Politechnika Wrocławska

Katedra Sztucznej Inteligencji

Wrocław, 20 grudnia 2024 r.

Prof. dr hab. inż. Przemysław Kazienko
Katedra Sztucznej Inteligencji
Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska
Email: kazienko@pwr.edu.pl
http://kazienko.eu

RECENZJA

rozprawy doktorskiej mgr Katarzyny Woźnicy
pt. „*Towards Trustworthy Automated Data Science*"

Rozprawę napisano pod kierunkiem dr hab. inż. Przemysława Biecka na Politechnice Warszawskiej i złożono w 2024 r.

Recenzję wykonano na zlecenie Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej.

I. Przedmiot, problematyka i charakter rozprawy

Tematyka rozprawy mieści się w szerokiej dziedzinie danologii (*data science - DS*) i ogólnie automatyzacji jej złożonych procesów (*Automated Data Science - AutoDS*), w tym także w jej najważniejszej składowej– w uczeniu maszynowym. Konkretniej, rozprawa dotyczy: (1) problemów metauczenia, tj. wiedzy, którą można pozyskać i wykorzystać w procesach uczenia i testowania wielu modeli, (2) dołączania i wykorzystania wiedzy dziedzinowej podczas budowania złożonych modeli wnioskujących, (3) wielokryterialnej oceny modeli uczenia maszynowego oraz (4) wyjaśnialności różnych etapów w automatycznych procesach danologii. Wszystko to, bez wątpienia, zawiera się w dyscyplinie informatyka techniczna i telekomunikacja, w której pracę złożono.

Ogólnie, rozprawa ma charakter koncepcyjno-eksperymentalny, tzn. zidentyfikowano w niej czasami nowe problemy lub inne spojrzenia na już istniejące a następnie je zwalidowano eksperymentalnie na wybranych modelach uczenia maszynowego, dla wybranych zbiorów danych, w ustalonych scenariuszach i z wykorzystaniem czasami nowych miar.

II. Hipotezy i cele pracy

Praca zawiera pięć głównych rozdziałów (rozdz. 3-7), dla każdego sformułowano odrębną hipotezę badawczą, czyli łącznie 5 hipotez, zaś rozdziały te są poprzedzone pewnym wprowadzeniem do tematyki automatycznej danologii (rozdz. 2). Same hipotezy mają sens, aczkolwiek można dyskutować o ich dość ogólnym brzmieniu i nieco oczywistym charakterze, np. „*Włączenie ontologii* […]

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

umożliwia wstrzyknięcie informacji specyficznych dla dziedziny"; to jest oczywiste (Hipoteza 5).

Każdy z owych pięciu rozdziałów ma odrębny charakter i jest w zasadzie niezależnym artykułem naukowym, co wynika z tego, że został napisany na podstawie jednej, ale innej publikacji naukowej.

Taki układ pracy powoduje to, że trudno sformułować jeden cel rozprawy, chyba, że będzie on bardzo ogólny, np. taki jak zawarto w streszczeniu „*zwiększenie zaufania do systemów AutoSD*" (automatycznej danologii).

W związku z powyższym, każdy z głównych rozdziałów można rozpatrywać odrębnie, co – w mojej ocenie – nie umniejsza wartości całej pracy i jest jak najbardziej dopuszczalne w rozprawach doktorskich.

Nie będę opisywał hipotez i celów poszczególnych rozdziałów, ale ważniejszy, wynikający z nich wkład Doktorantki do dyscypliny zawarłem w pkt. *IV Oryginalne osiągnięcia*.

III. Zawartość rozprawy

Rozprawa została napisana w języku angielskim. Jego poziom jest bardzo dobry a tekst zrozumiały. Tekst składa się z 8 rozdziałów oraz bibliografii zawierającej 241 pozycji.

