Învățare automată: de la reprezentarea datelor la modele

ZALÁN-PÉTER BODÓ

Teză de abilitare

REZUMAT

Facultatea de Matematică și Informatică
Universitatea Babeş–Bolyai

2023

1. Introducere

Inteligența artificială (IA) este, prin definiție, imitarea, reproducerea, etc., a inteligenței naturale, fără a însemna neaparat inteligență umană – cu toate acestea, credem că oamenii sunt cea mai inteligentă formă de viață pe care o cunoaștem până acum, prin urmare, IA este adesea considerată echivalentă cu inteligența umană. Nu este ușor de abordat în mod general (IA generală sau puternică, *AGI*), prin urmare, este mult mai obișnuit să ne concentrăm asupra unor probleme specifice (IA restrânsă sau slabă), cum ar fi recunoașterea imaginilor, sisteme de întrebare–răspuns, detectarea automată a înșelăciunii, etc. În unele dintre aceste domenii restrânse, mașinile depășesc clar oamenii: de exemplu, predicțiile algoritmilor de învățare automată sunt mai precise comparativ cu cele ale patologilor umani in diagnosticul cancerului [Dabeer et al., 2019]. Dar, în general, putem spune cu certitudine că IA nu există încă. Indiferent dacă va exista vreodată o inteligentă artificială de nivel uman, deja beneficiem de realizările tehnologice ale ultimilor decenii. Suntem deja ciborgi[1], telefoanele inteligente – pe care petrecem mai mult de 3 ore pe zi[2] – reprezintă extensii digitale ale noastre. Este suficient să menționam doar câteva dintre aceste realizări pentru a vedea progresul semnificativ: IBM Watson[3], rezultatele revoluționare ale Alibaba și Microsoft în probleme de întrebare–răspuns care depașesc performanțele umane[4], sau ChatGPT[5], care a dat peste cap lumea de la publicarea sa din noiembrie 2022.

Teza rezumă realizările autorului în domeniul invățării automate, unul dintre cele mai importante subdomenii ale IA. Conform lui [Jäkel et al., 2007]: "machine learning is now an independent and mature field that has moved beyond psychologically or neurally inspired algorithms towards providing foundations for a theory of learning that is rooted in statistics and functional analysis". Unii cercetători critică divergența metodelor de invățare automată față de procesele biologice, cu toate acestea, renunțarea la replicarea exactă a proceselor biologice nu nuapărat reprezintă o problemă, deoarece ar putea exista mai mult de o singură modalitate de a atinge același obiectiv.

Reprezentarea datelor este o parte crucială a învățării automate – nu poate exista învațare fără reprezentare. Deși la început poate părea dificil sau chiar imposibil, totul trebuie să fie atribuit o reprezentare numerică pentru a putea aplica metodele de învățare automată asupra lor – imagini, texte scrise în limbaje naturale, muzică, etc., trebuie transformate în numere. Dacă să folosim sau nu inginerie de caracteristici sofisticată – implicând în mod obișnuit experți în domeniu – pentru a înțelege cele mai importante aspecte ale datelor, sau să folosim abordări de tip end-to-end constituie o dezbatere continuă între cercetători [Glasmachers, 2017]. Conform teoremei *no free lunch* [Wolpert, 1996] precum și a dovezilor empirice ale cercetării în domeniul invățării automate din ultimele decenii, este puțin probabil să existe un *cuțit elvețian* al modelelor de învățare automate – cu toate acestea, învațarea umană indică într-un fel opusul acestui lucru. Presupunerile constituie, de asemenea, o parte importantă a învățării automate, de la reprezentarea datelor la algoritmii de învățare (de exemplu, presupunerea i.i.d., naivitatea in Bayes naiv, presupunerea de netezire în învatarea semi-supervizată, pentru a numi doar câteva). Un model de învățare automată este mai mult decât doar un algoritm (vezi Figura 1): reprezentarea datelor devine o parte integrantă a modelului, deoarece diferite reprezentări pot necesita algoritmi diferiți pentru ca modelul să funcționeze sau să ofere cea mai bună performanță.

