

PhD értekezés bírálata

Szekér Szabolcs:

SUPPORTING DATA ANALYSIS OF RETROSPECTIVE HEALTH EXAMINATIONS WITH DATA SCIENCE METHODS

Készítette: Vassányi István, PhD, PE MIK VIRT

Ez úton szeretném megköszönni a bírálati felkérést. A dolgozat munkahelyi vitára előterjesztett első változatát volt alkalmam már bírálni, és köszönettel vettem, hogy a korábbi megjegyzéseim, felvetéseim nagyrészt megoldásra kerültek a dolgozat végleges változatában. Ezért jelen bíráló csak a teljesen meg nem válaszolt kérdésekkel, illetve javaslatokkal foglalkozik, a szokásos bírálati szerkezetben.

A kutatási terület újdonsága

A bemutatott kutatás témája az adatelemzés, adatfeldolgozás támogatása informatikai módszerekkel, az egészségügy területén. Ez a terület világszerte aktív kutatás tárgyát képezi, és a nemzetközi tudományos együttműködés egyik aktív területe is. A dolgozatban bemutatott két, lazán kapcsolódó tézispont új koncepciókat, módszereket, eszközöket és konkrét eredményeket hoz ezen a területen. *Kétségtelen, hogy a problémafelvetés releváns, és minden eredmény közvetlenül alkalmazható a gyakorlatban.*

A disszertáció szerkesztése és stílusa

Általánosságban elmondható, hogy a szöveg formázása nagyon jó, és a tárgyalásmód megfelel a tudományos jellegű angol szövegek elvárásainak—bár a stílus és különösen a szóhasználat kissé nehézkes, „magyaros” az egész dolgozatban, ez nem rontja az érthetőséget. A nyelvtannal és a szöveg minőségével kapcsolatban nincsenek lényeges problémák. A táblázatok és ábrák könnyen olvashatók és érthetőek, a hivatkozások egyértelműek. A hivatkozások száma és jellege megfelel a PhD-s követelményeknek.

A fejezetek és a tézispontok tartalmi összefoglalása és értékelése

I. fejezet

Az I. fejezet (Introduction) két alfejezetében bemutatja a két tézispont kiindulópontját, eddigi irodalmát, a javasolt megoldás alapgondolatát és az elért eredményeket, tulajdonképpen a disszertáció kiterjesztett absztraktjának is tekinthető. Nagyon jó gondolatnak tartom az eredmények ilyenfajta előzetes összefoglalását.

II. fejezet

A fejezet a fogalmak definiálásával és az eddig publikált módszerek kritikai áttekintésével indul, különös tekintettel a PS illesztésre. Az áttekintés alapos, az értékelő megjegyzések helytállóak, ez a rész jól mutatja a jelölt felkészültségét ezen a területen.

A 2.2 fejezet három új, heurisztikus hasonlósági mértéket javasol a kontroll csoport és a kezelt csoport hasonlóságának mérésére: az NNI és a GDI indexek a páronkénti összehasonlításból vett különbségeket átlagolják, a DDI a hisztogramok hasonlóságát méri. A mértékek tesztelése alapján az NNI mérték az

extrém zajérzékenysége miatt véleményem szerint nem látszik alkalmasnak a halmazok összehasonlítására (tekintve, hogy a zaj természetesen megjelenik minden egészségügyi adatban). A PSM módszer tesztelése az újonnan javasolt mértékekkel megerősíti azt az ismert összefüggést, hogy a caliper érték növelése, bár nagyobb kontroll csoportot eredményez, rontja a kiválasztás minőségét.

A 2.3 fejezet új kontrollcsoport-kiválasztási módszereket javasol, ezek tekinthetők az I. tézispont magjának. Az első módszer, a WNNEM alap gondolata a logisztikus regressziós együtthatókból nyert, a változók relevanciájára jellemző odds értékek felhasználása a többdimenziós térben történő illesztés során a dimenziók súlyozására két vektor távolságának számítására, egy heurisztikus távolságfüggvénnyel (2.18). A módszer az azonos távolságból eredő konfliktust a második legközelebbi szomszéd figyelembe vételével oldja fel. A 2.3.2.1-4 fejezetekben néhány tipikusnak mondható konfigurációban szimulált adatokon mutatja be a dolgozat a WNNEM előnyeit a PSM-hez képest. A WNNSA módszer a konfliktus-feloldásra fejlettebb, távolabbi szomszédokat is figyelembe vevő megoldást (szimulált hűtést) vezet be. A fejezet végén mindeket módszer teljesítményét összeveti az irodalomban használt módszerekkel. Az alapos elemzés eredménye az, hogy a WNNSA, különösen az összes változóra értelmezett eloszlás alapú hasonlóság tekintetében, és különösen a sok konfliktust eredményező adathalmazokon a legjobban teljesítő módszer, tehát a szimulált hűtés alkalmazása célravezető.

