Theses of the doctoral (PhD) dissertation

Machine learning supported survival analysis

Róbert Csalódi

University of Pannonia
Doctoral School in
Chemical Engineering and Material Sciences

Supervisors:
János Abonyi, DSc
Tamás Ruppert, PhD

Department of Process Engineering
Veszprém
2024.

1. Introduction and the aim of the work

Survival analysis is a statistical methodology used to estimate the probability distribution of time until an event occurs. The method was initially applied in biostatistics and medical researches, where the event of interest is typically the time until death. However, this method has spread across various disciplines, as analogies can be drawn between survival times and a multitude of other quantities.

The method serves as an efficient tool in exploring inefficient operation of processes. However, relying solely on this method can be limiting when aiming to identify the root causes of losses. Therefore, integrating survival analysis with machine learning algorithms is crucial to explore the underlying factors in the necessary depth.

This thesis introduces three integrated algorithms that combine survival analysis with machine learning models, offering a more profound understanding of complex processes. The first algorithm presents an integrated survival analysis and expectation-maximization-based clustering framework, identifying clusters based on the similarity of survival times and explanatory variables. The second algorithm introduces an integrated survival analysis and frequent itemset-based association rule mining method, that identifies relevant triggering events defined from time-dependent categorical variables, that lead to consequent events of competing risks. The algorithm estimates the cumulative incidence function directly based on the rule supports and confidences. The third algorithm demonstrates an integrated survival analysis and sequential pattern mining framework, determining the time-dependent confidence function of event continuations.

Through these algorithms, the thesis not only showcases their adaptability but also highlights their potential to provide valuable insights, improve predictions, and enhance decision-making processes.

2. Experimental tools and technologies

The effectiveness of the proposed algorithms is demonstrated through diverse case studies, showcasing their applicability across various domains. The methods are applied to estimate:

1. Remaining useful life of Li-ion batteries based on their capacity, internal resistance and charging conditions

2. Survival times of patients with prostate cancer based on their age, serum hemoglobin level and treatment

3. Mortality rate per 100K population of countries related to COVID-19 pandemic based on demographical and economical data.

4. Dropout rate of university students based on uncompleted subjects patterns

5. Occurrence chances of disorders based on their already existing ones

These studies span across complex and dynamic scenarios in various domains, emphasizing the ability of algorithms to contribute to improved outcomes. The algorithms were developed in Matlab and Python environments.

3 Theses

1. **I developed an integrated survival analysis and expectation - maximization - based clustering framework.**

   (a) I demonstrated, how heterogeneous survival models can be grouped into homogeneous models based on the similarity of survival times and explanatory variables by utilizing an expectation maximization algorithm. This approach defines clusters and simultaneously identifies their survival probabilities using the Weibull distribution and the related continuous explanatory variables, leveraging multivariate Gaussian distributions.

   (b) I presented, how the cluster memberships can be represented using Takagi-Sugeno fuzzy rules, providing a framework to determine the operating domain of continuous variables. With this approach, survival characteristics can be described by considering the domain of continuous variables. The method is versatile and can be applied for categorizing continuous variables.

   Related publications: Róbert Csalódi, Zoltán Birkner, János Abonyi: Learning Interpretable Mixture of Weibull Distributions – Exploratory Analysis of How Economic Development Influences the Incidence of COVID-19 Deaths, Data, 2021. [1]
   Róbert Csalódi, Zsolt Bagyura, János Abonyi: Mixture of survival analysis models - Cluster-weighted Weibull distributions, IEEE Access, 2021. [2]

2. **I developed an integrated survival analysis and frequent itemset-based association rule mining algorithm.**

   (a) I demonstrated, how the probability of competing risks can be determined at specific time instances using event-driven frequent itemset-based association rules. This approach identifies relevant triggering events defined from time-dependent categorical variables, that lead to consequent events defined from competing risks. A sequence of frequent itemsets can be represented as a global, time-independent feature.

   (b) I presented, how the cumulative incidence function of a specific competing risk can be directly determined for the population with a given sequence of frequent itemsets. The method segments the dataset for subjects that supports all the frequent itemset of the selected sequence and estimates the cumulative incidence function based on the modified rule supports and confidences.

