

Doktori (PhD) értekezés tézisei

**GÉPI TANULÁSSAL TÁMOGATOTT
TÚLÉLÉSELEMZÉS**

Csalódi Róbert

Pannon Egyetem
Vegyéssz mérnöki- és Anyagtudományok
Doktori Iskola

Témavezetők:
Dr. Abonyi János
Dr. Ruppert Tamás



Folyamatmérnöki Intézeti Tanszék
Veszprém
2024.

1. Bevezetés és célkitűzések

A túléléselemzés olyan statisztikai módszer, amelyet egy esemény bekövetkezéséig eltelt idő valószínűségi eloszlásának becslésére használnak. A módszert eredetileg a biostatistikában és az orvosi kutatásokban alkalmazták, ahol a vizsgált mennyiség jellemzően a halál bekövetkezéséig eltelt idő. Ez a módszer azonban mára számos tudományágban elterjedt, mivel a túlélési idők és számos más mennyiség között analógiák állíthatóak fel.

A módszer kiváló eszközként szolgál a folyamatok nem hatékony működésének feltárására. Azonban önmagában gyengének bizonyul, ha a veszteségek kiváltó okait próbáljuk feltárni. Ezért a túléléselemzés és a gépi tanulási módszerek integrált alkalmazása szükséges a mögöttes tényezők kellő mélységben történő feltárásához.

Ez a disszertáció három ilyen integrált algoritmust mutat be, amelyek a túléléselemzést gépi tanulási modellekkel kombinálják, ezzel feltárva a komplex folyamatok mélyebb összefüggéseit. Az első algoritmus egy integrált túléléselemzés és expectation-maximization-alapú klaszterezési keretrendszert mutat be, amely a klasztereket a túlélési idők és a magyarázó változók közötti hasonlóság alapján identifikálja. A második algoritmus egy integrált túléléselemzés és gyakori elemhalmaz-alapú asszociációs szabálybányászati módszert mutat be, amely az időfüggő kategorikus változókból meghatározott releváns kiváltó eseményeket azonosítja, amelyek a versengő kockázatok következményi eseményeihez vezetnek. Az algoritmus a kumulatív előfordulási függvényt közvetlenül a szabályok támogatottságai és a konfidenciák alapján becsüli meg. A harmadik algoritmus egy integrált túléléselemzés és szekvenciális mintázatok bányászati keretrendszert mutat be, amely meghatározza az eseményrákövetkezések időfüggő konfidencia függvényét.

A dolgozat nem csupán az algoritmusok széleskörű felhasználhatóságát mutatja be, hanem rávilágít a bennük rejlő lehetőségekre is, hogy értékes betekintést nyújtsanak, javítsák az előrejelzéseket és fokozzák a döntéshozatali folyamatokat.

2. Kísérleti eszközök és módszerek

A bemutatott algoritmusok hatékonysága különböző esettanulmányokon keresztül került szemléltetésre, rávilágítva a széleskörű alkalmazhatóságukra. A vizsgált példák és a felhasznált változók az alábbiak:

1. A Li-ion akkumulátorok maradék hasznos élettartamának becslése a kapacitásuk, a belső ellenállásuk és a töltési körülmények alapján

2. A prosztatarákos betegek túlélési ideje életkor, szérum hemoglobinszint és kezelés alapján
3. Az országok COVID-19 világjárványhoz kapcsolódó 100.000 lakosra vetített halálozási rátája demográfiai és gazdasági adatok alapján.
4. Az egyetemi hallgatók lemorzsolódása a mintatanterv szerint nem teljesített tantárgyak alapján
5. A jövőbeli betegségek előfordulási esélyei a már meglévő betegségek alapján

Ezek az esettanulmányok különböző területek összetett és dinamikus jellegzetességeit mutatják be, rávilágítva azokra az elemekre, ahol az algoritmusok hozzájárulnak a pontosabb eredmények meghatározásában. Az algoritmusok Matlab és Python környezetben kerültek kifejlesztésre.

3. Tézisek

1. **Kifejlesztettem egy integrált túléléselemzés és expectation-maximization-alapú klaszterezési keretrendszert**
 - (a) Bemutattam, hogyan lehet heterogén túlélési modelleket homogén modellekbe klaszterezni a túlélési idők és a magyarázó változók hasonlósága alapján egy expectation-maximization algoritmus segítségével. Ez a módszer klasztereket határoz meg, és egyidejűleg azonosítja a túlélési valószínűségeiket a Weibull-eloszlás és a kapcsolódó folytonos magyarázó változók valószínűségeit többváltozós Gauss-eloszlás alapján.
 - (b) Bemutattam, hogyan lehet a klasztertagságokat Takagi-Sugeno fuzzy szabályok segítségével reprezentálni, amivel meghatározhatók a folytonos változók működési tartományai. Ezzel a megközelítéssel a túlélési tulajdonságok leírhatók a folytonos változók tartományai alapján. A módszer sokoldalú, és alkalmazható a folytonos változók kategorizálására.

