Warszawa, 30.01.2017

Prof. Dr hab. Henryk Rybinski
Instytut Informatyki Politechniki Warszawskiej
hrb@ii.pw.edu.pl

Recenzja rozprawy doktorskiej
mgr Mirona Bartosza Kursy
p/t. „Robust and Efficient Approach to Feature Selection and Machine Learning”

Dane ogólne

Przedstawiona do recenzji rozprawa jest napisana w języku angielskim. Składa się ze streszczenia w języku polskim i angielskim oraz pięciu rozdziałów i krótkiego podsumowania. Rozdział pierwszy jest poświęcony wprowadzeniu w problematykę, motywacji autora oraz omówieniu poszczególnych metod omawianych w pracy. W rozdziałach 2-5 zawarte są publikacje doktoranta (dwie autorskie i 3 współautorskie):

1. Rozdział 2 stanowi publikacja współautorska

   Kursa, Miron B., and Witold R. Rudnicki. "Feature Selection with the Boruta Package." Journal of Statistical Software 36.11 (2010): 1-13.

2. Rozdział 3 zawiera publikację

   Kursa, Miron B. "rFerns: An Implementation of the Random Ferns Method for General-Purpose Machine Learning." Journal of Statistical Software 61.i10 (2014).

3. W rozdziale 4 zawarta jest publikacja

   Kursa, Miron Bartosz. "Robustness of Random Forest-based gene selection methods." BMC Bioinformatics 15.1 (2014): 8.

4. Rozdział 5 zawiera 2 prace współautorskie:

   Wieczorkowska, Alicja A., and Miron B. Kursa. "A comparison of random forests and ferns on recognition of instruments in jazz recordings." International Symposium on Methodologies for Intelligent Systems. Springer Berlin Heidelberg, 2012.

   Kursa, Miron B., and Alicja A. Wieczorkowska. "Multi-label Ferns for Efficient Recognition of Musical Instruments in Recordings." International Symposium on Methodologies for Intelligent Systems. Springer International Publishing, 2014.

Łączna objętość rozprawy wynosi 76 stron. Ponadto rozprawie towarzyszą 3 oświadczenia wpółautorów o ich udziale (dotyczy trzech zamieszczonych w doktoracie publikacji współautorskich).

1. Problem badawczy i jego znaczenie

1.1 Problem rozważany w rozprawie

Tematyka rozprawy ściśle wiąże się z metodami odkrywania wiedzy ze zbiorów danych. Dziedzina ta jest od dawna w obszarze zainteresowań statystyków, nieco krócej w obszarze zainteresowań naukowców zajmujących się metodami sztucznej inteligencji, w szczególności metodami uczenia maszynowego. Zagadnienie to ma ogromne znaczenie, zarówno teoretyczne, jak też praktyczne, w szczególności w kontekście systemów akwizycji wiedzy, maszynowego uczenia, a także metod wnioskowania indukcyjnego przez analizę przykładów. Dlatego też od ponad 20-u lat daje się zaobserwować gwałtowny wzrost zainteresowań badaniami nad nowymi metodami analizy danych. Zainteresowania te dotyczą nie tylko środowisk akademickich, ale także przemysłowych laboratoriów badawczych. Wynika to przede wszystkim z zapotrzebowania na narzędzia w dziedzinie analizy dużych zasobów informacyjnych.

Jednym z istotnych problemów badawczych w dziedzinie analizy danych jest problem selekcji cech analizowanych obiektów. Tematyka ta jest szczególnie ważna z uwagi na ograniczenia obliczeniowe eksploracji danych i uczenia maszynowego z jednej strony, zaś z drugiej strony ze względu na ryzyko nadmiernego zwiększenia wpływu fałszywych wzorców na wynik, bądź czysto losowych zależności, co w efekcie może prowadzić do fałszywych rezultatów. Opiniowana praca mieści się w nurcie badań nad metodami doboru cech analizowanych obiektów. Celem, jaki postawił sobie doktorant, jest opracowanie stabilnych i wydajnych metod doboru cech.

Motywując swoje badania autor odniósł się we wstępie do problemów rozwiązań związanych doborem cech znanych z literatury. W szczególności autor stawia tezę, że tradycyjne podejście do problemu wyboru atrybutów, polegające na zidentyfikowaniu minimalnego zbioru atrybutów, prowadzącego do uzyskania najlepszej dokładności algorytmu klasyfikacji nie zawsze prowadzi do interesujących wniosków, przede wszystkim z uwagi na zależności przypadkowe, w szczególności w przypadku dużej wymiarowości danych, ponadto może prowadzić do usuwania istotnej informacji.

Tematyka rozważana przez doktoranta jest niezwykle ważna i niewątpliwie jest godna rozprawy doktorskiej.

