POKROČILÉ TECHNOLÓGIE SPRACOVANIA A ANALÝZY VEĽKÝCH DÁT

EDITOR

Adam DUDÁŠ

Autori

Mihaela Tinca Udriştioiu

Adam Dudáš

Alžbeta Michalíková

Fatih Kilic

Onder Tutsoy

Jarmila Škrinárová

Silvia Puiu

Slaveya Petrova

Táto publikácia bola financovaná Európskou komisiou v rámci projektu

Erasmus+ Uplatňovanie niektorých pokročilých technológií vo výučbe a výskume v súvislosti so skúmaním znečistenia ovzdušia

Kód projektu: 2021-1-RO01-KA220-HED-00003028

Podpora Európskej komisie na vydanie tejto publikácie nepredstavuje schválenie jej obsahu, ktorý vyjadruje len názory autorov, a národná agentúra a Európska komisia nenesú zodpovednosť za akékoľvek použitie informácií v nej obsiahnutých.

Financované Európskou úniou

University of Craiova

University of Plovdiv "Paisii Hilendarski"

Adana Alparslan Türkeş Science and Technology University

Matej Bel University Banská Bystrica

ISBN
978-625-399-465-5

Book Title
Pokročilé technológie spracovania a analýzy veľkých dát

Editor
Adam DUDÁŠ
ORCID iD: 0000-0001-5517-9464

Project manager
Mihaela Tinca UDRISTIOIU
ORCID iD: 0000-0002-5811-5930

Publishing Coordinator
Yasin DİLMEN

Page and Cover Design
Akademisyen Dizgi Ünitesi

Publisher Certificate Number
47518

Printing and Binding
Vadi Matbaacılık

Bisac Code
BUS070030

DOI
10.37609/akya.2892

Library ID Card
Tınca Udrıstıoıu, Mihaela and others.
Pokročilé technológie spracovania a analızy veľkıch dát / Mihaela Tınca Udrıstıoıu, Adam Dudaš, Alžbeta Michalikova [and others] ; editör : Adam Dudas.
Ankara : Akademisyen Yayınevi Kitabevi, 2023.
172 page. : figure, table. ; 195x275 mm.
Includes Bibliography.
ISBN 9786253994655
1. Information Technology.

GENERAL DISTRIBUTION
Akademisyen Kitabevi A.Ş.
Halk Sokak 5 / A Yenişehir / Ankara
Tel: 0312 431 16 33
siparis@akademisyen.com

www.akademisyen.com

OBSAH

ÚVOD

Táto učebnica predstavuje jeden z výsledkov dosiahnutých v rámci Erasmus+ projektu číslo 2021-1-RO01-KA220-HED-000030286 s názvom "Uplatňovanie niektorých pokročilých technológií vo výučbe a výskume v súvislosti so skúmaním znečistenia ovzdušia". Na dosiahnutí tohto cieľa spolupracovali štyri partnerské organizácie: Univerzita Mateja Bela v Banskej Bystrici zo Slovenska, Univerzita v Craiove z Rumunska, Univerzita Paisij Chilendarského v Plovdive z Bulharska a Adanská univerzita vied a technológií z Turecka. Cieľom učebnice je pomôcť inštruktorom STEM predmetov zlepšiť zručnosti študentov pri práci s dátami.

Sme zahltení informáciami, ktoré sú okolo nás. Pre získanie informácií relevantných pre každý zvolený cieľ skúmania je v dnešnej dobe potrebné vedieť spracovávať dáta. Počítače, senzorové siete a satelity každú sekundu zhromažďujú milióny hodnôt rôznych fyzikálnych alebo iných veličín a parametrov. Databázy uchovávajú a organizujú údaje a informácie, čím zlepšujú kvalitu dát. Keďže dôležitosť informácií rastie viac ako kedykoľvek predtým, študenti STEM predmetov sa musia naučiť pracovať s údajmi. Moderné spoločnosti vyžadujú vysokoškolské vzdelanie, aby poskytovali vysokokvalifikovaných absolventov schopných riešiť problémy na základe informácií získaných zo špecializovaných databáz alebo pomocou programov či algoritmov. Na univerzitách by STEM študenti mali študovať, ako sa množiny dát zhromažďujú, analyzujú a interpretujú – činnosti, ktoré pomôžu pri klasifikácii a aproximácii údajov a pri vytváraní kvalitných odhadov. Nakoniec, trh práce žiada absolventov STEM, aby vytvárali predpovede toho, ako sa procesy vyvíjajú v priestore a čase, alebo aby robili dôležité rozhodnutia. Strojové učenie sa a umelá inteligencia sú štandardné pojmy v každodennej slovnej zásobe študentov.

