

Białystok, 5.07.2022 r.

prof. dr hab. inż. Marek Krętowski
Wydział Informatyki
Politechnika Białostocka

Recenzja
w postępowaniu w sprawie nadania stopnia doktora habilitowanego
dr Krzysztofowi Gajowniczкови
w dziedzinie nauk inżynieryjno-technicznych
w dyscyplinie informatyka techniczna i telekomunikacja

Recenzję przygotowałem na podstawie informacji przekazanej 5.05.2022 r. przez prof. Macieja Romaniuka, który działając w imieniu Instytutu Badań Systemowych PAN, poinformował mnie o powołaniu w skład Komisji Habilitacyjnej dra Krzysztofa Gajowniczka i powierzeniu mi obowiązków recenzenta.

I. Sylwetka Habilitanta

Dr Krzysztof Gajowniczek jest związany ze Szkołą Główną Gospodarstwa Wiejskiego w Warszawie od początku swojej pracy zawodowej. Wcześniej był też studentem SGGW, gdzie w 2012 roku obronił pracę magisterską na kierunku informatyka i ekonometria. W 2016 roku w Instytucie Badań Systemowych PAN uzyskał stopień doktora nauk technicznych w dyscyplinie informatyka na podstawie rozprawy „Prognozowanie zapotrzebowania na energię elektryczną z wykorzystaniem inteligentnych systemów pomiarowych”. Krótco po tym awansował na stanowisko adiunkta i do dnia dzisiejszego pracuje w Katedrze Sztucznej Inteligencji Instytutu Informatyki Technicznej. Ponadto współpracuje jako „senior data scientist” z Departamentem Optymalizacji Procesów Monitorowania Przeciwdziałania Praniu Pieniędzy w Citibank Europe.

II. Ocena wskazanego przez Habilitanta osiągnięcia naukowego

Jako osiągnięcie naukowe Habilitant wskazał cykl powiązanych tematycznie 8 artykułów naukowych pod wspólnym tytułem: „Opracowanie niestandardowych funkcji błędu, niejednorodności i regularyzacji dla algorytmów uczenia maszynowego”.

Omówienie struktury cyklu

Przedstawiony cykl składa się z 8 artykułów, z których 7 opublikowanych zostało w czasopiśmie a jeden jest publikacją konferencyjną. Wszystkie artykuły są współautorskie, w pięciu jedynym współautorem jest dr Tomasz Ząbkowski z tej samej co habilitant katedry w

SGGW, w dwu kolejnych oprócz dr Ząbkowskiego występuje jeszcze jeden współautor, w ostatniej pracy jest 4 współautorów łącznie z dr Ząbkowskim. Analizując włączone do dokumentacji oświadczenia współautorów, wydaje się, że generalnie deklarowane i potwierdzone w oświadczeniach udziały procentowe wydają się niezbyt realistyczne i wręcz niepotrzebnie sztucznie korzystne dla habilitanta. Tym niemniej nie mam żadnych wątpliwości, że odegrał on wiodącą rolę w realizacji przedstawionego do oceny osiągnięcia.

Zgodnie z deklaracją habilitanta wszystkie czasopisma mają Impact Factor w granicach od prawie 2 do ponad 3, natomiast jeśli spojrzymy na czasopisma w cyklu przez pryzmat listy ministerialnej, to z pewnym zaskoczeniem odkryjemy, że 3 czasopisma (to jest Processes - 70 pkt., Energies - 140 pkt. oraz Electronics 20 pkt.) nie są przypisane do dyscypliny informatyka techniczna i telekomunikacja. Można wytłumaczyć np. publikację w Energies tematyką pracy, tym niemniej niezwrócenie uwagi na ten element i włączenie do cyku aż 3 takich prac wydaje mi się nieprzemyślane i w moim odczuciu jednak obniża rangę cyklu. Najwyżej punktowane są dwa artykuły w czasopiśmie SoftwareX, po 200 pkt., dwa artykuły ukazały się w Entropy, któremu aktualnie przypisane jest 100 pkt. Konferencja EANN ma tylko 20 pkt. co nie świadczy o rozpoznawalności konferencji.

Omówienie prac wchodzących do cyklu

Najpierw scharakteryzowane zostaną 4 prace dotyczące indukcji drzew decyzyjnych, a następnie 4 prace dotyczące uczenia sieci neuropodobnych, przy czym w ramach grup prace omawiane są w kolejności ukazywania się.

