POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI
ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcf@put.poznan.pl, www.put.poznan.pl

Poznań, 4 czerwca 2018 r.

dr hab. inż. Mikołaj Morzy, prof. nadzw.

Mikolaj.Morzy@put.poznan.pl

Recenzja rozprawy doktorskiej dla Instytutu Podstaw Informatyki Polskiej Akademii Nauk

Tytuł rozprawy: **Support Vector Machines for Uplift Modeling**

Autor rozprawy: **Łukasz Zaniewicz**

Jakie zagadnienie naukowe jest rozpatrzone w pracy (teza pracy) i czy zostało ono dostatecznie jasno sformułowane przez autora? Jaki charakter ma praca (teoretyczny, doświadczalny, inny)?

Recenzowana rozprawa dotyczy nowego podejścia do problemu modelowania różnicowego przy wykorzystaniu metodyki SVM. W rozprawie przedstawiono adaptację oryginalnej metodyki SVM do specyfiki modelowania różnicowego, zaprezentowano obszerną analizę matematyczną proponowanego rozwiązania, zaproponowano także modyfikacje algorytmu (L_p-USVM lub regularyzacja Szekely'ego) które zapobiegają wystąpieniu niekorzystnych konfiguracji, takich jak nieskończony margines między hiperpłaszczyznami decyzyjnymi. Praca jest bardzo mocno osadzona w dziedzinie uczenia maszynowego i rozwija obszar modelowania różnicowego. Jest to obszar trudny i wymagający, o dużym znaczeniu praktycznym w wielu różnych dziedzinach nauki i przemysłu. W przeciwieństwie do problemu klasyfikacji (który można umownie potraktować jako problem "płytki" znalezienia związków między zmiennymi niezależnymi i zmienną zależną), modelowanie różnicowe stara się rozwiązać problem "głęboki", wiążący obserwowany wynik (w formie zmiennej celu) z akcją, jakiej poddana była dana instancja ucząca. Poza oczywistymi zastosowaniami w medycynie czy marketingu, efektywne algorytmy modelowania różnicowego mogą znaleźć zastosowanie w handlu internetowym, systemach rekomendacyjnych, czy ogólnych systemach sterowania. Praca ma charakter oryginalny a zaprezentowane w niej badania cechują się bardzo wysokim poziomem merytorycznym i dojrzałością naukową, a także stanowią ważny cel, spełniający wymagania zwyczajowo stawiane rozprawom doktorskim.

Rozprawa jest napisana w języku angielskim i liczy, wraz z bibliografią, 88 stron. W przedmowie zawarto krótkie streszczenie w języku polskim. Na rozprawę składa się siedem rozdziałów, bibliografia liczy 50 pozycji, w tym trzy prace których współautorem jest Doktorant. Rozdział pierwszy wprowadza czytelnika w problem modelowania różnicowego, przedstawia motywacje do podjęcia badań w tym obszarze, a także prezentuje aktualny stan badań i oryginalny wkład recenzowanej rozprawy w domenie modelowania różnicowego. W rozdziale 2 Autor zamieszcza prezentację algorytmu wektorów

ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcf@put.poznan.pl, www.put.poznan.pl

wspierających oraz przedstawia matematyczne podstawy algorytmu uczenia. Tu także wprowadzona zostaje notacja matematyczna, która w sposób bardzo spójny jest wykorzystywana w całej rozprawie. Główne oryginalne propozycje zamieszczono w kolejnych trzech rozdziałach. Rozdział 3 przedstawia koncepcję USVM - adaptacji algorytmu SVM do rozwiązywania problemu modelowania różnicowego. Autor prezentuje nietrywialną, moim zdaniem, modyfikację oryginalnego algorytmu oraz formułuje problem optymalizacyjny na potrzeby uczenia modelu. W kolejnym podrozdziale Autor wprowadza kilka lematów przedstawiających najważniejsze z punktu widzenia modelowania różnicowego cechy algorytmu USVM. W tym miejscu pragnę podkreślić, że sposób przedstawienia algorytmu USVM oraz jego cech jest wzorowy: wywód jest przejrzysty, spójny, zrozumiały, starannie przygotowane ryciny czytelnie ilustrują opisywane pojęcia (jak choćby praktyczną interpretację ilorazu współczynników kary). W tym samym rozdziale Autor wprowadza modyfikację algorytmu która zapobiega zjawisku pojawienia się nieskończonego marginesu między klasami. Rozdział 4 w całości poświęcono zagadnieniu optymalizacji, której wynikiem są ostateczne definicje hiperpłaszczyzn decyzyjnych. Mimo, że w tej części pracy wywód matematyczny jest miejscami bardzo zawiły, to przyznać należy, że Autor w sposób staranny i umiejętny przedstawia ten wywód, opatrując go komentarzami i wyjaśnieniami w taki sposób, że cała prezentacja jest przystępna i zrozumiała. Rozdział 5 wprowadza metodę regularyzacji, która ma zapobiegać problemowi nie w pełni losowego przydziału instancji do obu zbiorów (zbioru uczącego, zawierającego instancje poddane pewnemu oddziaływaniu, oraz zbioru kontrolnego). Przedstawiona w pracy metoda jest nie tylko pomysłowa, ale też zachwyca prostotą i elegancją. W rozdziale 6 Autor zamieścił sprawozdanie z przeprowadzonych eksperymentów, opisując wykorzystane zbiory danych, przedstawiając szczegółowo protokół eksperymentu, prezentując wyniki i ich omówienie. Rozprawę zamyka rozdział 7 zawierający krótkie podsumowanie.

Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle, świadczący o dostatecznej wiedzy autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonujący?

Aktualny stan wiedzy w obszarze modelowania różnicowego został przedstawiony w rozdziale 1. Przegląd literaturowy nie jest obszerny, ale też przyznać należy, że domena modelowania różnicowego, mimo bardzo dużego znaczenia praktycznego, nie była przedmiotem licznych badań naukowych. W zamieszczonym w rozprawie przeglądzie Autor odnosi się do najważniejszych prac w domenie, opatrując je komentarzem i wskazując na związek z badaniami prezentowanymi w rozprawie. Taki zabieg ułatwia ocenę oryginalności prezentowanych w rozprawie propozycji. Zamieszczone w bibliografii odnośniki literaturowe odnoszą się w znakomitej większości do prac opublikowanych na przestrzeni ostatnich lat w czołowych periodykach i konferencjach tematycznych. Poza podrozdziałem 1.4 referencje bibliograficzne pojawiają się też w innych miejscach tekstu, szczególnie w rozdziałach prezentujących opracowane metody optymalizacji uczenia maszyn wektorów wspierających. Lektura rozdziału nie pozostawia najmniejszych wątpliwości odnośnie wiedzy Autora w obszarze poruszanym w rozprawie. Autor nie tylko biegle porusza się w obszarze modelowania różnicowego, ale też precyzyjnie i poprawnie posługuje się terminami związanymi z uczeniem maszynowym, o czym dobitnie świadczy choćby

POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI
ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcf@put.poznan.pl, www.put.poznan.pl

staranność z jaką zaprojektowano ewaluację eksperymentalną proponowanych w rozprawie algorytmów. Na pochwałę zasługuje też bardzo jasne i jednoznaczne przedstawienie oryginalnego wkładu rozprawy w dyscyplinę, zawarte w podrozdziale 1.4. Pozwolę sobie dodać, że w mojej opinii wkład ten jest istotny.

Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?

Nie mam najmniejszych wątpliwości, że Autor w sposób naukowo poprawny, rygorystyczny i przekonujący rozwiązał postawiony w rozprawie problem. Przedstawiona w rozdziale 6 eksperymentalna ewaluacja algorytmu USVM jest obszerna, wykorzystując prawie 30 zbiorów danych i porównując algorytm USVM z pięcioma alternatywnymi algorytmami. Autor nie tylko wykazuje wyższość zaproponowanej metody nad aktualnie znanymi, ale też interesująco ilustruje wpływ poszczególnych parametrów algorytmu na uzyskiwane wyniki. Co ważne, Autor nie ukrywa przypadków, w których algorytm USVM wydaje się działać gorzej od znanych metod, ale wyraźnie je wskazuje i tłumaczy, co powoduje obniżenie jakości algorytmu USVM. Warto w tym miejscu jeszcze podkreślić, że samo opracowanie protokołu eksperymentu było zadaniem nietrywialnym. Mimo, że oryginalny algorytm SVM jest najczęściej wykorzystywany w zadaniach klasyfikacji (ew. regresji lub klasyfikacji jednoklasowej) i istnieje zestaw powszechnie wykorzystywanych miar oceny jakości algorytmu, to w przypadku algorytmu USVM te miary nie znajdują zastosowania. W związku z tym Autor musiał wybrać inny sposób oceny algorytmu. Moim zdaniem wybór poczyniony przez Autora był trafny. Także w przypadku weryfikacji użyteczności zaproponowanej metody regularyzacji modelu Autor musiał opracować protokół eksperymentu i wywiązał się z tego zadania wzorowo. Poczynione założenia (np. decyzję o usunięciu tych atrybutów, których rozkłady różniły się pomiędzy zbiorem uczącym i kontrolnym) uważam za w pełni uzasadnione. W podsumowaniu stwierdzam, że część eksperymentalna pracy jest pozbawiona błędów metodologicznych i w sposób przekonujący dowodzi skuteczności prezentowanych metod.