Táto učebnica obsahuje jedenásť častí, niekoľko príloh a zoznam literatúry relevantnej pre opisované oblasti analýzy dát. Prvá časť učebnice je zameraná na rôzne typy dát, ich vlastnosti, metódy vzorkovania dát a spôsob spracovania a analýzy dát. Nasledujúce časti približujú jeden z najvýznamnejších procesov súvisiacich s veľkými dátami, analýzu dát. Pri analýze veľkých dát je potrebné vedieť používať vhodné metódy štatistickej analýzy, vizualizáciu dát a ďalšie exploratívne, prediktívne a odhadovacie metódy. Jednotlivé sekcie učebnice sa zameriavajú na prístupy, ako je strojové učenie sa, fuzzy inferencia a systémy využívajúce neurónové siete. Prílohy učebnice obsahujú opis datasetu Iris, príklady riešení niektorých problémov, opis datasetov o klimatických zmenách alebo znečistení ovzdušia a informácie o vplyve znečistenia ovzdušia na ľudské zdravie. Príručku uzatvára príklad učebného plánu pre kurz "Pokročilé technológie spracovania a analýzy veľkých dát".

ZOZNAM BIBLIOGRFICKÝCH ODKAZOV

References for sections 1 – 4:

- C.J. Date. An Introduction to Database Systems (8th. ed.). Addison-Wesley Longman Publishing Co., 2003. ISBN: 978-0-321-19784-9
- Felix Kutsanedzie, Sylvester Achio, Edmund Ameko. Practical Approaches to Measurements, Sampling Techniques and Data Analysis. Science Publishing Group, 2016. ISBN: 978-1-940366-58-6.
- William J. Lammers, Pietro Badia. Fundamentals of Behavioral Research Textbook. Online: https://uca.edu/psychology/fundamentals-of-behavioral-research-textbook/
- Jimin Quian et al. Introducing self-organized maps (SOM) as a visualization tool for materials research and education. Results in Materials, Volume 4, 2019, ISSN 2590-048X.
- Naseer Raheem. Big Data: A tutorial-based approach. Chapman and Hal-l/CRC, 2019. ISBN: 978-0-367-67024-5
- Lior Rokach, Oded Maimon. Data mining with decision trees. 2015.
- Steven S. Skiena. The Data Science Design Manual. Springer, 2017. ISBN: 978-3-319-55443-3
- Karthik Ramasubramanian, Abhishek Singh. Machine Learning Using R. Springer, 2019. ISBN: 978-1-4842-4214-8
- Patrik Očenáš. Parallel and distributed methods of big data sampling (in Slovak). 2023.
- Bianka Modrovičová. Decision trees for sizable graph datasets (in Slovak). 2023.
- Aneta Szoliková. Explorative data analysis in document databases (in Slovak). 2023.
- Adam Dudáš, Bianka Modrovičová. Decision Trees in Proper Edge k-coloring of Cubic Graphs. In Proceedings of 33rd FRUCT conference. 2023.

References for sections 5 – 8:

- ZADEH, L. A. Fuzzy Sets. In: Information and Control, 8, 1965, 338-353.
- MICHALÍKOVÁ, A.: Fuzzy množiny v informatike. rec. Mirko Navara, Martin Kalina, Martin Klimo. Belianum. Matej Bel University in Banská Bystrica, 1, 2020, 206p. ISBN 978-80-557-1707-4
- Sendai Subway. Japan Visitor [cit. 2023-02-02]. Online: https://www.japanvisitor.com/japan-transport/sendai-subway
- RUAN D.: Fuzzy Logic Applications in Nuclear Industry. Fuzzy Logic Foundations and Industrial Applications. 1996, 8, ISBN 978-1-4612-8627-1.
- TAKAGI, T., SUGENO, M. Fuzzy Identifications of Fuzzy Systems and its Applications to Modelling and Control. In: IEEE Transactions on Systems, Man, and Cybernetics, 15(1), 1985, 116-132.
- ROSS, T. J. Fuzzy Logic with Engineering Applications. John Wiley & Sons, 2005, 585s., ISBN 9780470743768.
- ZADEH, L. A., The Concept of a Linguistic Variable and its Application to Approximate Reasoning - 1, In: Information Sciences, 8, 1975, 199–249.