[c5] Gajowniczek K., Ząbkowski T. (2021) lmbTreeEntropy and lmbTreeAUC: Novel R Packages for Decision Tree Learning on the Imbalanced Datasets, Electronics, 10(6), art. 657.

Jest to pierwszy z 4 artykułów przedstawiających opracowane pakiety umożliwiające indukcję drzew decyzyjnych na podstawie danych niezbalansowanych. W artykule przypomniano definicje kilku uogólnionych miar entropii oraz obszaru pod krzywą ROC. Ponadto omówiono również klasyfikację czułą na koszty w kontekście różnych kosztów błędów oraz zagadnienia związane z poszukiwaniem testów na atrybutach nominalnych. Korzystając z wprowadzonych elementów opisano zaproponowane algorytmy budowy drzew, które poruszają się w ramach typowej zachłannej metody zstępującej. W wersji AUC podjęto próbę bardziej globalnego oceniania podziałów i generowanych drzew. W części eksperymentalnej przeanalizowano 10 niewielkich, głównie niezbalansowanych, zbiorów z UCI ML Repository (od 150 do 1484 obserwacji, od 2 do 10 klas). Uzyskane wyniki jakości klasyfikacji i rozmiary drzew zestawiono z rezultatami uzyskanymi przez inne algorytmy indukujące drzewa (C5.0, Ctree oraz Rpart) a także k-NN i sieć neuropodobną (NNET). Eksperymenty nie objęły kosztów, a także nie zaprezentowano analizy statystycznej rezultatów. Autorzy konkludują, że zaproponowane przez nich metody uzyskiwały podobną jakość klasyfikacji jak inne metody, natomiast często wygenerowane drzewa były mniejsze.

Autorzy zdają sobie sprawę, że większość opisanych elementów (w szczególności różne uogólnione miary entropii czy AUC) była już wykorzystywana w indukcji drzew decyzyjnych, tym niemniej siłę swojego rozwiązania widzą w integracji i udostępnieniu w postaci pakietu R.

[c3] Gajowniczek K., Ząbkowski T. (2021) Interactive Decision Tree Learning and Decision Rules Extraction Based on the `lmbTreeEntropy` and `lmbTreeAUC` Packages, *Processes*, 9(7), art. 1107.

W kolejnej pracy autorzy koncentrują się na opracowaniu i zaprezentowaniu w działaniu interaktywnej wersji generowania drzewa decyzyjnego. Zidentyfikowane zostały sytuacje, w których potencjalny użytkownik (ekspert dziedzinowy) mógłby zdecydować o wybraniu podziału z zaproponowanych opcji według własnych preferencji. Ponadto przedstawione zostały możliwości przekształcenia wygenerowanego drzewa w zestaw reguł decyzyjnych z podaniem typowych miar charakteryzujących znaczenie reguł. W części eksperymentalnej opisany został przykład interaktywnego tworzenia drzewa na bazie powszechnie znanego zbioru Iris (150 obserwacji, 4 cechy, 3 klasy po 50). Jakość klasyfikacji utworzonych w ten sposób drzew okazała się o kilka punktów procentowych niższa niż jakość najlepszych drzew generowanych automatycznie.

[c1] Gajowniczek K., Ząbkowski T. (2021) `lmbTreeEntropy`: An R Package for Building Entropy-based Classification Trees on Imbalanced Datasets, *SoftwareX*, 16, art. 100841.

[c2] Gajowniczek K., Ząbkowski T. (2021) `lmbTreeAUC`: An R Package for Building Classification Trees Using the Area Under the ROC Curve (AUC) on Imbalanced Datasets, *SoftwareX*, 15, art. 100755.

Dwa kolejne, najpóźniej opublikowane, artykuły są dość specyficznymi pracami (oznaczenie „original software publication” w C1), gdyż w zasadzie tylko prezentują stworzone i udostępnione w publicznym repozytorium (Github) oprogramowanie, w postaci dwu pakietów do R. Oprócz zwięzłego przedstawienia możliwości oprogramowania, zawierają po kilka przykładów ilustrujących zastosowanie pakietów. Wyniki klasyfikacji oraz rozmiary drzew odpowiednio na 7 i 4 niewielkich zbiorach z UCI ML Repository (od 106 do 1728 obserwacji, od 2 do 8 klas) zestawiono z rezultatami uzyskanymi przez C5.0, Ctree i Rpart. Nie przeprowadzano analizy statystycznej wyników.