Z drugiej strony nie sposób nie wspomnieć o rozwiązaniach analitycznych prezentowanych we wcześniejszych rozdziałach. Przedstawione tam wywody jasno pokazują użyteczność i poprawność opracowanych algorytmów. Uważam, że Autor rozwiązał postawiony problem nie tylko w sposób zadowalający, ale raczej w sposób kompletny. Wychodząc od problemu, przedstawił model, uzasadnił model w sposób teoretyczny, zaprezentował algorytm optymalizujący którego rozwiązanie stanowi model różnicowy, a na końcu przeprowadził praktyczną weryfikację. Można tylko sobie życzyć, aby więcej prac w obszarze informatyki podchodziło do rozwiązywania problemów obliczeniowych w sposób tak wyczerpujący.

POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI
ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcf@put.poznan.pl, www.put.poznan.pl

Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?

Główną wartością rozprawy jest niezwykle eleganckie zaadaptowanie znanego algorytmu uczenia maszynowego do rozwiązania trudnego problemu modelowania różnicowego. W pełni podzielam przekonanie Autora o dużym znaczeniu teoretycznym i praktycznym tego problemu. Tym bardziej cieszy mnie przedstawiona w rozprawie propozycja, która ma szanse przyciągnąć większą uwagę do tego obszaru uczenia maszynowego. Do głównych osiągnięć Autora zaliczam:

- adaptację metodyki SVM do problemu modelowania różnicowego,
- sformułowanie problemu optymalizacyjnego na potrzeby modelu USVM,
- przedstawienie teoretycznych przesłanek efektywności modelu,
- usprawnienie modelu poprzez jego stabilizację i zmniejszenie czułości modelu na niewielkie zmiany parametrów wejściowych,
- wprowadzenie nowej metody regularyzacji modelu minimalizującej skutki nielosowego przydziału instancji do zbioru uczącego i kontrolnego.

Każde ze wspomnianych wyżej osiągnięć jest istotne i oryginalne. Zaprezentowane w rozprawie rozwiązanie jest pierwszą propozycją, która w sposób jawny dokonuje przypisania do jednej z trzech klas (+1, 0, -1), a stojąca za algorytmem USVM teoria jest niebanalną adaptacją metodyki SVM. Oryginalność i znaczenie zaprezentowanych w rozprawie wyników z nadmiarem wyczerpują wymagania stawiane rozprawom doktorskim.

Czy autor wykazał umiejętność poprawnego i przekonującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Praca jest napisana wzorowo. Tekst jest całkowicie wolny od błędów stylistycznych, gramatycznych czy ortograficznych. Styl wywodu jest niezwykle jasny i przejrzysty, struktura pracy logicznie prowadzi czytelnika od jednego omawianego pojęcia do drugiego, wszystkie rozdziały układają się w spójnie skonstruowaną całość. Widać wyraźnie niezwykłą dbałość Autora o poprawność i spójność tekstu. Mimo, że fragmenty rozprawy zawierają długie i zawiłe wywody matematyczne, przyjęty sposób oznaczeń i konsekwencja ich stosowania, oraz właściwie dobrane komentarze tekstowe objaśniające kolejne kroki przekształceń matematycznych wyśmienicie ułatwiają lekturę. Dawno nie obcowałem z tekstem rozprawy doktorskiej który, mówiąc kolokwialnie, "czytałby się" tak dobrze i łatwo. Warto tu też wyraźnie powiedzieć, że stosunkowo niewielka objętość rozprawy w niczym nie umniejsza jej znaczenia czy jakości. Wręcz przeciwnie, uważam, że Autor zawarł w tekście wszystkie najważniejsze elementy, a żaden fragment tekstu nie jest zbędny.

POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI
ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcf@put.poznan.pl, www.put.poznan.pl

Jakie są słabe strony rozprawy i jej główne wady?