References for sections 9

- Ahmed, Z. H. (2010). Genetic algorithm for the traveling salesman problem using sequential constructive crossover operator. International Journal of Biometrics & Bioinformatics (IJBB), 3(6), 96.
- Aktaş, M., Yetgin, Z., Kılıç, F., & Sünbül, Ö. (2022). Automated test design using swarm and evolutionary intelligence algorithms. Expert Systems, 39(4), e12918.
- Bartz-Beielstein, T., Branke, J., Mehnen, J., & Mersmann, O. (2014). Evolutionary algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(3), 178-195.
- Bertsimas, D., & Tsitsiklis, J. (1993). Simulated annealing. Statistical science, 8(1), 10-15.
- Blickle, T. (2000). Tournament selection. Evolutionary computation, 1, 181-186.
- Cui, Y., Geng, Z., Zhu, Q., & Han, Y. (2017). Multi-objective optimization methods and application in energy saving. Energy, 125, 681-704.
- De La Iglesia, B. (2013). Evolutionary computation for feature selection in classification problems. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(6), 381-407.
- Gaivoronski, A. A., Lisser, A., Lopez, R., & Xu, H. (2011). Knapsack problem with probability constraints. Journal of Global Optimization, 49, 397-413.
- Glover, F., & Laguna, M. (1998). Tabu search (pp. 2093-2229). Springer US.
- Hansen P, Mladenovi'c N (1999) An introduction to variable neighborhood search. In: Voß S, Martello S, Osman IH, Roucairol C (eds) Metaheuristics: advances and trends in local search paradigms for optimization, chapter 30. Kluwer Academic Publishers, Dordrecht, pp 433–458
- Hayyolalam, V., & Kazem, A. A. P. (2020). Black widow optimization algorithm: a novel meta-heuristic approach for solving engineering optimization problems. Engineering Applications of Artificial Intelligence, 87, 103249.

- Hinson, J. M., & Staddon, J. E. R. (1983). Matching, maximizing, and hill-climbing. Journal of the experimental analysis of behavior, 40(3), 321-331.
- Holland JH. Outline for a logical theory of adaptive systems. J ACM. 1962;9(3):297–314
- Holland, J. H. (1973). Genetic algorithms and the optimal allocation of trials. SIAM journal on computing, 2(2), 88-105.
- Hoos, H. H., & Stützle, T. (2004). Stochastic local search: Foundations and applications. Elsevier.
- I. Rechenberg, Cybernetic solution path of an experimental problem. Royal Air-craft Establishment, Library Translation 1122, Farnborough, Reprint in: D.B. Fogel (Ed.), Evolutionary Computation, The Fossil Record, IEEE Press, Piscataway, NJ, 1965, pp. 301–309
- I. Rechenberg, Evolutionsstrategie—Optimisierung technischer Systeme nach Prinzipien der biologischen Evolution, Frommann-Holzboog, Stuttgart, 1973
- Kılıç, F., Yılmaz, İ. H., & Kaya, Ö. (2021). Adaptive co-optimization of artificial neural networks using evolutionary algorithm for global radiation forecasting. Renewable Energy, 171, 176-190.
- Kiliç, F., & Gök, M. (2013). A public transit network route generation algorithm. IFAC Proceedings Volumes, 46(25), 162-166.
- Li, X., Tang, K., Omidvar, M. N., Yang, Z., Qin, K., & China, H. (2013). Benchmark functions for the CEC 2013 special session and competition on large-scale global optimization. gene, 7(33), 8.
- Mirjalili, S. (2016). SCA: a sine cosine algorithm for solving optimization problems. Knowledge-based systems, 96, 120-133.
- Rossi, F., Van Beek, P., & Walsh, T. (Eds.). (2006). Handbook of constraint programming. Elsevier.
- Salkin, H. M., & De Kluyver, C. A. (1975). The knapsack problem: a survey. Naval Research Logistics Quarterly, 22(1), 127-144.
- Sharifi, A. A., & Aghdam, M. H. (2019). A novel hybrid genetic algorithm to reduce the peak-to-average power ratio of OFDM signals. Computers & Electrical Engineering, 80, 106498.
- Wang, L., Cao, Q., Zhang, Z., Mirjalili, S., & Zhao, W. (2022). Artificial rabbits optimization: A new bio-inspired meta-heuristic algorithm for solving engineering optimization problems. Engineering Applications of Artificial Intelligence, 114, 105082.
- Yang, J., & Soh, C. K. (1997). Structural optimization by genetic algorithms with tournament selection. Journal of computing in civil engineering, 11(3), 195-200.

