Prof. dr hab. inż. Andrzej Polański

Katedra Grafiki, Wizji Komputerowej i Systemów Cyfrowych

Politechniki Śląskiej

Gliwice 7.01.2025

Recenzja rozprawy doktorskiej

Autor: Mgr inż. Katarzyna Woźnica

Tytuł: Towards Trustworthy Automated Data Science

Promotor: Prof. dr hab. inż. Przemysław Biecek

Ogólna charakterystyka rozprawy

Przedłożona do recenzji rozprawa doktorska jest napisana w języku angielskim. Liczy 190 stron tekstu, składa się z ośmiu rozdziałów oraz bibliografii. Praca, zgodnie z tytułem ukierunkowana jest na badaniach związanych z automatycznym wykonywaniem procesów / potoków analiz danych związanych z przetwarzaniem, eksploracją, wnioskowaniem oraz podsumowaniami i wizualizacją danych. Ten obszar badań nazywany jest w pracy skrótowym terminem AutoDS (Automated Data Science). Z ośmiu rozdziałów pracy dwa pierwsze mają charakter wstępu i formalizacji pojęć i metod wykorzystywanych w pracy. Pięć kolejnych rozdziałów zostały podzielone na dwie części; pierwsza z nich (obejmująca trzy rozdziały) dotyczy metodologii oceny jakości różnych kroków w analizie / eksploracji danych, druga (składająca się z dwóch rozdziałów) – metodologii integracji wiedzy dziedzinowej w potoki analiz wchodzących w skład AutoDS. Ostatni, ósmy rozdział stanowi podsumowanie. Bibliografia pracy jest bardzo obszerna, składa się z 241 pozycji.

Tematykę pracy należy uznać za bardzo ważną i interesującą. Obszar badań i wiedzy nazywany angielskim terminem Data Science (DS) jest obecnie jednym z wiodących i najbardziej intensywnie rozwijanych. Obszar DS bardzo ściśle wiąże się z rozwojem uczenia maszynowego i sztucznej inteligencji. Charakteryzuje się interdyscyplinarnością, ma bardzo szerokie zastosowania, zarówno w nauce jak też właściwie w każdym obszarze związanym z zastosowaniami i komercjalizacją wyników badań. Obecnie, umiejętność / zdolność do odniesienia swoich prac lub badań naukowych do wyników z obszaru DS jest właściwie koniecznością. Dodatkowo, tempo rozwoju naukowego w obszarze DS jest na takim

poziomie, że formułowane są bardzo daleko idące prognozy dotyczące perspektyw oddziaływania wyników w obszarze DS na rozwój wielu kierunków prac i badań. Wpływ rozwoju DS na naukę i wiele innych obszarów jest też już oczywiście bardzo wyraźnie widoczny.

Strukturę DS Doktorantka przyjmuje jako składającą się z czterech części składowych, nazywanych angielskimi terminami Data Engineering (przygotowanie, przetwarzanie, odszumianie danych), Data Exploration (eksploracja danych, eksploracja wiedzy dziedzinowej, formułowanie celów analizy, wizualizacje), Model Building (wybór modelu, wybór algorytmu, optymalizacja parametryczna, ocena dopasowania) oraz Exploitation (interpretacje wyników analiz, wizualizacje, predykcje). Taki punkt widzenia Doktorantka przyjmuje za wielokrotnie cytowaną publikacją „Automating Data Science: Prospects and Challenges" opublikowaną w czasopiśmie Communications of the ACM w 2022 roku przez grupę profesorów pracujących w obszarze DS, z Wielkiej Brytanii, Belgii, Holandii, Szwecji i Stanów Zjednoczonych. Osiągnięcia naukowe zawarte w publikacjach, których współautorką jest Doktorantka są omówione w rozdziałach 3, 4, 5 w części pierwszej oraz 6, 7 w części drugiej.

W rozdziale 1 pracy formułuje się pięć tez badawczych. Każda z tych tez jest tezą jednej z publikacji przedstawionych w rozdziałach 3-7, lub też jest ściśle z nią związana. Doktorantka stara się zinterpretować wyniki swoich prac naukowych w taki sposób aby mogły być rozumiane jako części składowe struktury DS pochodzącej z powyżej wspomnianej publikacji.

W rozdziale 2 pracy Doktorantka wprowadza definicje i opisy pojęć, koncepcji związanych z tematyką pracy oraz formalizacje metod implementowanych w swoich publikacjach i w pracy. Rozdział 2 jest odniesiony do dużej liczby prac literaturowych.

