Wydział Automatyki, Elektroniki i Informatyki
Katedra Sieci i Systemów Komputerowych

dr hab. inż.
Prof. Politechniki Śląskiej
Aleksandra Gruca

Gliwice, 06.11.2023

Recenzja rozprawy doktorskiej mgr-a Radosława Piliszka

Development of methods for feature selection based on information theory

Ukończonej na Wydziale Informatyki i Telekomunikacji
Politechniki Poznańskiej

Pod opieką promotora dr hab. Witolda Rudnickiego, prof. UwB

Tematyka i cel pracy, problem badawczy i jego znaczenie

Przedstawiona mi do recenzji rozprawa doktorska powstała na Wydziale Informatyki i Telekomunikacji Politechniki Poznańskiej pod kierunkiem dr hab. Witolda Rudnickiego. Praca podejmuje istotny problem selekcji cech w uczeniu maszynowym. Tematyka ta jest bardzo rozległa na co wskazuje duża liczba pojawiających się co roku publikacji prezentujących nowe i ulepszone metody selekcji cech. Ponadto, w związku z coraz większą popularnością metod maszynowego uczenia i tak zwanej „sztucznej inteligencji" coraz więcej dziedzin nauk – nie tylko ścisłych – eksploruje możliwości związane z zastosowaniem takich metod, gdzie selekcja cech często odgrywa kluczową rolę. W temacie selekcji cech należy też zwrócić uwagę na istotny, szczególnie w badaniach biomedycznych, element stabilności wyników. Jest to problem istotny z dwóch głównych powodów. Po pierwsze, etap selekcji cech stanowi tylko początkowy krok w kierunku pełnego zrozumienia zjawiska, które wymaga dalszej pracy eksperckiej, która jest czasochłonna i kosztowna. Dlatego ryzyko otrzymywania tak zwanych "mylących wskazówek" powinno być możliwie ograniczane. Po drugie, ze względu na biomedyczny charakter odkryć, pojawiają się obawy dotyczące odpowiedzialności związanej z ich zastosowaniem. Jeśli wyniki są niestabilne, istnieje duże ryzyko wyciągnięcia błędnych wniosków, przy czym dodatkowo okazuje się również, że wybór cech ma istotniejsze znaczenie dla wyników klasyfikacji niż dostrojenie hyperparamterów wykorzystanego modelu.

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Katedra Sieci i Systemów Komputerowych
ul. Akademicka 16, pok. 414, 44-100 Gliwice
+48 32 237 21 51 / +48 237 27 33
Aleksandra.gruca@polsl.pl

NIP 631 020 07 36
ING Bank Śląski S.A. o/Gliwice 60 1050 1230 1000 0002 0211 3056

Cele zdefiniowane w recenzowanej rozprawie doktorskiej są następujące:

- Wdrożenie podejścia do odkrywania najlepszych (minimalnie optymalnych) cech nie ingerując w klasyfikator.

- Zaproponowanie miary odmienności cech z uwzględnieniem zmiennej decyzyjnej.

- Zaproponowanie protokołu walidacji wyboru cech pod względem ich stabilności.

- Wykorzystanie i ulepszenie istniejących podejść opartych na teorii informacji (MDFS; Multi Dimensional Feature Selection)

Charakterystyka rozprawy

Praca doktorska została napisana w języku angielskim i składa się z siedmiu rozdziałów. Układ rozprawy jest typowy dla tego rodzaju opracowań.

Rozdział pierwszy stanowi wprowadzenie do tematyki pracy i przedstawia jej cel.

Rozdział drugi pracy stanowi wprowadzenie teoretyczne do podjętej tematyki i zawiera kompleksowy przegląd zagadnień związanych z tematem pracy. Zaczyna się on od krótkiego przeglądu dziedzin uczenia maszynowego, a następnie skupia się na problemie redukcji wymiarowości i selekcji cech, w szczególności w klasyfikacji z binarną decyzją. Omówione zostały różne podejścia do selekcji cech, w tym porównanie metod wyszukiwania wszystkich zmiennych istotnych i minimalno-optymalnego zbioru cech, wraz z opisem filtrów, metod opakowujących i metod osadzonych. W dalszej części rozdziału opisane są metody grupowania, a także przedstawiono wybrane pojęcia z teorii informacji istotnych w kontekście metod selekcji cech, na których skupia się praca doktorska. Rozdział ten obejmuje również kilka zagadnień technicznych wykorzystanych w pracy, takich jak problem wielokrotnego testowania, miary stabilności klastrów i miary wydajności klasyfikatorów.