References for section 10:

- Basic Neural Networks 1 - https://docs.google.com/a/atu.edu.tr/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxpaHNhbnlhc3NpbjJ8Z3g6NGY4MjNjN2Y4ZTdhNWM2MQ
- Basic Neural Networks 2 - http://www.cs.stir.ac.uk/courses/ITNP4B/lectures/
- Basic Neural Networks 3

  https://www.cs.bham.ac.uk/~jxb/inn.html
- Basic Neural Network 4

  https://www.fer.unizg.hr/en/course/neunet_a/lecture_notes
- Basic Neural Network 5

  http://users.monash.edu/~cema/courses/FIT3094/lecturePDFs/

References for section 11:

- Paluszek, M., Thomas, S. Matlab machine learning recepies. 2019. Plainsboro, NJ, USA. ISBN-13 (pbk): 978-1-4842-3915-5. DOI 10.1007/978-1-4842-3916-2.
- Kim, P. MATLAB Deep Learning. With Machine Learning, Neural Networks and Artificial Intelligence. 2017. Apress Korea ISBN-13 (pbk): 978-1-4842-2844-9. DOI 10.1007/978-1-4842-2845-6.
- Get Started with Matlab. https://www.mathworks.com/help/matlab/getting-started-with-matlab.html
- Iris Clustering. https://www.mathworks.com/help/deeplearning/ug/iris-clustering.html

References for Appendices:

- Fisher, R.A. (1936) "The use of multiple measurements in taxonomic problems". Annual Eugenics, 7, Part II, pages 179-188
- Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule". IEEE Transactions on Information Theory, May 1972, pages 431-433
- Duda, R.O., Hart, P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1, page 218
- Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recogni-