IV. Oryginalne osiągnięcia

Praca zawiera kilka wartościowych osiągnięć, w szczególności:

1. Ewaluacja różnych metod wypełniania brakujących danych (rozdz. 3). Dotyczy ona jednego z procesów przygotowania danych dla różnych, dalszych zadań poprzez wnioskowanie o ich wartości na podstawie istniejących, innych danych. Wydaje się, że jest to istotna, ale jednak najmniej nowatorska część całej pracy.
2. Opracowanie narzędzia i jego szerokie eksperymentalne wykorzystanie do oceny znaczenia poszczególnych cech metadanych procesu uczenia, w tym profilu zbioru danych czy hiperparametrów (rozdz. 4). W szczególności ciekawa i wartościowa jest propozycja analizy hierarchicznych zależności pomiędzy cechami (w tym triploty) idąca w kierunku analizy wielokryterialnej a także podejście uogólniające znaczenie cech względem wielu modeli używanych jako czarne skrzynki. W dużym stopniu rozwiązuje to problem tego, że poszczególne metacechy mogą mieć różne znaczenie dla każdego zbioru, jak również dla każdego modelu, co wynika z możliwej dużej złożoności zależności istniejących w danych. Sam problem ma duży potencjał dalszych rozszerzeń, np. wykrywania anomalii czyli zbiorów o skrajnie różnej

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

charakterystyce, rozróżnianie trenowania od pretrenowania, uczenia wielozadaniowego (*multitask learning*), itd.

3. Bardzo ciekawa i innowacyjna koncepcja wykorzystania rankingu ELO używanego m.in. w szachach lub dla reprezentacji piłkarskich FIFA (Rosja jest w nim tuż przed Polską) jako miary *EPP score* (*Elo-based Predictive Power*) dla tworzenia rankingu modeli (rozdz. 5). Sama koncepcja, jej analiza i eksperymentalna weryfikacja jest moim zdaniem najważniejszym dokonaniem Doktorantki. W pracy zaprezentowano i zanalizowano także dziesięć właściwości nowej metody.
4. Zdefiniowanie w pewnym sensie nowego problemu łącznej optymalizacji hiperparametrów dla wielu zbiorów i zadań czyli dla wielu modeli jednocześnie (*consolidated learning*), rozdz. 6. Rozważano modele uczone na zbiorach mających różne zbiory obserwacji – scenariusz S2 oraz – co jest ciekawsze – podział rozszerzony o różne zbiory zmiennych predykcyjnych (scenariusz S3). Jest to – w pewnym sensie – nieco inaczej zdefiniowany problem transferu uczenia – *transfer learning* (zabrakło nieco odniesienia do tej dziedziny), ale nowością jest tutaj meta-uczenie, tj. optymalizacja hiperparametrów dla wielu modelów jednocześnie. Całość przetestowano eksperymentalnie dla jednego zbioru medycznego MIMIC-IV (odpowiednio dzielonego) i jednego modelu (XGBoost). Równocześnie w eksperymentach rozważano różne podproblemy i zależności, np. liczbę zbiorów czy zmiennych dostępnych w metauczeniu.
5. Koncepcja *Semantic Feature Net* (*SeFNet*), rozdz. 7, która przede wszystkim bazuje na zaproponowanej niesymetrycznej mierze DOSS (*Dataset Ontology-based Semantic Similarity measure*). Określa ona bliskość między zbiorami na podstawie relacji pomiędzy ich cechami składowymi, tj. podobieństwem związanych z nimi terminów, z wykorzystaniem zewnętrznych ontologii (tutaj medyczna terminologia *SNOMED Clinical Terms*). Koncepcja została zweryfikowana eksperymentalnie na danych medycznych.

Warto zwrócić uwagę, że wyniki ww. badań były dodatkowo zweryfikowane poprzez procesy recenzji oraz prezentacji z dyskusją dla społeczności międzynarodowej, gdyż treści poszczególnych rozdziałów były opublikowane na warsztatach bardzo dobrych konferencji (*ICML, PKDD*) lub w bardzo dobrych czasopismach (*Nature Machine Intelligence*, *Machine Learning* oraz *Konwledge-based Systems* - artykuł w recenzji).