A 2.4 fejezet lazán kapcsolódik az eddigiekhez, az ismeretlen változók hatását vizsgálja a kezelt, illetve a kontroll csoportból számított logisztikus regressziós modellek pontosságára. Ezt a problémakört az előzőekhez képest kisebb jelentőségűnek érzem, mivel a gyakorlatban nem mindig tudjuk *a priori*, milyen fontos változók hiányoznak egy adatkészletből (vagy éppen a fontos változók azonosítása az elemzés egyik célja), illetve utólagos pótlásukra általában nincs is lehetőség. Az elemzés eredménye az, hogy szélsőséges esetben igen nagy hibát (torzítást) is okozhat egy hiányzó, lényeges változó.

Megjegyzések:

- A dolgozat a 2.1 fejezetben meggyőzően érvel amellett, hogy több szempontból is jobb a új hasonlósági mértékek, mint az elterjedten használtak, de a pártatlan megítélés érdekében a WNNEM módszert az irodalomban jelenleg leginkább elterjedten használt mérték(ek)kel (lásd 2.1.2 fejezet) is értékelni lehetne, nem csak a Hansen-Bowers teszttel. A WNNSA tesztelése ilyen szempontból korrekt.
- A 2.4 táblázat több teszt-futtatás eredményét mutatja be, melyből a szerző arra következtet, hogy a WNNEM jobban teljesít, mint a PSM. Egy ilyen állítást szokás a két csoport szignifikáns különbségét vizsgáló statisztikai hipotézis-vizsgálattal alátámasztani, vagy legalább a szórásokat megadni.
- A 2. fejezet 3 tipikus konfigurációban szimulált adatokon mutatja be a dolgozat a WNNEM előnyeit a PSM-hez képest, de még sok hasonló konfiguráció elképzelhető lenne. Talán szerencsésebb lett volna a három támogatott adattípusra/eloszlásra (bináris, folytonos normális/egyenletes, nominális) külön-külön elvégezni a tesztet alacsony és magas zaj mellett, mert akkor jobban általánosíthatóak az eredmények.
- Mivel a WNNSA általában véve jobban teljesít, mint a WNNEM, ezért a WNNEM (mint a WNNSA elődje) mint külön módszer ilyen részletes bemutatását, tesztelését akár el is lehetett volna hagyni, illetve az elvégzett tesztek eredményeit elsősorban annak bizonyítására használni, hogy

a szimulált hűtés bevezetése a WNNSA-ban célravezető. Ilyen módon lehetségesnek tartom az 1.2 és 1.3 tézispont összevonását, és a tézispont eredményének tekinteni a legjobb kontrollcsoport-generáló módszer bemutatását.

- A 2.4 fejezetben sok, véletlenszerűen generált adathalmazon végez méréseket, ezáltal az eredmények rendkívül meggyőzőek. A korábbi WNNEM tesztelések során is lehetett volna ezt a megközelítést használni.

Kérdések:

- A 2.2.2.1-ben oldalon javasolt adathalmaz-generátor számára miért ezeket az eloszlásokat választotta? Mennyire tekinthetők ezek az egészségügyben tipikusnak? Veszítenek-e erejükből a későbbi következtetések, ha egy vagy több változó egyik modellezett eloszlást sem követi, hanem például bimodális eloszlású?
- A WNNEM módszer a logisztikus regressziós együtthatókból nyert, a változók relevanciájára jellemző odds értékekre alapul. A változók relevanciájára hasonló helyzetben igen gyakran alkalmazott mérőszám tipikusan az információnyereség (information gain) a változó és az osztálycímké (jelen esetben a halmaztagság) között—mi szólna ennek az alkalmazása ellen?
- Milyen minimális k -értékek fordultak, fordulhatnak elő (2.3 algoritmus) a gyakorlatban? Milyen a 2.3 algoritmus komplexitása/futásideje?
- Megtörtént-e a WNNSA módszer alkalmazása valós kutatásban—ha igen, milyen eredményekkel?