Rozdział 3 pracy jest powiązany z publikacją, Woznica, K., & Biecek, P. (2007). *Does Imputation Matter? Benchmark for Predictive Models.",* https://arxiv.org/pdf/2007.02837, Workshop on the Art of Learning with Missing Values, International Conference of Machine Learning. W publikacji tej dokonano eksperymentalnego porównania siedmiu metod imputacji brakujących danych, możliwych do zastosowania w algorytmach klasyfikacji. Wybrano 13 zbiorów danych z publicznie dostępnego repozytorium OpenML100, w których występowały brakujące dane. Procent brakujących wartości był różny dla różnych zbiorów danych, (od 0.6% do 33%). Użyto sześć algorytmów klasyfikacji oraz dwóch indeksów jakości klasyfikacji (indeksu F1 oraz AUC – pola powierzchni pod wykresem krzywej ROC). Wynikiem eksperymentów obliczeniowych były rankingi metod imputacji brakujących danych. Kolejność metod imputacji zmienia się zarówno w zależności od algorytmu klasyfikacji jak też od analizowanego zbioru danych. W artykule (rozdziale 3) stosuje się różne podsumowania (uśredniania) rankingów. Konkluzje pracy nie zawierają jednoznacznych rekomendacji (ocen jakości) dla różnych metod imputacji. Często w rankingach wysokie miejsca zajmują najprostsze metody imputacji, uśrednienie lub wpisanie losowych wartości w brakujące pola.

W rozdziale 4 pracy omawiane są wyniki publikacji, Woźnica, K., & Biecek, P. (2021, September). *„Towards explainable meta-learning"*. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 505-520). Cham: Springer International Publishing. Publikacja ta poświęcona jest opisowi nowej metodologii poprawy efektywności i intepretowalności algorytmów metauczenia. Kluczową oryginalną koncepcją pracy jest połączenie technik wykorzystania modeli zastępczych (surrogate models) optymalizacji / strojenia hiperparametrów algorytmów klasyfikacji / uczenia maszynowego, z metodami wyjaśnialnej sztucznej inteligencji (eXplainable Artificial Intelligence, XAI). Problem efektywnego doboru hiperparametrów w algorytmach uczenia maszynowego jest intesywnie studiowany w literaturze naukowej od co najmniej kilkunastu lat. W powiązaniu z tym problemem definiuje się techniki budowy modeli zastępczych dla szybszego przeszukiwania przestrzeni hiperparamterów. Rozwijane metody pozwalają studiować, proponować rozwiązania postulatu automatyzacji algorytmów uczenia maszynowego. W omawianym artykule Doktorantki i jej promotora proponuje się wzbogacenie istniejących podejść o nowy aspekt analizy, to znaczy o opisy i interpretacje wyników przeszukiwania przestrzeni hiperparamterów z wykorzystaniem XAI. W pracy w pierwszym kroku tworzy się model / system nazwany Meta-OpenML100, który jest modelem zastępczym typu czarnej skrzynki stworzonym na podstawie repozytorium benczmarkowego OpenML100. Techniki tworzenia tego systemu bazują na cytowanej w pracy literaturze. W drugim kroku wykorzystuje się techniki XAI do scharakteryzowania i interpretacji statystycznych oraz graficznych uzyskanego w pierwszym kroku systemu Meta-OpenML100. Metody XAI użyte są do zbudowania rankingów meta cech oraz grup meta cech, do opisu interakcji pomiędzy meta cechami, do charakteryzowania ważności cech w postaci dendrogramów, do interpretacji znaczenia hiperparametrów w postaci wykresów CP (Ceteris Paribus, profili warunkowych wartości oczekiwanych) oraz do odporności algorytmów uczenia na cechy / obserwacje odstające. Jako główny wniosek z przeprowadzonych badań formułuje się samo wskazanie możliwości i użyteczności zastosowania XAI w meta uczeniu.