Rozdziały trzeci i piąty to najważniejsze rozdziały rozprawy przedstawiający oryginalny wkład autora. Rozdział trzeci pracy skupia się na przedstawieniu nowej miary odmienności – symetrycznego wzrost informacji o celu STIG (ang. *Symmetric Target Information Gain*). Miara ta została już wcześniej wprowadzona w pracy *Sotoca & Pla, 2010* jednak doktorant słusznie zauważa, że przedstawiony przez autorów pracy dowód, iż miara jest metryką oparty jest na błędnych założeniach i w związku z tym miara ta nie spełnia warunku nierówności trójkąta. Doktorant w rozdziale trzecim przedstawia dowód matematyczny oraz przykład liczbowy obrazujący ten fakt.

Kolejny, czwarty rozdział przedstawia ulepszenia w bibliotece metod selekcji cech – MDFS (ang. *MultiDimensional Feature Selection*). Ulepszenia te obejmują implementację dodatkowych miar opartych o teorię informacji, poprawę wartości domyślnego parametru zakresy wykorzystywanego do dyskretyzacji,

stworzenie uniwersalnego silnika biblioteki oraz optymalizację wykorzystania zmiennych kontrastowych.

W rozdziale piątym Autor przedstawił nowy algorytm minimalno-optymalnej selekcji cech RAFS (ang. *Robust Aggregative Feature Selection*). Głównym założeniem RAFS jest wykorzystanie grupowania w celu znalezienia cech reprezentatywnych. Aby ograniczyć problem niestabilności wyników metoda stosuje schemat zagnieżdżonej wielokrotnej walidacji krzyżowej. W przedstawionym algorytmie zastosowano miarę STIG, jakkolwiek metoda ta może wykorzystywać dowolną miarę odmienności. Algorytm RAFS został zaimplementowany w języku R i udostępniony pod postacią pakietu biblioteki R. Doktorant w rozdziale piątym zawarł również szczegółowe opisy funkcji wraz z fragmentami kodu implementującego funkcje wykorzystane do przetestowania biblioteki. W mojej opinii są to informacje techniczne i jako takie powinny zostać umieszczone w dodatkowych załącznikach, a nie w głównej części dysertacji.

Szósty rozdział pracy przedstawia zastosowanie algorytmu RAFS do analizy trzech rzeczywistych zbiorów danych. Doktorant przedstawia zbiory danych, na których przeprowadzono eksperymenty, oraz metody oceny skuteczności selekcji cech. Następnie przedstawia wyniki porównania zaproponowanego algorytmu dla różnych parametrów selekcji cech oraz wyniki analizy stabilności wybranych klastrów.

Ostatni rozdział zawiera wnioski.

Opinia o rozprawie

Należy podkreślić, iż cel pracy jest jasno sformułowany a problem, który podjął się rozwiązać doktorant jest niewątpliwie ważny i wpisujący się w trendy najnowszych badań w dziedzinie. Przedstawione rozwiązanie ma wysokie zastosowanie praktyczne. Dodatkowo pozytywnie oceniam fakt, że zarówno zaproponowana miara jak i nowa metoda zostały udostępnione środowisku naukowemu w postaci bibliotek. Wyniki zaprezentowane pracy zostały opublikowane trzech publikacjach naukowych (w dwóch z nich doktorant jest pierwszym autorem), dwie kolejne znajdują się w przygotowaniu.

Cześć teoretyczna pracy, omówienie problemu i przegląd stosowanych metod pokazują wystarczająco głęboką i aktualną wiedzę Doktoranta dotyczącą tematyki podjętej w rozprawie doktorskiej. Bibliografia zawierająca 191 pozycji jest odpowiednio dobrana.