2. Conținutul tezei

Prima parte a tezei de abilitare, care prezintă contribuțiile științifice, este împărțită în patru capitole, care corespund celor patru teme principale în care se încadrează lucrările autorului de față, în timp ce secțiunile fiecărui capitol corespund unor lucrări separate publicate după susținerea tezei de doctorat (adică publicate

[1] Thomas Ricker. Elon Musk: We're already cyborgs. The Verge, June 2, 2016. https://www.theverge.com/2016/6/2/11837854/neural-lace-cyborgs-elon-musk

[2] Shelagh Dolan. How mobile users spend their time on their smartphones in 2022, Insider Intelligence, April 14, 2022. https://www.insiderintelligence.com/insights/mobile-users-smartphone-usage

[3] https://www.ibm.com/watson

[4] Rangan Majumder. Microsoft's AI now as good as humans on SQuAD Reading Test, January 16, 2018. https://www.linkedin.com/pulse/microsofts-ai-now-good-humans-squad-reading-test-rangan-majumder/

[5] https://openai.com/blog/chatgpt

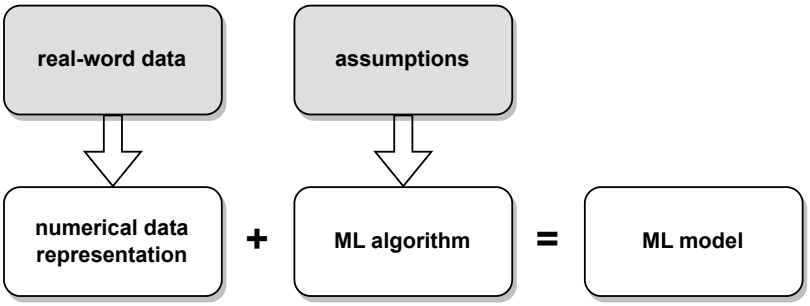

Figura 1: Schema de obținere a modelelor de învățare automată.

după 2009) – în ordine cronologică. Lucrările sunt scurtate la aproximativ 5 pagini și nu includ cursul universitar [Bodó, 2014a], lucrarea educațională [Bodó, 2014b] și nota de curs [Bodó and Csató, 2023].

În timp ce învățarea eficientă de date (*data-efficient learning*) este un termen generic pentru toate metodele care vizează învățarea cu puține date (învățare semi-supervizată, învățare activă, învățare prin transfer, augmentarea datelor etc.) [Adadi, 2021], acest capitol prezintă lucrările legate de învățarea semi-supervizată și învățarea activă și se bazează pe trei lucrări. În [Bodó et al., 2011] prezentăm algoritmul nostru – premiat – combinat, bazat pe învățare activă prin grupare, utilizat în cadrul Active Learning Challenge și prezentat la workshopul Active Learning and Experimental Design din cadrul AISTATS 2010. Lucrarea [Bodó and Csató, 2014b] prezintă o metodă de hashing care poate utiliza, de asemenea, etichete, dacă sunt disponibile și, astfel, poate fi aplicată în scenarii semi-supravegheate. Ultima lucrare inclusă în acest capitol, [Bodó and Csató, 2015], studiază două versiuni ale algoritmului de propagare a etichetelor – o celebră metodă de clasificare semi-supervizată transductivă, bazată pe grafuri [Zhu and Ghahramani, 2002]; diferența dintre cele două variante constă în construcția matricei de tranziție.

Metodele de hashing pot fi utilizate pentru a face eficientă căutarea celor mai apropiaţi k vecini, dar pot fi considerate și tehnici de reducere a dimensionalității [Charikar, 2002]. În lucrările [Bodó and Csató, 2012] și [Bodó and Csató, 2014a] propunem câteva modificări ale algoritmului LSH cu kernel din [Kulis and Grauman, 2009] și, respectiv, ale hashing-ului spectral din [Weiss et al., 2008]. Aceste modificări includ utilizarea de preimagini ale vectorilor de caracteristici gaussiene aleatorii și o liniarizare eficientă a algoritmului de hashing spectral pentru a reduce complexitatea. Lucrările [Mester and Bodó, 2021] și [Mester and Bodó, 2022] aplică hashing-ul sensibil la localitate (*locality-sensitive hashing*) în analiza malware-ului, permițând ca secvențe de instrucțiuni ușor diferite să fie considerate ca fiind aceleași obiecte sau exemple foarte asemănătoare. Versatilitatea acestor coduri este validată prin intermediul mai multor experimente folosind seturi de date din lumea reală.

