Opponensi vélemény Csalódi Róbert

„Gépi tanulással támogatott túléléselemzés”

című doktori disszertációjáról

Dr. Kovács Edith Alice

Budapesti Műszaki és Gazdaságtudományi Egyetem

A „*Gépi tanulással támogatott túléléselemzés*” című disszertáció három új algoritmust vezet be, melyek kombinálják a túlélés elemzést és a gépi tanulási módszerekkel kinyert információkat, ezzel jobb betekintést nyerve komplex folyamatokba, ami segít a megértésükben, elemzésükben, modellezésben és perdikcióban.

Az ***első algoritmus*** célja heterogén adatokban lokális csoportok felismerése, a túlélési idők és más folytonos és diszkrét magyarázó változók kimeneteleinek hasonlóságának alapján. Az algoritmus bevezetését a létező Kaplan-Meier és Cox eljárásokkal szembeni előnyeivel indokolják.
A klaszterek megtalálásával egyidejűleg, az egyes klasztereket karakterizáló azonos paraméterezésű túlélési eloszlások is meghatározásra kerülnek. Így egy klaszterbe azon elemek kerülnek majd, amelyeknek a diszkrét és folytonos attribútumai alapján hasonló túlélési tulajdonsággal bírnak.

Az algoritmus az elemeket nem csak különböző klaszterekbe sorolja, hanem a klaszterekhez való hozzátartozást is jellemzi Takagi-Sugeno- féle fuzzy szabályok alapján.
Az algoritmus a klaszterezést kevert eloszlás illesztésével határozza meg, feltételezve, hogy a magyarázó változók feltételesen függetlenek, a klaszterbe tartozást feltétele mellett.

A ***második algoritmus*** a versengő kockázatokat (competing risks) veszi figyelembe, amelyek diszkrét (kategorikus) értékeket felvevő időhöz kapcsolt item- halmazokként jelennek meg. Az algoritmus a gyakran előforduló tételeket összekötő ún. asszociációs szabályokra épül és felismeri a releváns mintázatokat, vagyis adott eseményeket kiváltó releváns eseményeket. Ezen kívül becsli a releváns események alapján a kérdéses esemény bekövetkezésének valószínűségét, amit itt megbízhatóságnak neveznek.

A ***harmadik algoritmus*** is idősorokon dolgozik, de ebben az esetben az idősorban lévő mintázatot keres, úgy nevezett szekvenciákat, mely releváns lesz a következő esemény bekövetkezésére. A módszer jóval túlmutat az eddig alkalmazott módszereken. Egy- egy időablak alatt bekövetkezett események gyakori megfordulásából ad becslést a következő események bekövetkezésre. Külön kiemelendő, hogy a pontos eloszlás ismerete hiánya miatt, bootstrap módszerrel ad konfidencia intervallumot a becslésre.

A disszertáció rendkívül jól felépített, jól követhető. Az új algoritmusok a 3. 4. és 5. fejezetben vannak felvezetve és bemutatva. Külön nagyon tetszik, hogy meg van indokolva, minden esetben, miért van szükség az új algoritmusokra, mi újat adnak, milyen további problémákat kezelnek más, a szakirodalomban alkalmazott algoritmusokhoz képest.

Egy másik nagyon fontos része a dolgozatnak, hogy valós alkalmazásokon szemlélteti az algoritmusokat. Az alkalmazások különböző területekről jönnek (Egészségügy, korházi ellátás, egyetemi- lemorzsolódás illetve fizikában/kémiában), ami úgy szintén alátámasztja az algoritmusok széleskörű lehetséges alkalmazhatóságát.

Kiemelem a fejezetekben szereplő színes folyamábrákat, amelyek érthetővé teszik az algoritmusok menetét.

Az szakirodalomkutatást nagyon alaposnak tartom, 157 cikket/ könyvet idéz, ezek közül 6 komoly cikknek a társszerzője.

Megjegyzések:

- A Jelölt kijavította illetve kipótolta az első két fejezetben fellelt hiányokat/elírásokat, illetve további magyarázatokat fűzött hozzájuk a jobb megértés érdekében.
- A 3. fejezetben a 3.6 és 3.7 képletek esetében nincsen kiemelve, hogy a klasztereken belül feltételes függetlenséget használ a klaszterbe való tartozást feltételezve, holott ezt felhasználja, erre épül az eljárás.
- A 3.4 képlet, nem következik a (3.2) és (3.3)-ból. Itt nyilvánvaló egy elírás történt. Szerintem a 3.2 képletben $p(t)$ van a baloldalon.
- A 3. fejezetben lévő prosztata-rákos példán, a kúrák hatását a klaszterekbe való sorolás elemzéséhez, a Jelölt, kibővítette az Shannon entrópia kiszámolásával (3.23 képletből hiányzik egy egy (-1) szorzó-tényező). Ennek alapján a magas dózis relatív kevésbé hat a klaszterekbe való tartozásra. Ugyanennél a résznél pozitívumként kiemelem, hogy a Jelölt felhívja a figyelmet, a körültekintő értelmezésre, esetleges háttérváltozók hatására.
  Kérdés a Jelölthöz: Milyen érték körüli lenne az entrópia, ha egy tényező felvett értéke nem lenne hatással a klaszterbe való tartozásra, az 5 klaszter esetén?
- A 3. fejezetben lévő Covid-halálozás alkalmazás esetén, a Jelölt készített egy színezett ábrát (3.14 ábra). A feltüntetett számok alapján, amik a 100K emberre jutó halálozások szerepelnek. Észrevehető, hogy a klasztereken belül ezek egyáltalán nem mondhatók homogéneknek, pedig ezt várnánk. Lila klaszter (2-es) van 20 alatti és 200 feletti értékeket is tartalmaz, piros klaszter sem tekinthető homogénnek, mégha Kínát nem is tartjuk mérvadónak.
  Kérdés a Jelölthöz: Mi lehet ennek a magyarázata?
- A 3. fejezet, 19. oldalán, a módszer egy megszorításaként, az is megemlítendő, hogy a figyelembe vett attribútumokról feltételezendő, hogy egymástól független föltéve a klasztert. Megjegyzem, hogy más eljárásokban, mint a naive Bayes is élnek ilyen megszorító feltételezéssel, és hatékony.
  Kérdés a Jelölthöz: Mit gondol, hogyan lehetne azt tesztelni, hogy vajon a figyelembe vett változók elegendők ahhoz, hogy egy hatékony eljárást kapjunk?
- A 4. és 5. fejezethez kért pontosítások megtörténtek.

Összegezve: A disszertációt egy nagyon komoly, szép munkának tartom, szép algoritmusokkal, sokrétű alkalmazásokkal. Fontosnak tartanám a jövőben az algoritmusok által kinyert modellek jóságának a számszerűsítését nem csak AKAIKE vagy loglikelihood alapján, hisz az két modell illeszkedésének összehasonlítására szolgál, illetve az első algoritmus esetében a klaszterek számának meghatározása.

A Jelölt az előzetes kérdésekre a műhelyvitán válaszolt. Hibákat, elírások nagy részét kijavította, további magyarázatokat fűzött, ahol szükség volt rá.

A PhD disszertációt nyilvános vitára javaslom.

Budapest

2024 07 17

Dr. Kovács Edith Alice