Łącznie, powyższa lista osiągnięć sprawia, że recenzowana monografia stanowi nowe i ważne osiągnięcie naukowe, w którym zastosowano właściwe, głownie eksperymentalne metody badawcze, jak również odpowiednie formalne definicje, w tym np. definicje miar.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

V. Pytania, uwagi dyskusyjne i problemy

Pytania i problemy

1. Definicje zbiorów (s. 22) dotyczą tylko liczbowych (liczby rzeczywiste) zmiennych opisujących przypadki i jedynie jednowartościowych zmiennych wyjściowych (kategoryczne lub numeryczne). Takie ograniczenie jest oczywiście dopuszczalne, ale nie wiem czy niezbędne. Ogólnie są rozwiązania i modele, które operują na innych typach danych, np. na sekwencjach (serie czasowe, sygnały) czy grafach (*structure output prediction*), na obrazach, czy tekście. Jakie miało by to konsekwencje dla przedstawionych rozwiązań?
2. Nie do końca jest jasny wkład Doktorantki do efektów rozdz. 5 i publikacji w *Nature Machine Intelligence*, zwłaszcza chodzi o porównanie z wkładem pierwszej autorki Alicji Gosiewskiej. W samym artykule mamy informację: „ *A.G. and K.W. […] studied and described theoretical properties of the EPP. […]. All authors participated in the conceptualization and preparation of the paper*". W zasadzie podobne pytanie dotyczy także publikacji związanych z rozdz. 6 i 7. W artykule w czasopiśmie *Machine learning* (rodz. 6), w którym Doktorantka jest pierwszą i korespondencyjną autorką mamy wpisany: „*All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Katarzyna Woźnica, Mateusz Grzyb and Zuzanna Trafas*". Zdanie „*The original draft of the manuscript was written by Katarzyna Woźnica and all authors commented on previous versions of the manuscript.*" nieco wyjaśnia problem.

Uwagi dyskusyjne

3. Wnioskowanie o brakujących wartościach (rozdz. 3). Eksperymenty przeprowadzono na grupie 14 zbiorów danych. W efekcie otrzymano 14 punktów pomiarowych dla każdego z pięciu testowanych algorytmów. Nawet gdyby zbiorów było więcej dalej nic nie możemy powiedzieć o *zjawiskach*. Czy w związku z tym rozważane były rozwiązania manipulacji danymi tak, aby potencjalne zjawiska zbadać? Dla przykładu, efektywność metod może zależeć od pokrycia dziedziny wartości przez znane i nieznane wartości. W najprostszym przypadku dla zmiennej binarnej możemy mieć niezbalansowanie i jego wpływ można badać sztucznie regulując ów poziom niezbalansowania. Brak lub bardzo ograniczone badanie zjawisk jest typowym ograniczeniem większości prac naukowych w dziedzinie SI. Dalej mamy testowane pięć algorytmów, ale nie bardzo wiadomo jaka ich charakterystyka ma tutaj znaczenie, np. na jakie błędy poszczególne z nich są mniej lub bardziej wrażliwe. Poziomem błędów także można sterować chcąc zbadać *zjawisko*. Chyba nie zgodzę się ze stwierdzeniem, że jest to „*the first empirical benchmark of imputation methods*", czyli problemem wyboru najlepszego algorytmu wypełniania brakujących danych (s. 60).

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

Wystarczy wpisać „*"missing values" "imputation method" benchmark*” w Google Scholar.