   (c) I introduced, how the student dropout rate can be estimated based on patterns of uncompleted subjects. The study has a sample curriculum that prescribes the recommended semester for each subject completion. Inability to meet this requirement marks an uncompleted subject event, a crucial factor associated with subsequent student dropout.

   Related publication: Róbert Csalódi, János Abonyi: Integrated Survival Analysis and Frequent Pattern Mining for Course Failure-Based Prediction of Student Dropout, Mathematics, 2021. [3]

3. **I developed an integrated survival analysis and sequential pattern mining algorithm.**

   (a) I demonstrated, how the time-dependent support and confidence functions of event transitions can be estimated using the integrated survival analysis and frequent sequence mining algorithm. The approach identifies relevant event continuations through frequent sequence mining and determines their support and confidence metrics. The temporal characteristics of the resultant rule confidences are determined using the Kaplan-Meier estimator. The multiplication of time distributions and rule confidences provides the time-dependent confidence function.

   (b) I presented, how sequential rule mining can be an alternative approach when handling a substantial volume of unique events that

are poorly distributed. Unlike providing a continuation of events, this method identifies sets of antecedent events that may occur in any order before another set of consequent events that also may occur in any order. The determination of temporal characteristics of the resultant rules can be made using the previous approach.

(c) I introduced, how the confidence intervals of the time-dependent confidences can be determined using the bootstrapping method. This involves randomly selecting input sequences and executing the method on this data. The process is executed repeatedly, resulting in a set of confidence functions from bootstraps. The percentile-based method estimates the confidence bounds in this set of functions, thereby establishing the confidence intervals

Related publications: Róbert Csalódi, Zsolt Bagyura, János Abonyi: Time-dependent sequential association rule-based survival analysis: A healthcare application, MethodsX, 2024. [4]
Róbert Csalódi, Zsolt Bagyura, Ágnes Vathy-Fogarassy, János Abonyi: Time-dependent frequent sequence mining-based survival analysis, Knowledge - Based Systems, 2024. [5]

4 Utilization of results

Survival analysis is a statistical methodology and serves as a pivotal tool across diverse fields for analyzing the time to event data. However, it often proves inefficient for exploring in-depth root causes. Therefore, the thesis presented three algorithms that integrated survival analysis with machine learning techniques.

The applicability of the proposed algorithms are demonstrated on diverse case studies. These applications serve as prototypes to establish a forecasting system for healthcare professionals, university managements, governments and manufacturing companies.

Publications by Róbert Csalódi

2024 May
Current h-index: **4**
Current i10-index: **3**
Current citation count: **84**

Publications related to theses

1. R. Csalódi, Z. Bagyura, and J. Abonyi, "Mixture of survival analysis models-cluster-weighted weibull distributions," *IEEE Access*, vol. 9, pp. 152288–152299, 2021
2. R. Csalódi, Z. Birkner, and J. Abonyi, "Learning interpretable mixture of weibull distributions—exploratory analysis of how economic development influences the incidence of covid-19 deaths," *Data*, vol. 6, no. 12, p. 125, 2021
3. R. Csalódi and J. Abonyi, "Integrated survival analysis and frequent pattern mining for course failure-based prediction of student dropout," *Mathematics*, vol. 9, no. 5, p. 463, 2021
4. R. Csalódi, Z. Bagyura, and J. Abonyi, "Time-dependent sequential association rule-based survival analysis: A healthcare application," *MethodsX*, p. 102535, 2024
5. R. Csalódi, Z. Bagyura, Á. Vathy-Fogarassy, and J. Abonyi, "Time-dependent frequent sequence mining-based survival analysis," *Knowledge-Based Systems*, p. 111885, 2024

Further publications

1. T. Ruppert, R. Csalodi, and J. Abonyi, "Estimation of machine setup and changeover times by survival analysis," *Computers & Industrial Engineering*, vol. 153, p. 107026, 2021
2. R. Csalódi, Z. Süle, S. Jaskó, T. Holczinger, and J. Abonyi, "Industry 4.0-driven development of optimization algorithms: A systematic overview," *Complexity*, vol. 2021, pp. 1–22, 2021
3. R. Csalódi, T. Czvetkó, V. Sebestyén, and J. Abonyi, "Sectoral analysis of energy transition paths and greenhouse gas emissions," *Energies*, vol. 15, no. 21, p. 7920, 2022