Rozdział drugi pracy zawiera opis najważniejszych algorytmów klasycznego uczenia, podejść do selekcji cech oraz miar niepewności na bazie teorii informacji. Z uwagi na rozległość tematyki, przedstawienie jej najważniejszych elementów w sposób jasny oraz skondensowany nie jest prosty, i w mojej doktorant poradził sobie w tym zakresie bardzo dobrze, a sam rozdział czyta się z przyjemnością. Najważniejsze wyniki rozprawy stanowiące oryginalny wkład autora zaprezentowano w rozdziale trzecim, gdzie Doktorant omawia miarę odmienności STIG i przedstawia dowód na niespełnienie warunku trójkąta, a także w rozdziale piątym

prezentującym nowy algorytm minimalno-optymalnej selekcji RAFS. Skuteczność nowej metody potwierdzona jest analizami rzeczywistych zbiorów danych, które stanowią zawartość rozdziału szóstego.

Przedstawiony w rozprawie cel jest jasno sformułowany, a zaprezentowane wyniki świadczą o tym, że Doktorantowi udało się go osiągnąć.

Uwagi krytyczne i dyskusyjne

Na początku tej części chciałbym podkreślić, że nie znalazłam w przedstawionych wynikach żadnych zasadniczych błędów merytorycznych czy niewłaściwych rozumowań. Wszystkie poniższe uwagi wynikają z chęci podjęcia dyskusji i dialogu na temat niektórych aspektów pracy. Uwagi te nie obniżają mojej pozytywnej oceny pracy.

- W rozdziale szóstym Doktorant porównuje wyniki zastosowania algorytmu RAFS z różnymi parametrami, w tym wartości miary AUC różnych algorytmów klasyfikacji dla top *n* cech. Myślę, że wartościowe byłoby porównanie wyników klasyfikacji nie tylko dla różnych parametrów metody RAFS, ale z również z wynikami istniejących w literaturze metod *state-of-the-art* selekcji cech na przykład z wynikami metody Boruta, która na pewno bardzo dobrze jest znana Doktorantowi.

- Rozdział czwarty, gdzie Doktorant przedstawia usprawnienia biblioteki MFDS w mojej ocenie nie do końca wpisuje się w logiczną strukturę logiczną rozprawy. W mojej opinii skupia się mocno na aspektach technicznych (opis usprawnień implementacyjnych) i jako taki powinien być raczej umieszczony w końcowej części pracy, a być może nawet w aneksie.

- Czy Doktorant może podać jaka jest teoretyczna złożoność obliczeniowa nowej metody RAFS? Sądzę też, że warto byłoby również podać czasy wykonywanych obliczeń dla porównywanych podejść.

- Czy przedstawione różnice w mierze AUC pomiędzy poszczególnymi metodami są istotne statystycznie?

- Dla zbioru danych PLDT w pracy dokonana została analiza znaczenia biologicznego genów wybranych przez metodę selekcji cech. Czy byłoby wartościowe wykonanie podobnych analiz dla dwóch pozostałych zbiorów danych?

Podsumowanie

Stwierdzam, że pan mgr Radosław Piliszek przedstawił rozprawę doktorską rozwiązującą aktualny problem naukowy, która przyczyni się do rozwoju reprezentowanej dyscypliny naukowej. Rozprawa zawiera

oryginalne rozwiązanie problemu naukowego, a kandydat wykazał, że zarówno posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja oraz umiejętność prowadzenia pracy naukowej.

Biorąc pod uwagę opinie zaprezentowane w powyżej, moja ocena rozprawy pod względem trzech podstawowych kryteriów Ustawy jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego? (wybierz jedną opcję stawiając znak **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka techniczna i telekomunikacja?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Mając na uwadze powyższe, stwierdzam, że przedstawiona do oceny praca doktorska w pełni odpowiada warunkom określonym w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2023 r. poz. 742 z późn. zm.) i na tej podstawie wnoszę do Wysokiej Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Poznańskiej o dopuszczenie mgr Radosława Piliszka do dalszych etapów przewodu doktorskiego.

Signed by / Podpisano przez:
Aleksandra Helena Gruca
Politechnika Śląska
Date / Data:
2023-11-06 23:05

dr hab. inż. Aleksandra Gruca