tion in Partially Exposed Environments". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 1, pages 67-71
- https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-3/data-products
- https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-4/data-products
- https://climexp.knmi.nl/
- https://www.uradmonitor.com/
- Velea L, Udriștioiu MT, Puiu S, Motișan R, Amarie (2023)D. A Community-Based Sensor Network for Monitoring the Air Quality in Urban Romania. Atmosphere; 14(5):840. https://doi.org/10.3390/atmos14050840
- https://bookdown.org/floriandierickx/bookdown-demo/climate-data-from-models.html#differences-between-climate-projections-predictions-and-scenarios
- https://ec.europa.eu/eurostat/web/climate-change/database
- https://ourworldindata.org/
- https://ourworldindata.org/data-review-air-pollution-deaths
- •https://ourworldindata.org/outdoor-air-pollution#outdoor-air-pollution-deaths-by-age
- •https://www.who.int/health-topics/air-pollution#tab=tab_1
- •https://www.eea.europa.eu/en/topics/in-depth/air-pollution
- •https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants
- https://www.who.int/publications/i/item/9789240034228
- •https://apps.who.int/iris/bitstream/handle/10665/345329/9789240034228-eng.pdf
- EEA, 2012, The contribution of transport to air quality, EEA Report no. 10/2012, European Environment Agency.
- EEA. A closer look at urban transport TERM 2013: transport indicators tracking progress towards environmental targets in Europe EEA Report No 11/2013 Copenhagen, ISSN 1725-9177.
- http://dx.doi.org/10.1016/j.envpol.2007.06.012
- https://www.who.int/health-topics/air-pollution#tab=tab_1
- Report no. 05/2022, Air quality in Europe 2022. doi: 10.2800/488115. https://www.eea.europa.eu/publications/air-quality-in-europe-2022
- Xin Zhang, X. Chen, Xiaobo Zhang. The impact of exposure to air pollution on cognitive performance. Proc. Natl. Acad. Sci. Unit. States Am., 115 (2018), pp. 9193-9197, 10.1073/pnas.1809474115
- J. Currie, J.S.G. Zivin, J. Mullins, M.J. Neidell. What do we know about short and long term effects of early life exposure to pollution? NBER Work. Pap., 6 (2013), pp. 217-247, 10.3386/w19571
- Escamilla-Nuñez M-C., Barraza-Villarreal A., Hernandez-Cadena L., Moreno-Macias H., Ramirez-Aguilar M., Sienra-Monge J-J., Cortez-Lugo M., Texcalac J-L., del Rio-Navarro B., Romieu I. Traffic-Related Air Pollution and Respiratory Symptoms Among Asthmatic Children, Resident in Mexico City: The EVA Cohort Study. http://www.medscape.com/viewarticle/585875.
- Juvin P., Fournier T., Boland S. et al. Diesel particles are taken up by alveolar type II tumor cells and alter cytokines secretion. Arch Environ Health. 2002; 57(1):53-60.
- Le Tertre A., S. Medina, E. Samoli et al: Short term effects of particulate air pollution on cardiovascular disease in eight European cities. J. Epidemiol Community Health, 2002; 56, (10):773-9.
- Nordling E., Berglind N., Melén E., Emenius G., Hallberg J., Nyberg F., Pershagen G., Svartengren M., Wickman M., Bellander T. Traffic related air pollution and childhood respiratory symptoms, function and allergies. Epidemiology. 2008; 19(3):401-8.
- Pan G., Zhang S., Feng Y., Takahashi K., Kagawa J., Yu L., Wang P., Liu M., Liu Q., Hou S., Pan B., Li J. Air pollution and children's respiratory symptoms in six cities of Northern China. Respiratory Medicine 2010;104(12):1903–11.
- Richardson E.A., Pearce J., Tunstall H., Mitchell R., Shortt N.K.: Particulate air pollution and health inequalities: a Europe-wide ecological analysis. Int J Health Geogr 2013;12:34
- I. Jáuregui, J. Mullol, I. Dávila, M. Ferrer, J. Bartra, A. Del Cuvillo, J. Montoro, J. Sastre, A. Valero. Allergic rhinitis and school performance. J Investig. Allergol. Clin. Immunol., 19 (2009), pp. 32-39
- D.P. Skoner. Allergic rhinitis: definition, epidemiology, pathophysiology, detection, and diagnosis. J. Allergy Clin. Immunol., 108 (2001), pp. 2-8, 10.1067/mai.2001.115569
- I. Beck, S. Jochner, S. Gilles, M. McIntyre, J.T.M. Buters, C. Schmidt-Weber, H. Behrendt, J. Ring, A. Menzel, C. Traidl-Hoffmann. High environmental ozone levels lead to enhanced allergenicity of birch pollen. PloS One, 8 (2013), 10.1371/journal.pone.0080147
- P. Sturdy, S. Bremner, G. Harper, L. Mayhew, S. Eldridge, J. Eversley, A. Sheikh, S. Hunter, K. Boomla, G. Feder, K. Prescott, C. Griffiths. Impact of asthma on educational attainment in a socioeconomically deprived population: a study linking health, education and social care datasets. PloS One, 7 (2012), pp. 1-8, 10.1371/journal.pone.0043977
- https://europa.eu/eurobarometer/surveys/detail/2660
- https://data.europa.eu/data/datasets/s2660_97_2_sp524_eng?locale=en
- https://www.surveymonkey.com/r/airpollutionperceptionsurvey
- https://apps.who.int/iris/rest/bitstreams/1350812/retrieve
- https://www.ab.gov.tr/files/ardb/evt/Attitudes_of_Europeans_towards_air_quality_2013.pdf