Pannon Egyetem
Műszaki Informatikai Kar
Informatikai Tudományok Doktori Iskola

Tézisfüzet

Szekér Szabolcs

RETROSPEKTÍV EGYÉSZSÉGÜGYI VIZSGÁLATOK ADATELEMZÉSÉNEK TÁMOGATÁSA ADATTUDOMÁNYI MÓDSZEREKKEL

Témavezető
Dr. Fogarassyné dr. Vathy Ágnes

Veszprém
Magyarország
2024

1. Bevezetés

Az egészségügyi adatbázisokban tárolt nagy mennyiségű információ reflektorfénybe helyezi a retrospektív klinikai vizsgálatok nyújtotta lehetőségeket. A nagy mennyiségű adathalmazok feldolgozása azonban sok esetben új módszereket és új algoritmusokat igényel, mivel az egészségügy működésének egyedi jellege és a humán biológiai rendszer összetettsége miatt a adatbányászati módszerek jellemzően csak területspecifikus kiterjesztések után alkalmazhatók. A továbbfejlesztett és az egészségügyhöz igazított adattudományi módszerek hatékonyan járulhatnak hozzá a retrospektív klinikai vizsgálatok megvalósításához, és alapot adhatnak az emberi biológiai rendszer működésének alaposabb megismeréséhez. Ez az új tudás segítheti az orvosokat az egyénre szabott orvoslás megvalósításában.

2. Célok és alkalmazott módszerek

Kutatásom célja olyan új, egészségügyhöz adaptált adattudományi módszerek és algoritmusok kidolgozása volt, amelyek hatékonyan hozzájárulhatnak a nagyméretű (esetenként strukturálatlan) egészségügyi adatállományokból történő információkinyeréséhez és az adatok közt rejlő információk feltárásához.

Kutatásom a következő témákat ölelte fel: új kontrollcsoport-kiválasztási módszerek kidolgozása retrospektív eset-kontroll vizsgálatokhoz; új hasonlósági mértékek kidolgozása a kontrollcsoport-kiválasztás eredményeinek kiértékelésére; a hiányzó változók hatásának elemzése a kontrollcsoport kiválasztási folyamat során; és információk kinyerése nagy, strukturálatlan egészségügyi adathalmazokból.

3. Új tudományos eredmények

Kontrollcsoport-kiválasztás

Tézis 1.1

Három új, kvantitatív különbözőségi mérőszámot javasoltam az eset- és kontrollcsoportok különbözőségének mérésére. A javasolt mérőszámok minden változótípus esetén alkalmazhatók. Közülük kettő a párosított egyedek hasonlósága alapján értékeli az eset- és kontrollcsoportok hasonlóságait, a harmadik pedig a csoportok jellemző jegyeinek eloszlását hasonlítja össze. A javasolt módszerek jellemzőit szintetikus adathalmazokon értékeltem ki. Az eredmények rámutattak arra, hogy az eset- és kontrollcsoportok kiértékelését különböző szempontok szerint kell elvégezni, és a kiértékelések során figyelembe kell venni mind a páronkénti, mind az eloszláson alapuló metrikákat is.

Tézis 1.2

Javaslatot tettem egy új, legközelebbi szomszéd alapú kontrollcsoport kiválasztási módszerre, melynek neve *Weighted Nearest Neighbours Control Group Selection with Error Minimization* (WNNEM). A javasolt módszer a független változók eredeti jellemzőterében számítja ki az egyedek különbözőségét. A különbözőség kiszámítása során az WNNEM módszer az egyes dimenziókat eltérő súllyal veszi figyelembe. Az egyes dimenziók súlyának meghatározása a kimeneti változóra illesztett logisztikus regressziós modell alapján történik. A legközelebbi szomszédok megtalálásához a WNNEM módszer a Vogel-féle közelítés segítségével oldja fel azokat az eseteket, amikor a jelöltcsoport egy egyede az esetcsoport egynél több egyedéhez lenne párosítva. A javasolt WNNEM módszer hatékonyságát benchmark és szintetikus adathalmazokon értékeltem ki. Az eredményeim azt mutatják, hogy a WNNEM módszer kiegyensúlyozottabb kontrollcsoportot képes kiválasztani, mint a Propensity Score Matching legszélesebb körben alkalmazott mohó formája.