1.2 Charakter rozprawy i znaczenie praktyczne badań

Rozprawa łączy cechy pracy eksperymentalnej i teoretycznej. Zaawansowany teoretyczny warsztat autora pozwala doktorantowi konstruować nowe autorskie rozwiązania algorytmiczne. Z drugiej strony algorytmiczne propozycje autora są wsparte zaawansowanymi implementacjami oraz eksperymentami. Rozprawa wskazuje na

1. bardzo solidny warsztat doktoranta w zakresie podstaw teoretycznych metod i narzędzi uczenia maszynowego;
2. ogromny potencjał praktycznego wdrażania opracowanych algorytmów;
3. szeroką wiedzę w zakresie możliwości zastosowań metod analizy danych; pozwoliło to autorowi przeprowadzić bardzo interesujące eksperymenty w różnych dziedzinach: W Rozdziale 2 jest przykład eksperymentów z realnymi danymi dotyczącymi warstwy ozonowej, jak też z benchmarkiem w postaci danych Madelon spreparowanym na potrzeby eksperymentów uczenia maszynowego; w Rozdziale 4 autor omawia zastosowanie algorytmu rFerns na potrzeby eksploracji danych genetycznych, zaś w Rozdziale 5 prezentuje możliwości zastosowania opracowanych metod do analizy danych muzycznych.

W szczególności, punkty (1) i (2) są wsparte treścią rozdziałów 2 i 3. W Rozdziale 2 autor koncentruje się na omówieniu algorytmu Boruta. Metoda Boruta pozwala uniknąć problemów charakterystycznych dla tradycyjnego podejścia do problemu selekcji cech, takich jak np. utrata istotnych informacji. Tu głównym osiągnięciem autora jest efektywne wdrożenie tego algorytmu. Pomimo intersujących własności tego algorytmu istotnym jego ograniczeniem są wymagania obliczeniowe. Dlatego w Rozdziale 3 autor proponuje metodę rFerns, która jest uogólnieniem metody paproci losowych. Istotnym osiągnięciem autora jest pokazanie, że z użyciem opracowanej metody rFerns można dokonywać selekcji atrybutów metodą Boruta na obszernych zbiorach danych w rozsądnym czasie.

2. Wkład autora

Omawiając wkład autora ograniczam się tylko do prac zamieszczonych w doktoracie. Zastrzeżenia to ma o tyle znaczenie, że dorobek publikacyjny doktoranta jest znaczący i

obejmuje uczestnictwo w wielu międzynarodowych badaniach, gdzie doktorant realizował badania w zakresie eksploracji danych.

Wkład autora w zakresie przedstawionego doktoratu obejmuje szereg ważnych elementów związanych z uczeniem maszynowym, w szczególności z problemem doboru atrybutów. Wyróżnić tu można kilka problemów, którymi autor się zajmuje, i dla których proponuje interesujące rozwiązanie. Zasadniczym wkładem autora jest:

1. zaproponowanie heurystycznego *wrappera* klasy *all relevant* w formie metody Boruta;
2. zaproponowanie rozwinięcia metody Boruta poprzez wykorzystanie paproci losowych (rFerns);
3. przeprowadzenie eksperymentów w kilku różnych dziedzinach w oparciu o metody Boruta i rFerns

Podejście Boruta bazuje na koncepcji rozszerzenia systemu informacyjnego o tzw. cienie (atrybuty nieistotne), które wykorzystuje się do oceny istotności atrybutów oryginalnych. Jego przydatność została poparta eksperymentami.

Problemy z wymaganiami obliczeniowymi podejścia Boruta skłaniają autora do badań nad jego modyfikacją. W efekcie autor proponuje w Rozdziale 3 metodę rFerns, jednocześnie pokazuje jej przydatność do rozwiązania niezwykle trudnego zagadnienia klasyfikacyjnego klasy $p >> n$. W szczególności pokazana jest duża przydatność metody w procedurze selekcji genów w zbiorze danych mikromacierzy (Rozdział 4). Dodatkowo w Rozdziale 5 autor pokazuje przydatność metody rFerns do rozpoznawania instrumentów w nagraniach audio.

Omawiane tu osiągnięcia zawarte są w publikacjach z lata 2010-2014. Jak zaznaczyłem na wstępie, dorobek autora w przypadku 3ch publikacji współautorskich jest potwierdzony oświadczeniami współautorów.

3. Poprawność

Wysoko oceniam przyjętą przez autora metodologię badań. Autor dokonuje krytycznej analizy stanu badań w dziedzinie doboru atrybutów, w oparciu o tę analizę proponuje rozwiązania, a następnie przeprowadza badania eksperymentalne. Badania te stanowią istotny element rozprawy. Autor bada skonstruowane algorytmy i miary na danych rzeczywistych.

Uzyskane eksperymentalnie wyniki potwierdzają słuszność proponowanej koncepcji nie tylko co do jakości klasyfikacji ale też stabilności.

4. Redakcja pracy

Praca jest przygotowana starannie. Zamieszczone teksty są napisane dobrym językiem angielskim (choć kilka drobnych usterek można znaleźć, co oznacza, że nawet w przypadku dobrych czasopism nie należy liczyć na gruntowną redakcję tekstu przez wydawcę i warto weryfikować język z anglistą).

Na szczególną uwagę zasługuje układ pracy. Pomimo tego, że praca jest złożeniem prac opublikowanych na przestrzeni lat 2010-2014, ich dobór jest logiczny i świetnie podporządkowany tezie rozprawy.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami) moja ocena rozprawy jest pozytywna. W szczególności uważam, że rozprawa z nadmiarem spełnia te wymagania, dlatego wnoszę o dopuszczenie pana Mirona Kursy do publicznej obrony. Jednocześnie uważam, że rozprawa zasługuje na wyróżnienie.