Rozdział 5 pracy bazuje na publikacji, Gosiewska, A., Woźnica, K., & Biecek, P. (2022, December). *„Interpretable meta-score for model performance"*. In ECMLPKDD Workshop on Meta-Knowledge Transfer (pp. 75-77). PMLR, Nature Machine Intelligence, 4(9):792–800. Publikacja ta proponuje budowę systemu rangowania algorytmów uczenia maszynowego wzorowanego na systemach porównywania / oceniania siły gry szachistów, Elo. W proponowanym, przejętym za literaturą systemie jakość algorytmów klasyfikacji ocenia się wykonując porównania pomiędzy parami algorytmów, dla wielu zbiorów danych. Bazuje się przy tym na modelu logistycznym zależności pomiędzy jakością algorytmu a prawdopodobieństwem uzyskiwania przewagi nad innymi w porównaniach parami. Jakość algorytmu jest opisywana wskaźnikiem / parametrem oznaczanym skrótem EPP, występującym w wykładniku funkcji logistycznej. Wielokrotna walidacja krzyżowa pozwala na symulowanie wygranych i porażek dla porównywanych parami algorytmów. Dla zbudowania rankingu pięciu algorytmów (gbm, kknn, glmnet, RF, ranger) użyto 30 zbiorów

danych z repozytorium OpenML100. Każdy z algorytmów był uruchamiany dla kilkuset różnych ustawień wartości jego hiperparametrów.

W rozdziale 6 pracy omawia się wyniki przestawione w artykule Woźnica, K., Grzyb, M., Trafas, Z., & Biecek, P. (2024). *„Consolidated learning: a domain-specific model-free optimization strategy with validation on metaMIMIC benchmarks"*. Machine Learning, 113(7), 4925-4949. W artykule tym autorzy wprowdzają oryginalną koncepcję uczenia konsolidowanego (consolidated learning) jako nowej techniki meta – uczenia. Zadanie meta uczenia jest tu rozumiane jako problem konstrukcji efektywnych metod/technik doboru/optymalizacji hiperparametrów dla algorytmów klasyfikacji. Efektywność polega na tym, że opracowane techniki powinny pozwalać na szybki dobór dobrych zestawów hiperparamterów dla nowych zadań/danych. Powinny przy tym wykorzystywać doświadczenia/oceny zebrane podczas realizacji wcześniejszych zadań klasyfikacji. Autorzy publikacji, w tym aspekcie, studiują problem doboru hiperparamterów algorytmu XGBoost w zastosowaniu do danych z publicznie dostępnego repozytorium danych medycznych MIMIC-IV (Medical Information Mart for Intensive Care). Proponowana metodologia uczenia konsolidowanego polega na sformułowaniu systemów/schematów podobieństwa pomiędzy zadaniami meta uczenia i meta testowania. Proponuje się kilka takich schematów, a następnie bada się jak ich wykorzystanie (wkomponowanie w procedurę optymalizacji hiperparametrów) wpływa na dynamikę meta uczenia. Jako wynik przeprowadzonych eksperymentów obliczeniowych pokazuje się dużą przewagę proponowanych algorytmów uczenia konsolidowanego szczególnie nad najprostszym algorytmem losowego przeszukiwania przestrzeni hiperparametrów. Wykazuje się także, że (tak jak się to opisuje we wstępie do artykułu) proponowana metodologia uczenia konsolidowanego jest odpowiednim narzędziem dla tworzenia dobrych zestawów hiperparamterów w sytuacji gdy repozytorium danych na których się pracuje ma charakter taki jak MIMIC-IV – są to dane z jednego obszaru, posiadające wspólne/podobne cechy (zmienne objaśniające i wyjściowe). Aby to zobrazować ta sama metodologia jest także zastosowana dla danych z repozytorium OpenML, o bardzo różnorodnym charakterze. Porównanie wyników klasyfikacji pomiędzy zbiorami danych z tych dwóch różnych repozytoriów pozwala autorom wykazać tezę, że podobieństwa pomiędzy zbiorami danych w MIMIC-IV pozwala na efektywniejszy transfer konfiguracji hiperparametrów pomiędzy podobnymi zadaniami klasyfikacji.

Wreszcie, rozdział 7 pracy formułuje wyniki badawcze uzyskane z udziałem Doktorantki w publikacji Woźnica, K., Wilczyński, P., & Biecek, P. Sefnet: „Linking Tabular Datasets with Semantic Feature Nets". SSRN 4811308. W artykule tym przyjmuje się, że wiersze w zbiorach danych analizowanego repozytorium indeksowane są terminami, dla których dane jest drzewo ontologii. Dzięki temu autorom udaje się na bazie odległości pomiędzy terminami w ontologii zdefiniować wskaźnik podobieństwa (DOSS) pomiędzy zbiorami danych. Analizowane jest repozytorium obejmujące kilkanaście (16) zbiorów danych, o charakterze medycznym, dla których odpowiednie jest zastosowanie ontologii SNOMED-CT. Na przykładzie porównania pomiędzy procedurami meta uczenia