To może zabrzmieć nadmiernie optymistycznie, ale w mojej ocenie praca jest pozbawiona wad. Od początku do końca jest to spójne, staranne i czytelne zaprezentowanie ciekawych i oryginalnych wyników naukowych. Jedyna "literówka" na jaką natknąłem się w tekście występuje na stronie 7 ("controlled trails", powinno być "controlled trials"), poza tym na stronie 52 zamiast d=1 powinno być chyba α=1. W związku z tym zamieszczone poniżej uwagi nie mają charakteru krytycznego, lecz stanowią zaproszenie do krótkiej dyskusji

- W rozdziale 6 do oceny użyteczności nowej metody regularyzacji wykorzystano zbiór RHC, który wydaje się być idealnym zbiorem do testowania modeli różnicowych. Jednak w rozdziale 6.4 nie natknąłem się na użycie tego zbioru danych, czy jest jakiś konkretny powód, dla którego ten zbiór został wyłączony z ewaluacji?
- Intuicja podpowiada mi, że prezentowana w rozprawie metoda mogłaby znaleźć też zastosowanie w domenie systemów rekomendacyjnych (a przy okazji zostać przetestowana na licznych zbiorach danych, które środowisko badaczy systemów rekomendacyjnych zgromadziło na przestrzeni ostatnich lat). Chciałbym prosić Autora o odniesienie się do tego pomysłu.
- Tradycyjnie oczekuje się, że w pierwszym rozdziale rozprawy doktorskiej zostanie sformułowana teza, która w dalszej części rozprawy zostanie potwierdzona lub odrzucona. W recenzowanej rozprawie brak jest jednoznacznie sformułowanej tezy (co niekoniecznie jest wadą, bo oryginalna kontrybucja jest oczywista dla czytelnika), jednak prosiłbym Autora o próbę sformułowania zwięzłej tezy.
- Wprowadzona metoda regularyzacji bazuje na porównywaniu rozkładów cech w zbiorze uczącym i kontrolnym, przy czym do porównania rozkładów jest wykorzystywana odległość energetyczna rozkładów. Nie jest to, jak powszechnie wiadomo, jedyna metoda porównania dwóch rozkładów. W rozprawie zabrakło mi uzasadnienia wyboru tej akurat funkcji odległości między rozkładami i chętnie dowiedziałbym się, czy w opinii Autora efektywność regularyzacji uległaby zmniejszeniu przy wykorzystaniu np. odległości Mahalanobisa.

Jaka jest przydatność rozprawy dla nauk technicznych?

Recenzowana praca przedstawia niezwykle kompletną i dojrzałą propozycję zaadaptowania znanej metody uczenia maszynowego do rozwiązania trudnego problemu modelowania różnicowego. W rozprawie wprowadzono oryginalne rozwiązanie, przedstawiono elegancki formalizm matematyczny, zaproponowano metody optymalizacyjne i dowiedziono ich poprawności. Sama metoda została poddana ewaluacji eksperymentalnej, której wyniki Autor zrecenzował w sposób krytyczny. Lektura pracy utwierdza mnie w przekonaniu, że Autor znakomicie orientuje się w omawianym obszarze, biegle posługuje się szerokim wachlarzem metod uczenia maszynowego i algebry, oraz bardzo dobrze opanował warsztat naukowy. Jestem przekonany, że przedstawione w rozprawie propozycje są bardzo wartościowe i spotkają się z dużym zainteresowaniem środowiska naukowego. Może o tym świadczyć fakt przyjęcia

POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI
ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcf@put.poznan.pl, www.put.poznan.pl

publikacji prezentujących częściowe wyniki prac badawczych przeprowadzonych w ramach przygotowywania rozprawy doktorskiej do druku w renomowanym czasopiśmie "*Knowledge and Information Systems*" oraz w materiałach warsztatu organizowanego przy prestiżowej konferencji ICDM.

W podsumowaniu recenzji z przekonaniem stwierdzam, że rozprawa doktorska Łukasza Zaniewicza spełnia z nadmiarem wymagania stawiane rozprawom doktorskim przez obowiązującą Ustawę o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 r. (Dz.U. Nr 65 poz. 595 z późniejszymi zmianami) i wnoszę o jej dopuszczenie do publicznej obrony. Ze względu na bardzo wysoki, w mojej ocenie, poziom merytoryczny recenzowanej rozprawy, rekomenduję jej wyróżnienie.