Generalnie jednak samą ideę udostępniania oprogramowania wytworzonego na potrzeby badań uważam za bardzo cenną i nawet jeśli same artykuły mają bardziej charakter popularyzatorski niż stricte badawczy, to oceniam je jak najbardziej pozytywnie. Na razie (połowa roku 2022) okres udostępnienia pakietów na Github-ie jest jeszcze krótki i zainteresowanie jest obiektywnie niewielkie, ale mimo wszystko pojawiły się już pierwsze cytowania.

[c8] Gajowniczek K., Orłowski A., Ząbkowski T. (2018) Simulation Study on the Application of the Generalized Entropy Concept in Artificial Neural Networks, *Entropy*, 20(4), art. 249.

Jest to pierwsza praca w drugiej części cyklu dotycząca wykorzystania niestandardowych funkcji błędu w sieciach neuropodobnych. Autorzy postanowili zbadać wpływ wykorzystania uogólnionej entropii Tsallisa na różne aspekty uczenia sieci. Praca ma charakter jedynie studiów symulacyjnych, opiera się na specjalnie przygotowanych sztucznie wygenerowanych zbiorach danych, które zostały zaprojektowane analitycznie i charakteryzują się oczekiwanymi własnościami (zarówno w odniesieniu do poszczególnych cech, jak również zależności między nimi czy stopnia zbalansowania klas). Badania przeprowadzono na prostej strukturze sieci (do kilkunastu neuronów ukrytych), zaś zbiory uczące miały 1000 obiektów, 15 cech i dwie klasy. Wyniki obrazujące m.in. wpływ parametru q na jakość klasyfikacji, potrzebną liczbę iteracji czy

liczbę neuronów ukrytych przy różnych poziomach zbalansowania zbiorów przedstawiono w postaci kolorowych map.

[c6] Gajowniczek K., Ząbkowski T. (2020) Generalized Entropy Loss Function in Neural Network: Variable's Importance and Sensitivity Analysis, L. Iliadis et al. (Eds.): EANN 2020, INNS 2, str. 535–545.

Jest to krótka praca konferencyjna będąca kontynuacją badań symulacyjnych prowadzonych w [c8], przy czym tym razem autorzy postanowili skupić się na interpretacji uzyskanych sieci. Przy wykorzystaniu algorytmów Garsona i Oldena oraz profili Leka badali m.in. wpływ wartości q na znaczenie cech. Tym razem wykorzystano tylko jeden zbiór uczący (10000 obiektów, 5 cech, 2 równoliczne klasy) a analizowana sieć miała 5 neuronów wejściowych, 4 ukryte i jeden na wyjściu.

Przedstawione w tych dwóch pracach badania symulacyjne dotyczące wykorzystania uogólnionej entropii Tsallisa są niewątpliwie interesujące i pożyteczne, z drugiej jednak strony stwierdzenie w autoreferacie (załącznik 3), że przedstawione wyniki „dostarczają dowodu, że entropia Tsallisa może z dużym powodzeniem zastępować inne standardowe funkcje błędu ...” wydaje się niefortunne.

[c7] Gajowniczek K., Liang Y., Friedman T., Ząbkowski T., Van den Broeck G. (2020) Semantic and Generalized Entropy Loss Functions for Semi-Supervised Deep Learning, Entropy, 22(3), art. 334.

W pracy skoncentrowano się na coraz popularniejszym w ostatnim czasie częściowo-nadzorowanym uczeniu głębokich sieci neuronowych. Idea polega na wykorzystaniu w procesie uczenia również obserwacji, których przynależność do klas nie jest znana. Autorzy zdecydowali się zbadać wpływ dodania do głównej funkcji straty błędu semantycznego lub uogólnionej entropii Renyi. Tym razem przeprowadzono badania zarówno w oparciu o zbiór wygenerowany sztucznie jak i dwa znane zbiory obrazów z rodziny MNIST (po 60000 obrazów 28x28 pikseli, 10 klas). Analizowano rozbudowaną wielowarstwową sieć (perceptron) zawierającą tysiące neuronów w poszczególnych warstwach i usprawnioną typowymi rozwiązaniami uczenia głębokiego. Przeprowadzono eksperymentalne badania m. in. zależności jakości klasyfikacji od kombinacji kilku wybranych parametrów. Na podstawie przeprowadzonych badań autorzy stwierdzają, że „regularyzacja oparta o entropię Renyi dostarcza lepszych wyników niż regularyzacja z funkcją semantyczną”. Ponadto potwierdzono, że wykorzystanie częściowo-nadzorowanego uczenia poprawia jakość klasyfikacji.