4. W badaniach eksperymentalnych wykorzystano głównie relatywnie proste i małe modele uczenia maszynowego. Jaki wpływ na uzyskane wyniki ma pojemność / wielkość modelu, np. liczba parametrów sieci neuronowej?
5. Czy i jak można by wykorzystać zaproponowane metody wyjaśniania w trakcie meta uczenia w złożonych procesach pretrenowania modeli podstawowych / fundamentalnych (*foundation models*), np. wielkich modeli językowych?
6. W rozdz. 4 badano różne metacechy danych i procesu uczenia. Na ile ów zbiór badanych cech jest kompletny i jak można by rozszerzyć, np. tworząc coraz bardziej złożone konstrukty statystyczne (zauważmy, że korelacje są już uwzględnione np. w triplotach).
7. Ranking ELO ma duży sens szczególnie dla dynamicznych procesów, tj. zmieniających się w czasie skuteczności porównywanych obiektów (tutaj modeli). Oznacza to, że obejmuje on zmiany, np. to, że dany model był słaby a obecnie się poprawił. Czy w takim razie kolejność poszczególnych zawodów (tj. zbiorów) i potyczek ma dla *EPP score* znaczenie, a jeżeli tak to jakie? Analizy z załącznika dot. stabilności *EPP* nie do końca odpowiadają na to pytanie.
8. W jakich przypadkach uczenie skonsolidowane ma potencjalny sens i może mieć praktyczne zastosowanie?
9. Czym scenariusz S2 (podział obserwacji) w uczeniu skonsolidowanym różni się od 2-foldowej walidacji krzyżowej? Czy chodzi o łączną optymalizację hiperparametrów dla obu foldów a nie każdego niezależnie? Jak duży wpływ na wyniki ma sama metoda podziału?
10. Ponieważ ontologie mają charakter grafów, czy nie lepszym byłoby wykorzystanie miar grafowych, w tym także charakteru relacji pomiędzy terminami do ustalenia podobieństwa terminów (wzór 7.1)? Patrz Słowosieć jako przykład takiego grafu.

Chciałbym zwrócić uwagę na to że powyższe pytania i uwagi dyskusyjne nie są świadectwem niskiej jakości rozprawy, ale inspiracją wynikającą z dojrzałości zaprezentowanych analiz, z których mogą wynikać dalsze badania.

Uwagi formalne i redakcyjne

11. Hipoteza 1 (s. 17) może być nieco kontrowersyjna, gdyż sformułowanie „*can simplify the exploration of potential operations*” jest mało precyzyjne.
12. Brak wyjaśnienia oznaczenia dla *G*; także indeks *train* nie jest spójny z poprzednimi oznaczeniami s. 23 wzór 2.1.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

13. Definicja walidacji krzyżowej może być dyskusyjna, gdyż zwykle przyjmuje się rozłączny podział zbioru, co nie jest zapewnione we wzorze 2.2 (s.23).
14. Ponieważ Algorytm 1 i 2 pochodzą z literatury warto by umieścić odpowiednie odesłanie w samym algorytmie (np. po nazwie) a nie tylko w opisie pracy. W efekcie czytelnik nie jest pewien, czy może Algorytm 2 jest dokonaniem Autorki? Dodatkowo, w zadaniu następnym mamy A-SMFO zaś w nazwie algorytmu SMFO.
15. Niektóre podrozdziały zaczynają się z małej litery, np. 4.7, 4.8, 4.8.3.

VI. Podsumowanie i ocena rozprawy

Podsumowując, należy stwierdzić, że rozprawa jest bardzo ciekawa i dotyczy ważnej, użytecznej i różnorodnej tematyki metauczenia. Bez wątpienia stanowi istotny wkład w dyscyplinę informatyki technicznej i telekomunikacji. Osiągnięcia wymienione powyżej w pkt. IV są nowe i wartościowe. Świadczą one o bardzo dobrze dobranym warsztacie badawczym i sporych umiejętnościach Doktorantki. Treść jest metodologiczne poprawna, na wysokim poziomie i mieści się w aktualnych kierunkach badań na świecie. Tezy i wyniki zostały dodatkowo zweryfikowane za pomocą publikacji na bardzo dobrych konferencjach i w czasopismach naukowych.

Pomimo pewnych wątpliwości dotyczących wkładu Autorki, szeroki zakres rozprawy oraz poziom i dojrzałość badań moim zdaniem same się bronią, bez względu na interpretację owego wkładu.

W związku z powyższym stwierdzam, że opiniowana rozprawa doktorska mgr Katarzyny Woźnicy spełnia wymagania stawiane w obowiązujących przepisach ustawy o stopniu naukowym doktora i wnoszę o dopuszczenie jej Autorki do publicznej obrony.

Jednocześnie biorąc pod uwagę aktualność, nowatorskość oraz wysoki poziom przedstawionych badań proponuję rozważyć wyróżnienie rozprawy.

Kwiecko

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434