Válaszok bírálói kérdésekre

Tisztelt Prof. Dr. Hajdú András!

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. "Egyformán hatékonyak-e ezek a mérőszámok mind a kategorikus, mind a folytonos változók esetében, vagy mutatnak-e torzításokat bizonyos adattípusok felé?"

Válasz: Ha minden változóhoz azonos súlyt rendelünk, akkor matematikailag kimutatható, hogy a kisebb értékkészlettel rendelkező változók nagyobb hatással vannak a javasolt metrikák értékeire. Ezt a mérőszámok kifejlesztése során is figyelembe vettem, és úgy gondoltam, hogy ez a megfelelő megközelítés: egy bináris változó esetén mért eltérés jelentsen nagyobb különbséget, mint egy folytonos változó esetén mért eltérés (ez csak szélsőséges esetben éri el a változó terjedelmét).

K2. "Tekintettel arra, hogy a javasolt mérőszámok lineárisak, hogyan kezelik a csoportokon belüli változók közötti esetleges nem lineáris kapcsolatokat?"

Válasz: Mindegyik mérőszámot úgy fejlesztettem ki, hogy az egyes változókat egymástól független változóként értelmeztem. Emiatt ha valamely változók függnek egymástól, akkor a változók mentén mérhető eltérést többszörösen vesszük figyelembe. Sajnos igaz, hogy a mérőszámok számítása nem reflektál a változók kimeneti változóra gyakorolt hatásának mértékére, így a javasolt mérőszámok valóban torzíthatnak. Mivel a változók közti kapcsolat a kimeneti változó értékét is többszörösen befolyásolja, úgy vélem, hogy ez nem okoz nagy torzítást. Ugyanakkor köszönöm a felvetést, mert új kutatási kérdést nyitott meg számomra: a változók közti összefüggések hogyan befolyásolják az eset- és kontrollcsoport hasonlóságának megfelelő mérését.

T2.

K1. "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 erőforrásfelhasználás tekintetében?"

Válasz: A módszer futási ideje a korpusz méretével lineárisan arányos, de konstans értéke párhuzamosítással csökkenthető. A módszer Python implementációja egy pipeline-nal dolgozza fel a leleteket, így könnyen párhuzamosítható. A dolgozatban ismertetett esettanulmányt tekintve a javasolt, majd kifejlesztett módszer futási 1 szálon dolgozza fel a leleteket, és így a futási ideje 10 perc nagyságrendű. Sajnos az algoritmus kiértékelése során nem végeztem erőforrás-felhasználással kapcsolatos méréseket.

A hagyományos módszerek futási ideje jelentősen függ a vizsgálandó reguláris kifejezésrendszer komplexitásától, de ezek a módszerek is párhuzamosíthatók.

Összességében úgy vélem, hogy a futási idő mindkét esetben könnyen redukálható, az általam javasolt módszer előnye éppen ezért nem a futási időben, hanem az alkalmazott módszertan korpusz-független jellegében rejlik.

K2. "Vannak-e a Jaro-Winkler-távolság használatával kapcsolatos lehetséges korlátok vagy torzítások, különösen a magyar nyelvű orvosi szövegek esetében?"

Válasz: Mivel a Jaro-Winkler távolság kisebb értéket rendel azokhoz a szavakhoz, amelyek az első X karakterben megegyeznek, így a magyar nyelvben alkalmazott igekötők okozhatnak némi torzítást. A Jaro-Winkler távolság 2 alapvető műveletet különböztet meg: egyezés (matching) és tranzpozíció (transposition), így hasonló alakú, de ellentétes jelentésű szavakkal is problémák lehetnek, pl. le-fel, hipo-hiper (hypo-hyper).

K3. "Mennyire általánosíthatók ezek az eredmények más nyelvekre vagy orvosi területekre?"

Válasz: A módszer alkalmazhatósága leginkább a szótár tartalmától függ, így egy kielégítő szótár összeállítása után azonnal alkalmazható más nyelvekre vagy orvosi területekre is.

A dolgozatban bemutatott hatékonysági eredmények általánosítása már nem ennyire triviális, tekintve, hogy a módszer során alkalmazott Jaro-Winkler távolságot magyar nyelvű szívultrahang leleteken végzett vizsgálat alapján választottam ki. Más témakör esetében érdemes lenne a hasonlóságmértékek összehasonlító vizsgálatát újfent elvégezni, majd a módszertanba a legjobb eredményeket felmutató hasonlósági mértéket integrálni.

Ugyanakkor a korábbi bírálatában javasolt ensemble módszer alkalmazása ezt a manuális lépést kiválthatná, és általánosabb alkalmazhatóságot biztosítana. Bár ebben az esetben az ensemble modell újratanítására lenne szükséges minden korpusz esetében.

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

Veszprém, 2024. június 2.

Szekér Szabolcs