[c4] Nafkha R., Ząbkowski T., Gajowniczek K. (2021) Deep Learning-Based Approaches to Optimize the Electricity Contract Capacity Problem for Commercial Customers, Energies, 14(8), art. 2181.

Ostatnia z omawianych prac ma odmienny charakter od pozostałych, gdyż jest to interesujący przykład wykorzystania rozwijanych przez autorów metod prognostycznych w rzeczywistym problemie z branży elektroenergetycznej. Opracowana została dwuetapowa metoda wykorzystująca sieć neuronową typu LSTM oraz algorytm genetyczny, a także zaproponowana nowa architektura głębokiej sieci neuronowej, która pozwala na wprowadzenie modelu hybrydowego łączącego dwa etapy. Obie opracowane metody zostały sprawdzone na rzeczywistych zbiorach danych odnośnie prognozy wolumenu mocy dwu komercyjnych taryf opłat za energię i w obu przypadkach potencjalna korzyść dla klientów byłaby zauważalna.

Podsumowanie oceny cyklu

Przedstawione jako osiągnięcie naukowe badania w zakresie uczenia (głębokich) sieci neuronowych oraz indukcji drzew decyzyjnych są niewątpliwie interesujące i aktualne. W swoich pracach Habilitant dotyka wielu istotnych problemów uczenia maszynowego i eksploracji danych takich jak analiza zbiorów niebalansowanych / wieloklasowych, klasyfikacja czuła na koszty, skuteczność strategii zachłannych w indukcji drzew czy uczenie częściowo-nadzorowane. Na szczególne docenienie zasługuje przygotowanie i udostępnienie zainteresowanym w wolnym dostępie dwu bibliotek umożliwiających indukcję drzew decyzyjnych. Otwiera to możliwości zarówno korzystania z zaproponowanych rozwiązań jak i kontynuowania badań rozpoczętych przez Habilitanta.

Można jedynie żałować, że Habilitant nie zdecydował się na przygotowanie monografii, która z pewnością mogłaby być bardziej uporządkowana i jednorodna niż cykl artykułów, a przez to mogłaby pełnić też rolę popularyzującą osiągnięcia Habilitanta oraz byłaby bardziej dostępna i łatwiejsza w odbiorze. Pozwoliłoby to też zapewne zaprezentować bardziej zaawansowany warsztat naukowy w postaci analizy statystycznej podczas porównywania działania klasyfikatorów¹. Niezależnie od powyższych uwag, należy jednoznacznie stwierdzić, że przedstawiony przez Habilitanta cykl publikacji wnosi znaczący wkład do uczenia maszynowego, a tym samym do informatyki technicznej i telekomunikacji.

III. Ocena istotnej aktywności naukowej – pozostałych osiągnięć naukowo-badawczych

Poza publikacjami wchodzącymi do cyklu, kandydat opublikował po doktoracie co najmniej kilkanaście prac badawczych w czasopismach posiadających Impact Factor i znajdujących się na listach ministerialnych (Entropy, Energies, Electronics, PLoS ONE, Complexity, Sensors, ...). W większości są to szybko procedujące czasopisma otwarte. Można trochę żałować, że nie próbuje przedstawiać swoich osiągnięć naukowych w bardziej tradycyjnych czasopismach. Cieszy natomiast to, że w większości prac dr. Gajowniczek jest pierwszym autorem. Dodatkowo był po doktoracie współautorem kilku prac opublikowanych na międzynarodowych konferencjach (np. ICCS, ICANN czy INISTA). W dorobku nie znalazłem ani jednej samodzielnej publikacji, co z jednej strony jest zapewne znakiem czasu, ale z drugiej strony warto byłoby i takie doświadczenie w pracy badawczej kiedyś zdobyć. Generalnie tematyka prac badawczych pokrywa się w dużym zakresie z tematyką cyklu, sporo jest też prac z zakresu zastosowań (modelowania predykcyjnego, analizy szeregów czasowych) czy to w sektorze elektroenergetycznym czy biomedycynie.

Jeśli chodzi o aktywność w zdobywaniu i realizacji grantów, to należy zwłaszcza zauważyć kierowanie grantem NCN (w panelu ST8) oraz aktualnie realizowany grant NCBiR (jako główny wykonawca). Tematyka obu jest związana z zastosowaniami informatyki w energetyce.

¹ Przykładowo, można wykorzystać cytowaną ponad 11000 razy pracę Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7, 1-30.