Fiind cel mai lung capitol al acestei teze, capitolul 4 include lucrări legate de *text mining*: o abordare hibridă – aici însemnând utilizarea atât a algoritmilor nesupravegheați, cât și a celor supravegheați – pentru extragerea metadatelor din articolele academice [Bodó and Csató, 2017], experimente de clasificare a software-ului folosind caracteristici distribuționale bazate pe codul programului [Bodó and Indurkhya, 2017], detectarea genului muzical pe baza versurilor cântecelor folosind diferite caracteristici (de exemplu n-grame și caracteristici de rimă) [Bodó and Szilágyi, 2018], două lucrări care prezintă procesele de compilare a două seturi de date pentru corelarea înregistrărilor [Bodó, 2019] și, respectiv, detectarea știrilor false [Gencsi et al., 2021], experimente de clasificare a știrilor false fără a utiliza cunoștințe externe, cum ar fi modele lingvistice preinstruite [Bodó, 2021], și, nu în ultimul rând, un sistem capabil să genereze scurte secvențe de emoji care descriu ploturile filmelor [Bajcsi et al., 2022]. Metodele propuse sunt validate prin intermediul unor experimente amănunțite care implică seturi de date din lumea reală, în timp ce seturile de date compilate sunt *legitimate* prin compararea rezultatelor experimentale cu rezultatele cunoscute din literatura de specialitate.

Capitolul 5 conține lucrări privind prelucrarea semnalelor și viziunea computerizată. Acesta este al doilea capitol ca lungime al tezei și este, de asemenea, o reflectare fidelă a domeniului de interes recent al autorului tezei de abilitare. Acesta include lucrările care prezintă experimente privind clasificarea regiunilor de muzică populară cu ajutorul rețelelor neuronale convoluționale [Kiss et al., 2019], abordări evolutive de compoziție muzicală folosind cunoștințe specifice domeniului, adică un limbaj specific domeniului (DSL) în

acest caz [Sulyok et al., 2019], un sistem ortogonalizat pentru editarea semantică a fețelor [Antal and Bodó, 2021] și două lucrări legate de modele de învățare automată interpretabile: aplicarea modelului BagNet [Brendel and Bethge, 2019] pentru detectarea tipului de țesut histopatologic [Galiger and Bodó, 2023], și un studiu comparativ al unor modele de clasificare a imaginilor autointerpretabile de succes [Bajcsi et al., 2023].

Ultimul capitol al tezei constituie planul de dezvoltare a carierei, care prezintă obiectivele didactice ale candidatului, obiectivele științifice și academice, în care colaborarea cu studenții de la masterat și doctorat joacă un rol important și, nu în ultimul rând, este prezentată o analiză de risc cu privire la întregul plan.

Bibliografie selectată

A. Adadi. A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8(1):24, 2021.

L. Antal and Z. Bodó. Feature axes orthogonalization in semantic face editing. In *Proceedings of ICCP 2021 (IEEE 17th International Conference on Intelligent Computer Communication and Processing)*, Cluj-Napoca, Romania (online event), 2021.

A. Bajcsi, B. Botos, P. Bajkó, and Z. Bodó. Can you guess the title? Generating emoji sequences for movies. *Studia Universitatis Babeş–Bolyai Informatica*, LXVII(1):5–20, 2022.

A. Bajcsi, A. Bajcsi, S. Pável, A. Portik, C. Sándor, A. Szenkovits, O. Vas, Z. Bodó, and L. Csató. Comparative study of interpretable image classification models. *Infocommunications Journal*, pages 20–26, 2023. Special Issue on Applied Informatics.

Z. Bodó. *Fordítóprogramok szerkesztése Flex és Bison segítségével (Compiler construction using Flex and Bison – in Hungarian)*. Societatea Muzeului Ardelean, Kolozsvár, 2014a.

Z. Bodó. Gépi tanulás gráfokkal (machine learning with graphs – in hungarian). In *Tíz éves az ELTE Eötvös József Collegium Informatikai Műhelye*, pages 61–78. Eötvös József Collegium, 2014b.

Z. Bodó. A CiteSeerX-based dataset for record linkage and metadata extraction. In *Proceedings of the 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 230–236, Timişoara, Romania, 2019. IEEE.