Kapcsolódó publikációk: 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. Kifejlesztettem egy integrált túlélésanalízis és gyakori elemhalmaz - alapú asszociációs szabálybányászati algoritmust

- (a) Bemutattam, hogyan határozható meg a versengő kockázatok valószínűségei egy adott időpillanatban, eseményvezérelt, gyakori elemhalmaz-alapú asszociációs szabályok segítségével. Ez a megközelítés azonosítja az időfüggő kategorikus változókból meghatározott releváns kiváltó eseményeket, amelyek a versengő kockázatokból meghatározott következményi eseményekhez vezetnek. A gyakori elemhalmazok szekvenciája globális, időfüggetlen jellemzőként értelmezhető.
- (b) Bemutattam, hogy egy adott versengő kockázat kumulatív előfordulási függvénye hogyan határozható meg közvetlenül a populációra, akikre egy adott gyakori elemhalmazok szekvenciája jellemző. A módszer a kiválasztott szekvencia összes gyakori elemhalmazát támogató alanyok adathalmazát szegmentálja, és a kumulatív előfordulási függvényt a módosított szabály támogatottságok és konfidenciák alapján becsli.
- (c) Bemutattam, hogyan lehet a hallgatói lemorzsolódást a nem teljesített tantárgyak alapján megbecsülni. A tanulmány felhasználja a mintatantervet, amely előírja az egyes tantárgyak elvégzéséhez ajánlott féléveket. Ha valaki nem tudja teljesíteni ezt a követelményt, az egy tantárgyi hiányt jelent, ami a későbbi lemorzsolódással összefüggő döntő tényező.

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

3. Kifejlesztettem egy integrált túlélésanalízis és szekvenciális mintázatok bányászati módszert

- (a) Bemutattam, hogyan lehet az eseményátmenetek időfüggő támogatottsági és konfidencia függvényeit egy integrált túlélésanalízis és gyakori szekvencia bányászati algoritmus segítségével megbecsülni. A megközelítés a gyakori szekvenciák bányászata révén azonosítja a releváns eseményrácokkövetkezéseket, és meghatározza azok támogatottság és konfidencia értékeit. Az így kapott konfidenciák időbeli dinamikáját a Kaplan-Meier becslő segítségével határozzuk meg. Az időbeli eloszlások és a szabálykonfidenciák szorzata adja az időfüggő konfidencia függvényt.

- (b) Bemutattam, hogy a szekvenciális szabályok bányászata hogyan lehet alternatív megközelítés az aránytalanul eloszló egyedi események kezelésénél. Az események folytatásának megadásával ellentétben ez a módszer az előzményes események olyan halmazait azonosítja, amelyek bármilyen sorrendben előfordulhatnak egy másik, szintén bármilyen sorrendben előforduló következményi eseményhalmaz előtt. Az eredő szabályok időbeli jellemzőinek meghatározása az előző megközelítés alapján végezhető el.
- (c) Bemutattam, hogyan lehet az időfüggő konfidencia függvények konfidenciaintervallumait bootstrapping módszerrel meghatározni. Ennek során véletlenszerűen választunk ki bemeneti szekvenciákat visszatevéssel, és a módszert ezeken az adatokon futtatjuk le. A folyamatot ismételtén végrehajtjuk, aminek eredményeképpen bootstrapekből konfidenciafüggvények halmazát kapjuk. A percentilis alapú módszer megbecsüli a konfidenciahatárokat ebben a függvény értékkészletben, és ezáltal megállapítja a konfidenciaintervallumokat.

Kapcsolódó publikációk: 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. Eredmények gyakorlati hasznosítása

A túlélési elemzés egy statisztikai módszertan, és a különböző területeken kulcsfontosságú eszközként szolgál egy adott esemény bekövetkezéséig eltelt idő jellemzésére. Gyakran azonban nem bizonyul hatékonyak a mélyreható okok feltárására. Ezért a szakdolgozat három olyan algoritmust mutatott be, amelyek a túlélési elemzést gépi tanulási technikákkal integrálták.

A dolgozat a javasolt algoritmusok alkalmazhatóságát különböző esettanulmányokon mutatja be. Ezek az példák pedig prototípusként szolgálnak különböző előrejelző rendszerek létrehozásához az egészségügyi szakemberek, az egyetemi vezetés, a kormányok és a gyártó vállalatok számára.

Csalódi Róbert publikációi

2024 május

Jelenlegi h-index: 4

Jelenlegi i10-index: 3

Jelenlegi hivatkozások száma: 84

Tézisekhez kapcsolódó publikációk

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

További publikációk

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