University of Pannonia
Faculty of Information Technology
Doctoral School of Information Science and Technology

Thesis Book

Szabolcs Szekér

SUPPORTING DATA ANALYSIS OF RETROSPECTIVE HEALTH EXAMINATIONS WITH DATA SCIENCE METHODS

Supervisor
Dr. Ágnes Vathy-Fogarassy, PhD

Veszprém
Hungary
2024

1 Introduction

The large amount of information stored in health databases shines a spotlight on the possibilities offered by retrospective clinical studies. However, the processing of large amounts of data sets requires new methods and new algorithms in many cases, since due to the unique nature of the operation of healthcare and the complexity of the human biological system, data mining methods can typically only be applied after area-specific extensions. Advanced data science methods adapted to health care can effectively contribute to the implementation of retrospective clinical studies and can provide a basis for a more thorough understanding of the functioning of the human biological system. This new knowledge can help doctors implement personalised medicine.

2 Goals and applied methods

The aim of my research was to develop such new healthcare-adapted data science methods and algorithms, which can effectively contribute to the extraction of information from large (sometimes unstructured) healthcare data files and to the discovery of the information hidden in the data.

My research covered the following topics: development of new control group selection methods for retrospective case-control studies; developing new similarity measures for evaluating the results of the control group selection; analysing the effect of missing variables during the control group selection process; and extracting information from large, unstructured healthcare datasets.

3 New scientific results

Control group selection

Thesis 1.1

I proposed three quantitative dissimilarity measures to measure the dissimilarity of case and control groups regardless of the types of variables. Two of them evaluate the similarities of case and control groups based on the similarities of the paired individuals, and the third one compares the distributions of the characteristic features of the groups. The characteristics of the proposed methods was shown on synthetic datasets. All proposed measures are linear but their responsiveness is different. Results pointed out the fact that evaluating case and control groups must be made from different aspects, using both pairwise and distribution-based measures.

Thesis 1.2

I proposed a novel nearest neighbour-based control group selection method called *Weighted Nearest Neighbours Control Group Selection with Error Minimization* (WNNEM). The proposed method calculates the dissimilarities of the

individuals in the original feature space of the independent variables. The independent variables are weighted based on a logistic regression-fit. For finding the nearest neighbours, WNNEM uses Vogel's approximation to solve such cases where an individual of the candidate group would be paired to more than one individual of the case group. The effectiveness of the WNNEM method was evaluated on benchmark and synthetic datasets. Evaluation results showed that the proposed WNNEM method is able to select a more balanced control group than the most widely applied greedy form of the propensity score matching method.

Thesis 1.3

As the previously developed WNNEM method utilises local optimisation, I proposed a novel simulated annealing-based control group selection method called *Weighted Nearest Neighbour Control Group Selection with Simulated Annealing* (WNNSA). The WNNSA method utilises simulated annealing to achieve a global optimum during control group selection to find the nearest neighbours. The effectiveness of the WNNSA method was evaluated on benchmark and synthetic datasets. Evaluation results showed that the proposed WNNSA method is able to select a more balanced control group than the WNNEM method if numerous conflicted situations arise in the selection process of similar individuals.

Thesis 1.4

I analysed the effect of missing binary independent variables on the results of case-control studies using logistic regression-fit. Using Monte Carlo simulations, my empirical results showed that there is a correlation between missing binary independent variables and the model accuracy. The Monte Carlo simulations revealed, that the selection of independent variables is a critical step in case-control studies as a biased control group in regard to the missing variable may crucially affect the analyses results.

Information extraction from echocardiography documents

Thesis 2.1

I examined and compared different text similarity metrics applied in the field of NLP to determine which similarity metrics present the highest gain in terms of extracting medical terms from echocardiography documents. The examined metrics were the following: Longest Common Subsequence, Levenshtein distance, weighted Levenshtein distance, Jaro-Winkler distance and cosine distance. I established that the Jaro-Winkler distance is the most suitable to identify medical terms in echocardiography documents written in Hungarian language.

Thesis 2.2

By utilising the findings of the comparison of different text similarity metrics, I proposed a text mining-based information extraction method to extract numerical measurement results from echocardiography documents. The proposed method performs generally applicable, language-independent text-cleaning preprocessing activities, automatically identifies measurement names and results, and returns them in a structured way. The methodology is also able to identify, correct and unify synonyms, acronyms, and typos. Since the method does not contain any language-dependent implementation elements, it is suitable for processing echocardiography findings written in any language.

The proposed text mining-based information extraction method was evaluated on a document set containing more than 20,000 echocardiography reports. During the evaluation, 12 relevant echocardiography parameters were extracted from the documents. As a result, an average sensitivity of 0.904, an average specificity of 1.0 and an average F1 score of 0.948 were obtained. The evaluation sufficiently demonstrated the broad applicability of the method, also confirmed by the experts.

4 Utilisation of results

My dissertation includes novel nearest neighbour-based control group selection methods and a novel general text mining-based information extraction method.

The developed control group selection methods can be widely used in case-control studies, irrespective of the field of the study. This statement is evidenced by the fact, that in a recent study, Pouwels et al used the WNNEM method to select healthy participants from different sites [1]. The method is also mentioned and applied in the Pachama study [2] and in a dissertation [3] written at the University of Duisburg-Essen. The purpose of the research described in the professional paper was to create a dynamic baseline that algorithmically selects a regulatory area as an appropriate comparative reference for a carbon project, while the dissertation analyses the financial situation of Chile.

The developed information extraction method is able to extract numerical measurement results from the echocardiogram reports, regardless of the language of the document. The method was published only recently, so its application has not been mentioned until now. Due to the set of tools used, however, it is suitable for processing echocardiograms in any language, so I am confident that its use can be implemented on a wider scale in the near future. However, since the proposed method uses general text mining procedures, its application is not necessarily limited to the processing of echocardiogram reports, but its application in other areas is also conceivable.

Publications

Control group selection

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 Számítógéptudományi 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)

Information extraction from echocardiography documents

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)

Abstracts

Control group selection

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 profile

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

References

[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.