skonstruowanego repozytorium danych medycznych oraz meta uczenia repozytorium OpenML. W procedurach tych dokonywano optymalizacji hiperparamterów algorytmu XGBoost. W przypadku wykorzystania wskaźnika podobieństwa DOSS pomiędzy zbiorami danych (dla repozytorium indeksowanego terminami ontologii SNOMED-CT) udawało się uzyskać lepszą dynamikę poprawy wartości wskaźnika ADTM, a także lepsze wartości ustalone tego wskaźnika niż dla repozytorium OpenML, dla którego nie wykorzystywano wskaźnika DOSS.

Ocena rozprawy

Recenzowaną rozprawę oceniam bardzo pozytywnie. Doktorantka wykazuje się szeroką wiedzą dotyczącą współczesnych technik uczenia maszynowego. Publikacje, których wyniki i tezy zawarte są w pracy, powstałe z udziałem Doktorantki, zawierają bardzo wiele oryginalnych i interesujących pomysłów. Są one opisywane oraz testowane na rzeczywistych, szerokich zbiorach danych. Formułuje się także koncepcje ich rozwijania.

Za najciekawsze z oryginalnych osiągnięć naukowych w pracy uważam sformułowanie koncepcji / metody uczenia konsolidowanego (przedstawionej w rozdziale 6) i wykazanie jej przydatności na przykładzie rzeczywistego zbioru danych medycznych. Rozdział 7 może być rozumiany jako rozwinięcie metodologii uczenia konsolidowanego przed dodanie do niego aparatu ontologii terminów dziedzinowych.

Publikacje, na których bazuje praca ukazały się w renomowanych czasopismach naukowych lub / i w materiałach międzynarodowych konferencji naukowych.

Za bardzo cenne uważam podjęcie przez Doktorantkę wysiłku polegającego na dokonaniu (lub bardziej precyzyjnie przyjęciu na podstawie literatury) koncepcji opisu struktury DS oraz umieszczeniu w tej strukturze wyników pięciu oryginalnych publikacji naukowych powstałych ze swoim współudziałem. Taki schemat podkreśla orientację Doktorantki w szerokim obszarze (albo w przynajmniej w dużych częściach) uczenia maszynowego i klasyfikacji danych.

W ramach publikacji przedstawionych w pracy zostało opracowane oprogramowanie dostępne na platformie Github.. Jakość opracowania tej dokumentacji wykazuje doświadczenie programistyczne Doktorantki.

Uwagi i pytania do doktoratu

Do rozdziału 3: Wysokie pozycje najprostszych algorytmów imputacji wydają się sugerować, że uniwersalność metod imputacji danych jest dużym problemem. Czy nie powinno się poświęcać więcej uwagi metodom określania zakresów stosowalności poszczególnych algorytmów imputacji?

Do rozdziału 4: Teza sformułowana w związku z tym rozdziałem deklaruje poprawę zrozumienia wagi różnych meta cech. Czy uzyskanie wiedzy o wagach meta cech daje możliwość poprawy jakości klasyfikacji?

Do rozdziału 5: Czy rankingi algorytmów otrzymywane z zastosowaniem zaproponowanej metody są stabilne? Inaczej mówiąc – czy zależą mocno od kolejności analizowanych zbiorów danych?

Do rozdziału 6: Najprostszy algorytm przeszukiwania siatki wartości hiperparametrów daje bardzo złe wyniki. Czy badano zastosowanie do przeszukiwania przestrzeni hiperparamterów alagorytmu Metropolisa – Hastingsa (MCMC)? Czy algorytm BO (optymalizacji bayesowskiej) da się jakoś porównać z algorytmami MCMC?

Do rozdziału 6 i 7: Czy dałoby się zilustrować skuteczność wykorzystywania zaproponowanej metodologii z użyciem sztucznie wygenerowanych danych?

Konkluzja

Praca stanowi podsumowanie oryginalnych, interdyscyplinarnych badań naukowych, w których formułowano i weryfikowano oryginalne hipotezy badawcze. Dokumentuje wiedzę i kompetencje Doktorantki. Osiągnięcia i oryginalne elementy pracy są na pewno wystarczające do jej ogólnej pozytywnej oceny. Stwierdzam, że rozprawa spełnia warunki stawiane pracom doktorskim i wnioskuję o jej dopuszczenie do publicznej obrony.

A. Rucin