Tézis 1.3

Új, szimulált hűtésen alapuló kontrollcsoport kiválasztási módszert javasoltam, melynek neve *Weighted Nearest Neighbour Control Group Selection with Simulated Annealing* (WNNSA). A WNNSA módszer szintén a legközelebbi szomszédok elvén működik, de a WNNEM módszerrel ellentétben nem lokális, hanem globális optimalizáláson alapul. A WNNSA módszer az eset-kontroll párok kiválasztása során szimulált hűtést használ a globális optimum eléréséhez. A javasolt WNNSA módszer hatékonyságát benchmark és szintetikus adathalmazokon szemléltettem. Az eredményeim azt mutatják, hogy a WNNSA módszer kiegyensúlyozottabb kontrollcsoportot képes kiválasztani, mint a WNNEM módszer azokban az esetekben, ha számos konfliktusos helyzet adódik a hasonló egyedek kiválasztási folyamatában.

Tézis 1.4

Logisztikus regressziós illesztéssel elemeztem a hiányzó bináris független változók hatását az eset-kontroll vizsgálatok eredményeire. Monte Carlo szimulációkon alapuló empirikus eredményeim azt mutatják, hogy összefüggés van a hiányzó bináris független változók és a modell pontossága között. A szimulációk kimutatták, hogy a független változók kiválasztása kritikus lépés az eset-kontroll vizsgálatokban, mivel a hiányzó változó által befolyásolt kontrollcsoport döntően befolyásolhatja az elemzési eredményeket.

Információ kinyerése echokardiogramokból

Tézis 2.1

A természetes nyelvfeldolgozás területén gyakorta alkalmazott szöveghasonlósági mérőszámok alkalmazhatóságát vizsgáltam, hogy megállapítsam, mely hasonló-

sági mérőszámok segítségével lehet a legnagyobb bizonyossággal kinyerni orvosi kifejezéseket szívultrahang leletekből. A vizsgált mérőszámok a Longest Common Subsequence, a Levenshtein távolság, a súlyozott Levenshtein távolság, a Jaro-Winkler távolság és a koszinusz távolság voltak. Vizsgálataim során megállapítottam, hogy a magyar nyelven írt szívultrahang leletek esetében a Jaro-Winkler távolság a legalkalmasabb metrika az orvosi szakkifejezések azonosítására.

Tézis 2.2

A szöveghasonlósági mérőszámok összehasonlításának eredményeit felhasználva szövegbányászaton alapuló információkinyerési módszert javasoltam a numerikus mérési eredmények szabad szöveges szívultrahang leletekből történő kinyerésére. A javasolt módszer általánosan alkalmazható, nyelvfüggetlen szövegtisztító előfeldolgozási tevékenységeket végez, automatikusan azonosítja a mérési kifejezéseket és eredményeket, és azokat strukturált formára alakítja át. A módszertan alkalmas a szinonimák, betűszavak és elírások azonosítására, javítására és egységesítésére. Mivel a módszer nem tartalmaz nyelvfüggő elemeket, így bármilyen nyelven írt szívultrahang lelet feldolgozására alkalmas.

A javasolt szövegbányászaton alapuló információkinyerési módszert egy több mint 20 000 szívultrahang leletet tartalmazó dokumentumkészleten értékeltem ki. A kiértékelés során 12 releváns szívultrahang paraméter kinyerésének eredményességét vizsgáltam. A javasolt módszer átlagos érzékenysége 0,904, átlagos specificitása 1,0 és átlagos F1 értéke 0,948-es értéknek adódott a vizsgált dokumentumhalmazon. Az értékelés kellőképpen igazolta a módszer széleskörű alkalmazhatóságát, amit a szakértők is megerősítettek.

4. Az eredmények hasznosulása

A disszertációm új, legközelebbi szomszéd alapú kontrollcsoport kiválasztási módszereket és egy új általános szövegbányászat alapú információkinyerési módszert tartalmaz.

A kidolgozott kontrollcsoport kiválasztási módszerek széles körben alkalmazhatók eset-kontroll vizsgálatokban, a vizsgálati területtől függetlenül. Ezt az állítást bizonyítja, hogy Pouwels és társai legutóbbi tanulmányukban a WNNEM módszert használták, hogy egészséges résztvevőket válasszanak ki különböző telephelyekről [1]. A módszert a Pachama tanulmány [2] és egy, a Duisburg-Essen Egyetemen [3] írt disszertáció 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.

A kifejlesztett információkinyerési módszer a dokumentum nyelvétől függetlenül képes numerikus mérési eredményeket kinyerni a szívultrahang leletekből. A módszer publikálása csupán a közelmúltban történt meg, így alkalmazására

mindeddig nem érkezett említés. Az alkalmazott eszközkészletből adódóan azonban tetszőleges nyelvű szívultrahang feldolgozására alkalmas, így bízom benne, hogy felhasználása a közeljövőben szélesebb körben is megvalósulhat. Mivel azonban a javasolt módszer általános szövegbányászati eljárásokat használ, így alkalmazása nem feltétlen korlátozódik szívultrahang leletek feldolgozására, hanem egyéb területen történő hasznosítása is elképzelhető.

