

Válaszok bírálói kérdésekre

Tisztelt Dr. Vassányi István!

Először is szeretném ismét megköszönni a bírálatát. A szakmai megjegyzései további kutatási irányokat nyithatnak meg számomra a disszertáción túl is.

A bírálatban feltett kérdésekre a következő válaszokat adom.

T1.

K1. “A 20. 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?”

Válasz: A változók számossága és eloszlásuk problémánként és vizsgálat típusonként valóban eltérő lehet. A dolgozatban és a kapcsolódó publikációkban egy olyan komplex és változatos adathalmaz összeállítása volt a cél, amely olyan alapeloszlásokat tartalmaz, amelyek gyakran előfordulnak egészségügyi adatbázisokban. Egy-egy konkrét alkalmazás esetén természetesen a változók számossága és eloszlása is nagymértékben eltérő lehet. Kutatásom elején csak olyan szintetikus adathalmazokkal foglalkoztam, ahol a változók eloszlása azonos típusú volt, ezeket külön-külön vizsgáltam. Mivel az algoritmus hatékonyságát tekintve jelentős eltéréseket nem tapasztaltam, ezért a kutatás következő lépésében vegyes adathalmazokon vizsgáltam tovább az algoritmus működését, amikből egyet be is mutattam a dolgozatban.

Bimodális és multimodális eloszlást sajnos nem vizsgáltam, erre vonatkozóan nincsenek eredményeim. A kérdés felvetése azonban jogos, mivel ilyen eloszlások is előfordulhatnak az egészségügyi adatokban (pl. testmagasság), ezért érdemes lenne a továbbiakban ilyen eloszlások mentén is megvizsgálni az mérőszámok teljesítményét.

K2. “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 osztálycímke (jelen esetben a halmaztagság) között—mi szólna ennek az alkalmazása ellen?”

Válasz: Köszönöm a felvetést, valóban alkalmazhattam volna az information gain-t az egyes változók relevanciájának meghatározására. Az alkalmazásának hátránya

véleményem szerint az lehet, hogy nagyobb súlyt rendel azokhoz a kategorikus változókhoz, amelyek több értékkel rendelkeznek. Ezen attribútumok információnyereségének értéke túlreprezentált lehet. A logisztikus regresszióból kinyert odds értékek esetén ilyen probléma nem áll fenn.

K3. "Milyen minimális k -értékek fordultak, fordulhatnak elő (2.3 algoritmus) a gyakorlatban? Milyen a 2.3 algoritmus komplexitása/futásideje?"

Válasz: A k minimális értéke adathalmaz függő. Az általam megfigyelt értékek 6-14 közé sorolhatók, de ennél nagyságrenddel nagyobb is lehet. Az algoritmus komplexitása k függő ($O(kN)$, $k = 1, \dots, N - 1$), így a legjobb esetben, amikor $k = 1$, a futási idő lineáris ($O(1N)$), a legrosszabb esetben, amikor $k = N - 1$, a futási idő négyzetes ($O((N - 1)N) \sim O(N^2)$).

T2.

K1. „Megtörtént-e a WNNSA módszer alkalmazása valós kutatásban—ha igen, milyen eredményekkel?”

Válasz: Saját kutatás során nem történt meg a módszer alkalmazása, de mások felhasználták a WNNEM és WNNSA módszert is. Pouwels és társai legutóbbi tanulmányukban¹ a WNNEM módszert használták, hogy egészséges résztvevőket válasszanak ki különböző telephelyekről. A módszert a Pachama tanulmány² és egy, a Duisburg-Essen Egyetemen írt disszertáció³ is említi és alkalmazza. A szakmai tanulmányban ismertetett kutatás célja egy dinamikus alapvonal létrehozása volt, amely algoritmikusan kiválaszt egy szabályozási területet megfelelő összehasonlító referenciaként egy karbon projekthez, miközben a disszertáció Chile pénzügyi helyzetét elemzi.

¹ Pouwels, Petra JW, et al. "Global multi-center and multi-modal magnetic resonance imaging study of obsessive-compulsive disorder: Harmonization and monitoring of protocols in healthy volunteers and phantoms." *International journal of methods in psychiatric research* 32.1 (2023): e1931.

² Golmant, Noah, et al. "Pachama Research Brief: A description and initial validation of a dynamic baseline for avoided deforestation projects."

³ Durán Sanhueza, Gonzalo Ignacio. Marginalisation and fragmentation of collective bargaining in Chile. Impacts on workers' power resources and income distribution. Diss. Dissertation, Duisburg, Essen, Universität Duisburg-Essen, 2022.

K2. "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?"

Válasz: A szakkifejezés-mérési eredmény párosok kinyerése több lépésben zajlik. Először az egyszerűbbnek nevezhető kifejezéseket azonosítjuk be a Jaro-Winkler hasonlósági metrika alkalmazásával, majd a be nem azonosított kifejezések további feldolgozása történik meg. Ezen további feldolgozás szakértői szabályrendszer alapján történik. Az összetett szakkifejezés-mérési eredmények feldolgozása során a dolgozatban leírt szabályrendszert alkalmazzuk (72. o.). Ez a működés csupán egy fejlesztői döntés eredménye, korábban is fel lehetett volna bontani az összetett szekvenciákat, amennyiben a szótárt ilyen összetett kifejezésekkel is kiegészítjük és hozzá a felbontási szabályrendszert is definiáljuk. A javasolt módszertanban a szótár azonban csak egyszerű kifejezéseket tartalmaz, melyek felbontása nem szükséges, így felbontási szabályrendszer sincs a szótárban megadott kifejezésekhez definiálva.

K3. "A validáció csak 12 előre meghatározott szakkifejezést azonosítását ellenőrizte. Mit gondol az eredmények általánosíthatóságáról? Hasonló értékeket kapna, ha ugyanezzel a szótárral egy másik magyar kórház leleteit vizsgálná?"

Válasz: A szótár összeállítása a vizsgálatba bevont leletektől függetlenül történt. A szótárat egy olyan szakorvos állította össze, aki nem abban a kórházban dolgozik, ahonnan a leletek származnak, és a leleteket az összeállításhoz nem tanulmányozta. A szótár összeállításának vezérelve az volt, hogy a szakma általános szakkifejezéseit és a szakkifejezésekhez kapcsolódó szinonimákat tartalmazza. Ezen tények ismeretében úgy vélem, hogy az eredmények általánosíthatók. De természetesen előfordulhat egy-egy olyan kórház, ahol olyan szakkifejezést is használnak, melyhez hasonló szakkifejezést nem tartalmaz a szótár. Ebben az esetben célszerű lenne szótárt a hiányzó kifejezéssel bővíteni.

K4. "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?"

Válasz: Amennyiben a metrikák túl alacsonyak, akkor célszerű a dolgozatban bemutatott módon megvizsgálni, hogy mi okozza az alacsony találati arányt (3.2.2-es fejezet). Amennyiben ennek oka a szótár hiányossága, akkor célszerű a szótárt a megfelelő szinonimával bővíteni, s az algoritmust újrafuttatni.

Amennyiben egy orvosi kifejezéshez értelmezhető eredményt tartalmaz a lelet, úgy ezt a bővítési lépést automatizálni is lehetne, de az automatikusan kibővített szótárat mindenképpen validálni kell.

Végül szeretném megköszönni a kérdéseket és még egyszer szeretném megköszönni az értékes szakmai bírálatot.

Veszprém, 2024. június 2.



Szekér Szabolcs