Ponadto m.in. kilkakrotnie kierował zadaniami w pracach (statutowych) finansowanych przez SGGW.

Wskaźniki bibliometryczne habilitanta są wysokie, we wszystkich bazach indeks Hirscha wynosi powyżej 12, liczba cytowań (bez autocytowań) w najbardziej konserwatywnej bazie Web of Science jest powyżej 200.

Pan dr Gajowniczek jest aktywnym recenzentem, zgodnie z informacjami na portalu Publons zrecenzował ponad 100 artykułów w czasopismach posiadających Impact Factor. Co ważne przygotowuje recenzje nie tylko dla czasopism, w których sam publikuje, ale jest dostrzegany i zapraszany przez redaktorów szerokiej gamy czasopism, w tym specjalistycznych informatycznych czasopism wydawanych przez uznanych wydawców (np. IEEE czy Elsevier).

Bardzo pozytywnie oceniam aktywność dr Gajowniczka w zakresie odbytych staży naukowych, zwłaszcza w wiodących zagranicznych instytucjach naukowych. Na podkreślenie zasługują zwłaszcza 3 dłuższe staże w uczelniach plasujących się wysoko w uznanych rankingach dziedzinowych dotyczących informatyki i które to staże zostały udokumentowane współautorskimi publikacjami :

- miesięczny staż w Computer Science Department, University of Texas, Austin, Stany Zjednoczone Ameryki;
- miesięczny staż w Statistical and Relational Artificial Intelligence lab, Computer Science Department, University of California, Los Angeles, Stany Zjednoczone Ameryki;
- 3-miesięczny staż w Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malezja.

Ponadto pan dr Gajowniczek odbył krótsze staże w Niemczech (na Uniwersytecie w Bamberg) oraz Australii (na Swinburne University of Technology). Niewątpliwie można stwierdzić, że dr K. Gajowniczek wykazuje się istotną aktywnością naukową realizowaną w więcej niż jednej uczelni poprzez współpracę z instytucjami zagranicznymi.

IV. Ocena osiągnięć dydaktycznych, organizacyjnych oraz popularyzujących naukę

Aktywność dydaktyczną dr Gajowniczek realizuje na swojej macierzystej uczelni prowadząc zajęcia na kierunku informatyka oraz informatyka i ekonometria. Są to zarówno przedmioty podstawowe (jak np. Wstęp do programowania czy Ekonometria) jak i specjalistyczne (jak np. „Analiza dużych zbiorów danych”, „Techniki eksploracji danych” czy prowadzone w języku angielskim „Processing of massive datasets”). Prowadzi również zajęcia na studiach podyplomowych Systemy informacyjne i analiza danych. W swoim stosunkowo krótkim okresie promotorstwa opiekował się już 24 dyplomantami przygotowującymi prace magisterskie, inżynierskie i licencjackie.

Pan dr Gajowniczek mimo absorbującej pracy analitycznej w sektorze bankowym (Citibank Europe) angażuje się w różnorodną działalność organizacyjną w jednostce uczelni. Jest członkiem kilku komisji wydziałowych (np. komisji ds. dydaktyki dla kierunku informatyka i ekonometria, komisji rekrutacyjnej do Szkoły Doktorskiej). Został wybrany na członka rady

dyscypliny Informatyka Techniczna i Telekomunikacja. Przez 4 lata był pełnomocnikiem Dziekana do spraw ochrony danych osobowych. Udziela się również w organizacji konferencji naukowych, przykładowo od 2013 roku jest członkiem komitetu organizacyjnego konferencji „Systemy Informatyczne w Zarządzaniu”.

Odnosnie popularyzacji nauki, warto zauważyć zaproszone wystąpienie pana dr Gajowniczka na SAS Global Forum 2019 w Dallas dotyczące zastosowań metod sztucznej inteligencji w analizie danych pochodzących z systemów inteligentnego opomiarowania.

Podsumowując, aktywność dr K. Gajowniczka w zakresie działalności dydaktycznej, organizacyjnej oraz popularyzującej naukę oceniam jako wystarczającą.

V. Podsumowanie

Na podstawie oceny przedłożonego osiągnięcia naukowego w postaci cyklu publikacji oraz na podstawie oceny istotnej aktywności naukowej realizowanej w więcej niż jednej jednostce naukowej, a także uwzględniając całokształt działalności, **pozytywnie** opiniuję wniosek dr Krzysztofa Gajowniczka i wnoszę o dopuszczenie go do dalszych etapów postępowania habilitacyjnego.

M. Kozłowski