding an optimal separating line/hyperplane between different classes of data points, essentially maximizing the margin between these classes. However, survival analysis presents a unique challenge due to censoring - this is when the actual survival time is unknown because the event occurs beyond the time limit of the study period.[2] By adapting SVMs to handle such data, the authors offer an alternative to traditional methods in survival analysis.

The first approach is the regression approach. While traditional regression penalizes all predictions that deviate from true values, censored and uncensored data are treated differently under this approach. The model penalizes the predictions that differ from actual survival time only for uncensored observations. For censored observations, the model only penalizes predictions that fall below the censoring time, as

any prediction above this time could potentially be accurate. This allows the model to learn from incomplete data while preserving information from censored observations and improving overall accuracy.

The second method they introduced is the ranking approach. "This approach considers survival analysis based on SVMs as a classification problem with an ordinal target variable."[3] - unlike the previous method, where we predict the exact survival time, this method focuses on predicting the order or ranking of survival times among individuals. In this manner, they predict which individuals are more likely to experience an event sooner than others. The outcome is not just a binary classification but an ordered sequence of who is at higher risk.

Under this approach, the model aims to maximize the C-Index or Concordance Index defined by Van Belle[4], which is essentially a measure of how well the model ranks survival times. This approach of relative ordering rather than predicting exact survival times is applicable in the medical sphere and its finite resources.

Lastly, we have the hybrid approach, which combines the strengths of both regression and ranking methods. This is to balance the output between precise time estimates and relative risk ordering. While this output is certainly more informative, it comes at a computational cost. This is because it results in a much more complex quadratic optimization problem, as laid out in the paper. Additionally, two regularization parameters are needed. While this method is more computationally expensive, it also achieves superior performance by leveraging the strengths of both the previous methods.

Coming to the actual implementation, the authors allow the user to choose one of the four kernels: linear, additive, radial bias, or polynomial. While this allows the data to be solved in higher-dimensional spaces, which is better as the classes are more separable, choosing an appropriate kernel is equally important as it can impact both accuracy and performance. The authors also mention the pros and cons of using different packages, namely `quadprog` and `ipop`, and how each can impact performance and runtime. Once again, the user can choose which solver they use to solve the quadratic optimization problem.

The authors tested their methods using five different datasets, each chosen to represent different scenarios in survival analysis. The primary datasets include Veteran's lung cancer trial study, the Interpretation of Trial Stopped Early study, the Germany Breast Cancer Study Group 2, and the Mayo Clinic Lung Cancer study. This allowed them to test their methods across different dataset sizes and medical fields (which is crucial for generalization!).

In this process, each dataset was randomly divided into five approximately equal subsamples. For each iteration, one subsample served as the test set while the remaining data was used for training. Performance was primarily evaluated based on the C-index, a widely used metric in survival analysis. They compared their results against previous reference methods, including the Cox proportional hazards model, random survival forests (RSF), and gradient boosting. The authors tested three kernel functions

for each survival SVM approach: linear, additive, and radial basis function (RBF). The largest dataset, **GBSG2**, contained 686 patients, while the smallest datasets (**leuk_cr** and **leuk_death**) had 130 patients.

The runtime analysis revealed massive differences between approaches and kernel choices. The hybrid approach had superior prediction accuracy across most datasets. Looking at the C index, the hybrid approach ranked higher in almost all cases. The hybrid approach with additive kernel achieved the highest C-index of 0.71 on the **veteran** dataset. On the **leuk_cr** dataset, it achieved a C-index of 0.72 with both linear and RBF kernels, comparable to the Cox PH model and gradient boosting. In the **leuk_death** dataset, it tied for the best performance with a C-index of 0.72, matching the random survival forest method. The hybrid approach with additive kernel achieved a C-index of 0.68 on the **GBSG2** dataset, though slightly below the random survival forest's 0.69. Finally, the hybrid approach with RBF kernel achieved the highest overall C-index of 0.64 on the **MCLC** dataset.

However, this did not come without a computational cost. The hybrid model consistently took much longer to compute its results across all datasets. Moreover, due to its intensive calculation, it was unable to finish computing on the **GBSG2** dataset using the RBF kernel and had to be interrupted. This is an important factor to keep in mind when using this model. Regardless, it can be said with certainty that the consistently strong performance of the hybrid method shows us that the additional computational cost does translate to meaningful improvements in accuracy.

While the development of such a sophisticated and complex survival analysis tool is technically very impressive, it raises deep concerns about their potential misuse in healthcare resources and allocation. When healthcare systems face a lack of resources, perhaps during a crisis like a pandemic or in an everyday scenario like ICU availability, there might be an inclination to use a predictive algorithm such as this to automate the decision-making process. I firmly believe that the purpose behind such tools is to assist medical professionals in their decision-making process, not replace it in its entirety. Such misuse of this model represents a serious ethical concern that demands scrutiny.

The fundamental issue lies in reducing human life to a mathematical probability. There is a high risk of algorithmic bias in survival predictions. Historical medical data used to train these models often shows social biases and healthcare inequities. Communities of people that have historically had worse health outcomes due to the unavailability of care in their physical community would have lower survival predictions. The algorithm might interpret this as something inherent to the community rather than take into account the inequities they face. This can create a feedback loop with such members receiving less priority and further amplify the issue of disparate healthcare provisions.

Moreover, the use of such an algorithm also raises the subject of human dignity and autonomy. If such an algorithm were to become mainstream, healthcare workers might be pressured to follow its output even if their professional or ethical judgment suggests otherwise. This is very dangerous as it is diminishing the human element of decision-making, which I believe to be absolutely crucial in a field like healthcare.

Despite their statistical dominance, these algorithms cannot capture the complexity of human existence. A patient is much more than just a probability on a piece of paper. They are a human first. They have an identity to them - a parent, a spouse, a child, someone supporting their family, contributing to the community, etc. While one may argue that an algorithm can make a more “objective” decision as opposed to a human, perhaps it is this very ability to make questionable, non-optimal, emotional decisions that makes us human at all. An algorithm cannot replace the human essence.

References

1. Huecker, Shreffler, Jacob. “Survival Analysis.” StatPearls [Internet]., U.S. National Library of Medicine, 22 May 2023, www.ncbi.nlm.nih.gov/books/NBK560604/.↩
2. Clark, T G, et al. “Survival Analysis Part I: Basic Concepts and First Analyses.” *British Journal of Cancer*, vol. 89, no. 2, July 2003, pp. 232–238, dx.doi.org/10.1038%2Fsj.bjc.6601118, https://doi.org/10.1038/sj.bjc.6601118. (https://doi.org/10.1038/sj.bjc.6601118.) Accessed 28 June 2019.↩
3. R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In 9th International Conference on Artificial Neural Networks (ICANN99), pages 97–102, 1999. URL https://dx.doi.org/10.1049/cp:19991091. (https://dx.doi.org/10.1049/cp:19991091.) [p414]↩
4. V. Van Belle, K. Pelckmans, J. Suykens, and S. Van Huffel. Support vector machines for survival analysis. In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), pages 1–8, 2007. [p413, 414, 415]↩