Z. Bodó. Fake news detection without external knowledge. In *Proceedings of MDIS 2020*, volume 1341 of *Communications in Computer and Information Science (CCIS)*, pages 202–221. Springer International Publishing, 2021.

Z. Bodó and L. Csató. Improving kernel locality-sensitive hashing using pre-images and bounds. In *Proceedings of IJCNN*, pages 2710–2717, 2012.

Z. Bodó and L. Csató. Linear spectral hashing. *Neurocomputing*, 141:117–123, 2014a.

Z. Bodó and L. Csató. Augmented hashing for semi-supervised scenarios. In *Proceedings of the 22th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 53–58, 2014b.

Z. Bodó and L. Csató. A note on label propagation for semi-supervised learning. *Acta Universitatis Sapientiae*, 7(1): 18–30, 2015.

Z. Bodó and L. Csató. A hybrid approach for scholarly information extraction. *Studia Universitatis Babeş–Bolyai Informatica*, 62(2):5–16, 2017.

Z. Bodó and B. Indurkhya. Software categorization using low-level distributional features. In *New Trends in Intelligent Software Methodologies, Tools and Techniques. (Proceedings of the 16th International Conference on Intelligent Software Methodologies, Tools, and Techniques*, volume 297 of *Frontiers in Artificial Intelligence and Applications*, pages 88–98, Kitakyushu, Japan, 2017. IOS Press.

Z. Bodó and E. Szilágyi. Connecting the Last.fm dataset to LyricWikia and MusicBrainz. Lyrics-based experiments in genre classification. *Acta Universitatis Sapientiae, Informatica*, 10(2):158–182, 2018.

Z. Bodó, Z. Minier, and L. Csató. Active learning with clustering. In *Active Learning and Experimental Design Workshop 2010*, volume 16 of *JMLR Workshop and Conference Proceedings*, pages 127–139, Sardinia, Italy, 2011.

Z. Bodó and L. Csató. Code optimization with vectorization in data mining and machine learning. In *SusTrainable: Promoting Sustainability as a Fundamental Driver in Software Development Training and Education, 2nd Teacher Training*, pages 2–6, Juraj Dobrila University of Pula, Croatia, January 2023.

W. Brendel and M. Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet, 2019.

M. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, 2002.

S. Dabeer, M. M. Khan, and S. Islam. Cancer diagnosis in histopathological image: CNN based approach. *Informatics in Medicine Unlocked*, 16:100231, 2019.

G. Galiger and Z. Bodó. Explainable patch-level histopathology tissue type detection with bag-of-local-features models and data augmentation. *Acta Universitatis Sapientiae, Informatica*, 15(1):60–80, 2023.

M. Gencsi, Z. Bodó, and A. Szenkovits. Compilation and validation of a large fake news dataset in Hungarian. In *Proceedings of SISY 2021 (IEEE 19th International Symposium on Intelligent Systems and Informatics)*, pages 125–130, 2021.

T. Glasmachers. Limits of End-to-End Learning. In *Proceedings of the 9th Asian Conference on Machine Learning*, pages 17–32, Seol, Korea, 2017. PMLR.

F. Jäkel, B. Schölkopf, and F. A. Wichmann. A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, 51(6):343–358, 2007.

A. Kiss, C. Sulyok, and Z. Bodó. Region prediction from Hungarian folk music using convolutional neural networks. In *Artificial Neural Networks and Machine Learning – ICANN 2019*, volume 11730 of *Text and Time*, pages 581–594. Springer, 2019.

B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*, pages 2130–2137. IEEE, 2009.

A. Mester and Z. Bodó. Validating static call graph-based malware signatures using community detection methods. In *Proceedings of ESANN 2021 (29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning)*, pages 429–434, Bruges, Belgium (online event), 2021.

A. Mester and Z. Bodó. Malware classification based on graph convolutional neural networks and static call graph features. In *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence (IEA/AIE 2022)*, pages 528–539, Cham, 2022. Springer International Publishing.

C. Sulyok, C. Harte, and Z. Bodó. On the impact of domain-specific knowledge in evolutionary music composition. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 188–197, 2019.

Y. Weiss, A. B. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, pages 1753–1760. MIT Press, 2008.

D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.