Simultaneous Multi-Task Survival Analysis for Multiple Diseases

Li Xing[1], Yuying Huang[2], Grace Yi[3], Xuekui Zhang[2]

[1]University of Saskatchewan, Saskatchewan, CA [2]University of Victoria, Victoria, CA [3]University of Western Ontario, London, CA

Introduction

To tackle multivariate outcomes prediction problems, the stacking algorithm combines the predictions of multiple tasks, which works better than fitting individual prediction tasks separately[1]. Our research team proposes three variations of the stacking algorithm demonstrate satisfactory performance superior to other methods on continuous and binary outcomes. In this regard, an R package, MTPS, is developed to implement the standard stacking as well as the variant stacking algorithms[2]. We propose to extend our R package MTPS to accommodate analyzing survival data with the implementation of simultaneous multiple survival predictions.

Objective

The goal of this research is to bridge these two research fields of survival analysis and our stacking algorithms:

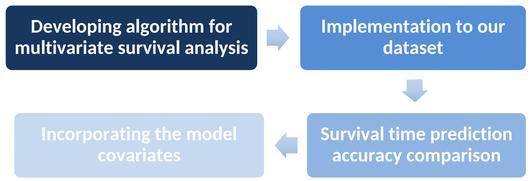

Figure 1. Flowchart of our work design.

- We aim to extend the R package, MTPS, to further handle survival outcomes.
- We compare the survival time prediction results between our extended algorithms and traditional methods of survival analysis to show the improvement.
- We analyze the relatedness between different chronic diseases and their contributions to the model performance in addition to our input covariates.

Data Source

Our sample from the Survey of Health, Ageing and Retirement in Europe (SHARE) dataset was collected over a 9-year span and consists of 11,940 participants. The survival time of five targeted diseases includes heart attack, hypertension, high blood cholesterol, diabetes, and COPD. Furthermore, we use the covariates such as patients' country, age, gender, BMI, depression scale, quality of life, education attainment, marital status, smoking and drinking history as the model input.

Methodology

Stacking is an ensemble machine learning algorithm, which consists of two-steps prediction: it trains a combiner algorithm to integrate information from the first step predictions and re-predicts the final outcome. Various base learners model can be chosen for each step.

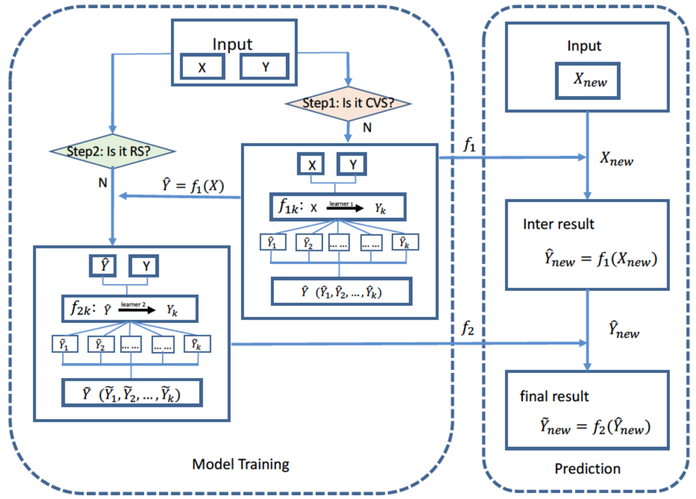

Figure 2. Heuristic description of the Standard Stacking (SS) algorithm, where X is the input covariate matrix, Y is the response variable and f_1, f_2 are the models in the first and second prediction steps, respectively. Residual Stacking (RS), Cross-Validation Stacking (CVS), and Cross-Validation Residual Stacking (CVRS) are variants of Standard Stacking.

Results

We use the Mean Squared Error (MSE) to measure the prediction accuracy of both single outcome and multiple outcomes models. In stacking algorithm, we use the parametric Accelerated Failure Time (AFT) model in the first step to predict survival times and linear regression in the second step.

Table 1. Median values of 100 model fitting MSEs with different partitions of training and testing data.

	Heart attack	Hypertension	High cholesterol	Diabetes	COPD
Single prediction	24.6	5.41	10.69	43.2	42.6
SS	3.23	2.79	2.92	3.08	3.10
RS	3.46	2.80	2.92	3.48	3.05
CVS	3.16	2.81	2.90	3.05	3.02
CVRS	3.44	2.81	2.91	3.50	3.02

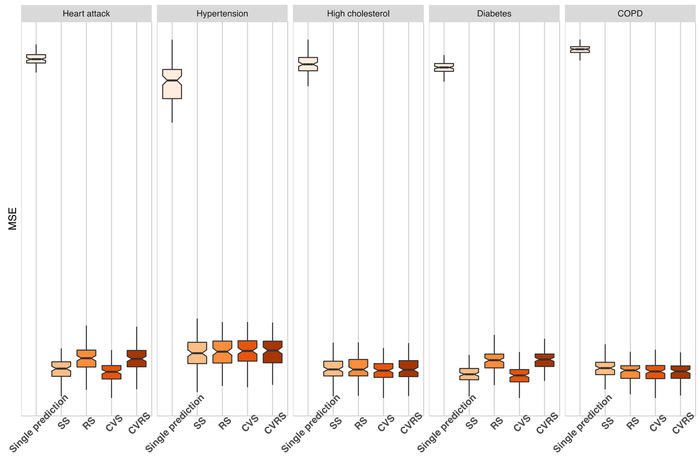

Figure 3. Boxplots of all 100 model fitting log 10 MSEs with different partitions of training and testing data. The single prediction model and the multivariate outcomes prediction model (SS, RS, CVS, CVRS) are performed for the five diseases of our interest.

Conclusions

In this project, we propose a framework to accommodate analysis of multivariate survival outcomes analysis, which extends the scope of multitask learning algorithm. We demonstrate the performance of the proposed stacking survival analysis models using the SHARE dataset to model the survival times of five different diseases (heart attack, hypertension, high cholesterol, diabetes, and COPD), which shows a success in improving prediction accuracy from the traditional model. In the future, we plan to develop refined stacking algorithms to handle longitudinal data or genomic data of complex structure.

Discussion

The successful application of the simultaneous multiple predictions in survival analysis alleviates insufficient use of data and improves prediction performance[3]. Moreover, we are allowed to look into the task relatedness among multiple related survival prediction problems which are rarely considered in survival analysis.

Though the multiple outcomes stacking algorithm improves the prediction results, it entails less transparency in interpreting the covariates in the model, which we are still trying to address. Moreover, the prediction quality in the first step affects the accuracy of the results and the concordance index in the second step. It is worth thinking whether there is a need for multiple outcomes prediction when a certain disease already achieve a high performance in a single prediction model. In this regard, we may need to use a different strategy to handle different diseases.

References

1. Theocharis, A., Pierce, M., & Tzanetakis, G. (2011, December). An empirical investigation of stacking for music tag annotation. In *2011 10th International Conference on Machine Learning and Applications and Workshops* (Vol. 1, pp. 90-95). IEEE.
2. Xing, L., Lesperance, M. L., & Zhang, X. (2020). Simultaneous prediction of multiple outcomes using revised stacking algorithms. *Bioinformatics*, *36*(1), 65-72.
3. Zhou, J., Yuan, L., Liu, J., & Ye, J. (2011, August). A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 814-822).

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Visual and Automated Disease Analytics (VADA) graduate training program. This research was enabled in part by the support provided by Compute Canada (www.computecanada.ca).

The VADA Program
Visual and Automated Disease Analytics Graduate Training Program

NSERC CRSNG