III. fejezet

A III. fejezet a szabad szöveges formátumú szív-UH leletek mérési adatokat tartalmazó részének automatizált feldolgozását tűzi ki célul. Az első bemutatott eredmény a szokásos szöveghasonlósági mértékek értékelő összehasonlítása a célba vett doménben, 10 kiválasztott szakkifejezés alapján. Tekintettel a kézi szakértői kiértékelés szükségességére, a 10 bár alacsony érték, de elfogadható. A tesztek alapján az igaz/hamis pozitív alapú kiértékelés szerint a Jaro-Winkler távolság tűnik a legalkalmasabbnak, legalábbis ezen a mintán. Bár ez a módszer meglehetősen empirikus, azonban ez a megközelítés az adatelemzés szakterületén bevett gyakorlat.

A fejezet, illetve a 2. tézispont fő eredménye a 3.3-ban bemutatott módszer az UH mérési eredmények kinyerésére. A módszer szinonimasztórra alapul, és a szövegbányászatban szokásos lépéseket tartalmazza, azonban a szakterülethez illesztett újdonsága az összetett EPC-k automatizált feloldása, melyet a tézispont legfontosabb új eredményének tartok. A módszer hatékonyságát 100 véletlenszerűen kiválasztott leleten kézzel mérve a 3.3. táblázat foglalja össze. A hamis negatív hibák okának vizsgálata továbbfejlesztési lehetőségeket mutat.

Megjegyzések

- Az eredmények összevetése más publikált módszerekkel nehézségekbe ütközik, hiszen ugyanazon a korpuszon és szótárral kellene futtatni minden összehasonlítani kívánt módszert. A hatékonyság (F1 score stb.) numerikus összevetését ezért nem teljesen tartom helyesnek (92. o). Azonos korpuszon a dolgozatban bemutatott módszer összehasonlítása csak az egyszerű

szövegillesztéses módszerrel történik meg (3.6 táblázat), amely—szöveghasonlósági mértékek híján—nyilván rosszabb eredményre vezet. Az az érv viszont teljesen elfogadható, hogy a problémátér specialitásai nem teszik lehetővé a közvetlen összehasonlítást más módszerekkel.

- Nem teljesen értek egyet azzal az érveléssel, hogy a javasolt módszer nem tételez fel semmit az eredmények rögzítési formátumáról („assumptions about the form in which the results are recorded”). Természetesen, akár az irodalomban javasolt más módszerek, a javasolt módszer is tartalmaz egyfajta a priori, domén-specifikus tudást, sőt anélkül nem is tudna működni: ezek a mértékegységek, a hihetőségi tartományok, illetve a szinonimaszótár, melyhez a készítő feltehetőleg felhasználták a korpuszt. Ez véleményem szerint nem hiba, az adatelemzési módszerek tipikusan konkrét problémára szabottak.

Kérdések

- 83. o. Az összetett szakkifejezés-mérési eredmény szekvenciák feldolgozása pontosan hogyan működik? A például felhozott „ejection fraction Teichholz 56% Simpson 52%” miért nem esik szét már a korábbi lépésben egy EPC-PMR (=56%)-EPC-PMR (=52%) sorozatra?
- A validáció 12 előre meghatározott szakkifejezést azonosítását ellenőrizte. Mit gondol az eredmények általánosíthatóságáról? Hasonló értékeket kapna, ha ugyanezzel a szótárral egy másik magyar kórház leleteit vizsgálná?
- A javasolt módszer egyik alapvető hibaforrása az, ha a szótár hiányos, nem tartalmaz egy fontos szinonimát, mivel ez fals negatív hibához vezethet. Mit gondol, ilyen esetben visszafelé haladva az értékből és a mértékegységből lehetne következtetni?

Összefoglalás

A disszertáció új tudományos eredményeket mutat be az orvosi informatika két fontos területén, amelyek jelenleg is aktív nemzetközi kutatás alatt állnak. A javasolt módszereket a jelölt generált illetve benchmark vagy orvosi gyakorlatból vett adatokon validálta. A tárgyalásmód rendkívül körültekintő, alapos. A két fő tézispont közül az elsőt tartom szakmai szempontból értékesebbnek. Az eredmények meggyőzőek, bizonyítják a javasolt megközelítés és a kidolgozott módszerek életképességét és gyakorlati alkalmazhatóságát, és a területen mérvadó, jelentős nemzetközi tudományos folyóiratokban és konferenciákon is publikálva lettek. A jelölt publikációs tevékenysége példás, hat nemzetközi folyóiratban megjelent publikáció első szerzője.

Összegzésként, a jelen bírálati dokumentumban megadott megjegyzésekkel a téziseket új tudományos eredményként elfogadom, és javaslom a disszertáció nyilvános vitájának megtartását, illetve sikeres vita esetén a PhD fokozat megítélését.

Veszprém, 2024. március



Vassányi István