DEBRECENI EGYETEM, INFORMATIKAI KAR

4028 Debrecen, Kassai út 26., 4002 Debrecen, Pf. 400.
☎ 52/512-700, 🖃 hivatal@inf.unideb.hu

Bírálói vélemény

Szekér Szabolcs

SUPPORTING DATA ANALYSIS OF RETROSPECTIVE HEALTH EXAMINATIONS WITH DATA SCIENCE METHODS

című PhD disszertációjáról

Bírálói véleményemben először a disszertáció formai jellemzőiről szólok, majd annak tartalmi vonatkozásaira térek ki. Áttekintem továbbá a kapcsolódó tudománymetriai paramétereket, majd összegzem véleményemet. A dolgozattal kapcsolatban felmerült kérdéseimet a szövegben aláhúzással jelzem.

Dolgozatában a szerző az adattudomány területén elért eredményeit mutatja be az egészségügyi adatokból való információkinyerés támogatásához. Az egészségügyi terület egyik fontos kihívása, hogy az adatok gyakran strukturálatlanok. A dolgozat két fő témakörre fókuszál két vonatkozó fejezetben. Az egyik a kontrollcsoportok hatékony kiválasztása, a másik pedig szabad szöveges leletekből történő információkinyerés. A dolgozat a kapcsolódó eredményeket foglalja össze angol nyelven, összesen 113 oldal terjedelemben; a dolgozat mellé angol és magyar nyelvű tézisfüzet is készült. A dolgozat három fejezetből, angol, magyar és német nyelvű absztraktból, valamint irodalomjegyzékből áll. A publikáció eredmények egyszerűbb áttekinthetőségéhez a szerző közleménylistát is beillesztett az egyes fejezetek végére. Igen hasznos, hogy az ábra- és táblázatjegyzék mellé a szerző jelölésjegyzéket is készített.

A bevezetés kellően alapos, mindkét témakör hátterét és a kapcsolódó szakirodalmat jól bemutatja a szerző.

A második fejezet a kontrollcsoport kiválasztásának feladatát helyezi a központba. Ebben a szakaszban három különböző kvantitatív módszert tárgyal a szerző, melyeket az eset- és kontrollcsoportok közötti különbségek objektív értékelésére vesz igénybe. A javasolt módszerek megkülönböztető jellemzője az érzékenység különbsége, melyet szintetikus adatkészleteken történő kísérletekkel validál. Bemutatásra kerül továbbá egy új módszer a kontrollcsoport kiválasztására, mely a súlyozott legközelebbi szomszédság elvén és hibaminimalizáló algoritmusok alkalmazásán alapul. Ezen módszer során a független változók súlyozása logisztikus regressziós illesztés segítségével történik, míg az elemek hatékony párosítása a legközelebbi szomszédok alapján valósul meg. A módszer továbbfejlesztéséhez

szimulált hűtést alkalmaz a szerző a globális optimum kereséséhez. Végül, a hiányzó bináris független változók hatását logisztikus regressziós illesztés segítségével elemzi az esetkontroll-vizsgálatok eredményeire. A második fejezettel kapcsolatos kérdéseim és megjegyzéseim a következők:

- Egyformán hatékonyak-e ezek a mérőszámok mind a kategorikus, mind a folytonos változók esetében, vagy mutatnak-e torzításokat bizonyos adattípusok felé?
- Tekintettel arra, hogy a javasolt mérőszámok lineárisak, hogyan kezelik a csoportokon belüli változók közötti esetleges nem lineáris kapcsolatokat?

A harmadik fejezet kiterjedt elemzéseket végez a kardiológiai szabad szöveges leletekből történő információkinyerés terén alkalmazott természeti nyelvfeldolgozási technikákkal. Az értekezésben alapos megvizsgálásra és összehasonlításra kerülnek a különféle szöveghasonlósági mérőszámok, melyek célja az orvosi kifejezések echokardiográfiás dokumentumokból való hatékony kinyerése. Az elméleti és gyakorlati elemzések során különös figyelmet szentel a szerző az egyes metrikák pontosságának, érzékenységének és alkalmazhatóságának vizsgálatára az adott szakterületen. Ezen összehasonlító elemzések alapján a Jaro-Winkler távolságot határozza meg, mint a leginkább megfelelőt az orvosi kifejezések megbízható azonosítására a kardiológiai leletekben. A fejezetben egy innovatív szövegbányászaton alapuló módszert is javasol a szerző az echokardiográfiás dokumentumokból származó numerikus mérési eredmények kinyerésére. Ez a módszer, bár nyelvfüggetlen, mégis rendkívül hatékonyan képes azonosítani és strukturált formában visszaadni a különböző mérési neveket és eredményeket, miközben rugalmasan kezeli a szinonimákat, rövidítéseket és elírásokat is. Mindezeket figyelembe véve, a harmadik fejezet hasznos eredményeket tartalmaz az orvosi leletek automatizált feldolgozására, és számos érdekes kérdést vet fel az NLP technikák további fejlesztése és alkalmazása tekintetében ezen a specifikus területen. A harmadik fejezettel kapcsolatos kérdéseim és megjegyzéseim a következők:

- Milyen skálázhatósági és teljesítménybeli következményekkel jár ez a módszer, ha nagy mennyiségű orvosi szöveges adatra alkalmazzák, és hogyan viszonyul a hagyományos módszerekhez a feldolgozási sebesség és az erőforrás-felhasználás tekintetében?
- Vannak-e a Jaro-Winkler-távolság használatával kapcsolatos lehetséges korlátok vagy torzítások, különösen a magyar nyelvű orvosi szövegek esetében?
- Mennyire általánosíthatók ezek az eredmények más nyelvekre vagy orvosi területekre?

DEBRECENI EGYETEM, INFORMATIKAI KAR

4028 Debrecen, Kassai út 26., 4002 Debrecen, Pf. 400.
☎ 52/512-700, 🖃 hivatal@inf.unideb.hu

Mind a két vizsgált területnél elmondható, hogy a szerző értékes eredményeket ért el, amiket megfelelő nemzetközi fórumokon is publikált. A vonatkozó tézispontokat megfelelőnek találom. A szerző törekedett rá, hogy eredményeit másokéval megfelelően összehasonlítsa. A dolgozat szerkezeti felépítése és stílusa jó, olvasmányos, a tartalom könnyen követhető.

A szerző összesen hét folyóiratban megjelent idegen nyelvű közleménnyel, valamint hét konferenciakiadványban (három idegennyelvű, négy magyar) megjelent közleménnyel rendelkezik. Mind a folyóiratok, mind a konferenciák között találunk nemzetközileg is színvonalas fórumokat, például D1-es közleményt. Így a disszertáció a publikációk szempontjából eleget tesz az elvárásoknak.

Összefoglalásként, a szerző szakmai munkássága és a dolgozat tartalmi színvonala véleményem szerint teljesíti a PhD fokozatszerzéshez szükséges követelményeket. A disszertáció nemzetközi mércével mérve is új eredményeket tartalmaz, a szerző elegendő mértékben és minőségben publikálta az elért tudományos eredményeket, amelyek a jelölt munkáját képezik. Javaslom az értekezés doktori fokozat megadásának alapjául történő elfogadását és sikeres védés esetén a fokozat megítélését.

Debrecen, 2024. május 07.

.....

Dr. Hajdu András
tszv. egyetemi tanár
Adattudomány és Vizualizáció Tanszék
Debreceni Egyetem, Informatikai Kar