Publikációk

Kontrollcsoport-kiválasztás

P1 Szabolcs Szekér and Ágnes Fogarassyné Vathy. Kontrollcsoport generálási lehetőségek retrospektív egészségügyi vizsgálatokhoz. *Orvosi Informatika 2016 A XXIX. Neumann Kollokvium konferenciakiadványa*, Neumann János Számítógép-tudományi Társaság, pages 135-139, 2016.

P2 Szabolcs Szekér, György Fogarassy, and Ágnes Vathy-Fogarassy. Comparison of control group generating methods. *Studies in Health Technology and Informatics*, Vol. 236, pages 311-318, 2017. (Q3)

P3 Szekér Szabolcs, Fogarassyné Vathy Ágnes. Látens változók hatása dichotom kimenetű vizsgálatok kiértékelésére. *Orvosi Informatika 2018 A XXXI. Neumann Kollokvium konferencia-kiadványa*, Neumann János Számítógép-tudományi Társaság, pages 37-42, 2018.

P4 Szabolcs Szekér and Ágnes Vathy-Fogarassy. The effect of latent binary variables on the uncertainty of the prediction of a dichotomous outcome using logistic regression based propensity score matching. *Studies in Health Technology and Informatics*, Vol. 248, pages 1-8, 2018. (Q3) *Best PhD Paper Award*

P5 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Measuring the similarity of two cohorts in the n-dimensional space. *The 11th Conference of PhD Students in Computer Science: Volume of short papers CS2*, pages 151-154, 2018.

P6 Szekér Szabolcs, Fogarassyné Vathy Ágnes. Kontrollcsoport kiválasztása súlyozott k-nn módszer alkalmazásával. *Orvosi informatika A XXXII. Neumann Kollokvium konferencia-kiadványa*, Neumann János Számítógép-tudományi Társaság, pages 7-12, 2019.

P7 Szabolcs Szekér and Ágnes Vathy-Fogarassy. How can the similarity of the case and control groups be measured in case-control studies? *Proceedings of IEEE International Work Conference on Bioinspired Intelligence IWOBI 2019*, IEEE, pages 33-40, 2019.

P8 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Weighted nearest neighbours-based control group selection method for observational studies. *Plos One*, 15(7): e0236531, 2020. (D1, IF: 3.24)

P9 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Optimized weighted nearest neighbours matching algorithm for control group selection. *Algorithms*, 14(12): 356, 2021. (Q2)

Információ kinyerése szívultrahang leletekből

P10 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Application of Text Mining Methods on Unstructured Hungarian Echocardiogram Documents. *Proceedings of the Pannonian Conference on Advances in Information Technology (PCIT 2019)*, University of Pannonia, pages 187-193, 2019.

P11 Szabolcs Szekér, György Fogarassy, Károly Machalik, and Ágnes Vathy-Fogarassy. Application of named entity recognition methods to extract information from echocardiography reports. *Studies in Health Technology and Informatics*, Vol. 260, pages 41–48, 2019. (Q3)

P12 Ágnes Vathy-Fogarassy, Szabolcs Szekér, Balázs Szolár, and György Fogarassy. The efficiency of different distance metrics for keyword-based search in medical documents: A short case study. *Studies in Health Technology and Informatics*, Vol. 271, pages 232–239, 2020. (Q3)

P13 Szabolcs Szekér, György Fogarassy, and Ágnes Vathy-Fogarassy. A general text mining method to extract echocardiography measurement results from echocardiography documents. *Artificial Intelligence in Medicine*, 143: 102584, 2023. (D1, IF: 7.5)

Absztraktok

Kontrollcsoport-kiválasztás

A1 Szekér Szabolcs, Ágnes Vathy-Fogarassy. Novel k Nearest Neighbour-based Control Group Selection Methods. *13th Miklós Iványi International PhD & DLA Symposium - Abstract Book: Architectural, Engineering and Information Sciences*, Pollack Press, page 124, 2017.

MTMT profil

https://m2.mtmt.hu/gui2/?type=authorsmode=browsesel=authors10063045

Hivatkozások

[1] Petra JW Pouwels, Chris Vriend, Feng Liu, Niels T de Joode, Maria CG Otaduy, Bruno Pastorello, Frances C Robertson, Ganesan Venkatasubramanian, Jonathan Ipser, Seonjoo Lee, 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):e1931, 2023.

[2] Noah Golmant, Martha Morrissey, Carlos Silva, Felix Dorrek, Bernhard Stadlbauer, Rachel Engstrand, and Dick Cameron. Pachama research brief: A description and initial validation of a dynamic baseline for avoided deforestation projects.

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