Prof. dr hab. Dominik Ślęzak
Wydział MIM, Uniwersytet Warszawski
ul. Banacha 2, 02-097 Warszawa
e-mail: slezak@mimuw.edu.pl

2023-10-30

Recenzja rozprawy doktorskiej

mgr. Radosława Piliszka

zatytułowanej:

Development of methods for feature selection based on information theory

1. Problem badawczy i jego znaczenie

Recenzowana rozprawa dotyczy zagadnienia selekcji cech w uczeniu maszynowym. Przedstawione w rozprawie badania koncentrują się na minimalno-optymalnym podejściu do selekcji cech, którego celem jest znalezienie najmniejszego zbioru cech pozwalającego jak najdokładniej opisać badane zjawisko. W ogólności jest to zadanie NP-trudne, zatem zastosowanie metod heurystycznych ma tutaj kluczowe znaczenie. Opracowana w rozprawie metoda ma postać binarnej stabilnej klasyfikacji hierarchicznej bazującej na teorii informacji wykorzystującej informację dotyczącą zmiennej decyzyjnej. Stabilność klasyfikatora zapewnia, iż niewielka zmiana informacji na wejściu klasyfikatora nie powoduje znaczącej zmiany informacji na jego wyjściu. W opracowanej metodzie redukcja wymiarów przeprowadzana jest z wykorzystaniem selekcji cech za pomocą hierarchicznego grupowania, w którym cechy organizowane są w rozłączne grupy, tzn. indukowany jest podział przestrzeni cech, a następnie konstruowany jest selektor dla tego podziału, tzn. z każdej grupy wybierana jest jedna cecha. W ten sposób konstruowany jest minimalny zbiór cech możliwie dokładnie opisujący daną binarną zmienną decyzyjną.

Motywacji do badań przedstawionych w rozprawie dostarcza troska o stabilność modeli klasyfikacji stosowanych w badaniach biomedycznych. Warto zaznaczyć, że klasyfikacja danych biomedycznych jest niezwykle ważnym społecznie zastosowaniem metod uczenia maszynowego. Zbyt mała stabilność wyników klasyfikacji może skutkować ryzykiem postawienia przez lekarza błędnej diagnozy. Tymczasem wybór cech ma często większy wpływ na odchylenia wyników klasyfikacji niż jakże modne dostrajanie parametrów modelu. Co więcej wybór cech, na których zbudowany jest model, może mieć znaczący wpływ na to, jak eksperci dziedzinowi będą ten model postrzegać i czy będą ufać jego działaniu.

Kolejna istotna motywacja dla badań przedstawionych w rozprawie płynie z konieczności szybkiej eksploracji bardzo dużych i dynamicznie przyrastających zbiorów danych. Odpowiedzią mgr. Piliszka były w tym przypadku badania mające na celu zwiększanie wydajności opracowywanych metod oraz zmniejszanie ich złożoności obliczeniowej. W ten nurt wpisuje się wspomniana wyżej metoda redukcji liczby wymiarów z wykorzystaniem hierarchicznego grupowania cech. Metoda ta może być uznana za przykład rozwiązania zadania nadzorowanej selekcji cech (ang. *supervised feature selection*). Łączy ona w istocie dążenie do zwiększenia stabilności procesu selekcji z przyspieszeniem tego procesu.

Ślęzak

2. Wkład autora

Badania przedstawione w recenzowanej rozprawie sytuują się na pograniczu zagadnienia selekcji cech oraz teorii informacji. Wśród osiągnięć przedstawionych w rozprawie znajdują się zarówno wyniki teoretyczne, jak i empiryczne. Do osiągnięć teoretycznych należy zaproponowanie przez mgr. Piliszka nowej miary odmienności cech, gdzie odmienność cech jest ograniczona do zmiennej decyzyjnej, oraz wprowadzenie nowego algorytmu minimalno-optymalnej selekcji cech. Zaś do osiągnięć empirycznych należy przetestowanie zaproponowanego algorytmu na rzeczywistych danych biomedycznych.

Zasadniczym osiągnięciem rozprawy jest wprowadzenie przez mgr. Piliszka nowego algorytmu selekcji cech RAFS (ang. *Robust Agglomerative Feature Selection*). RAFS jest odpornym na przeuczenie i stabilnym algorytmem minimalno-optymalnej selekcji cech, uwzględniającym informację istotną dla zmiennej decyzyjnej. W celu redukcji liczby cech wykorzystywanych w klasyfikacji, RAFS organizuje cechy w grupy, a następnie wyłania przedstawicieli poszczególnych grup. W celu uzyskania stabilnego zestawu cech, RAFS wykorzystuje zespół wyników uzyskanych w schemacie z walidacją krzyżową. Natomiast w celu pomiaru przyrostu informacji o zmiennej decyzyjnej, RAFS wykorzystuje miarę STIG, która jest kolejnym osiągnięciem teoretycznym rozprawy zaproponowanym przez Doktoranta. Miara STIG, czyli miara symetrycznego wzrostu informacji o celu (ang. *Symmetric Target Information Gain*), jest nową miarą odmienności cech uwzględniającą relację między cechami a zmienną decyzyjną. STIG mierzy przyrost informacji o zmiennej decyzyjnej zarówno gdy znamy wartości obu cech, jak i w przypadku gdy znamy wartość tylko jednej z nich. Miara STIG została zaczerpnięta z pracy [156][1], gdzie została ona wprowadzona pod inną nazwą. Nowatorski wkład mgr. Piliszka w przypadku miary STIG ma dwojaki charakter. Po pierwsze, Doktorant wykorzystał miarę zaczerpniętą z pracy [156] w zupełnie nowym celu, mianowicie do opracowania nowego algorytmu selekcji cech RAFS. Po drugie, Doktorant udowodnił w podsekcji 3.3.1, że miara STIG nie spełnia nierówności trójkąta, zatem nie jest metryką, co jest błędnie udowodnione w pracy [156] i co zostało wskazane w pracy [172].

Zasadniczym osiągnięciem empirycznym rozprawy jest przeanalizowanie eksperymentalne algorytmu RAFS na trzech zbiorach danych biomedycznych dotyczących pacjentów z różnymi typami i podtypami nowotworów, które to zbiory obejmują pełen zakres trudności klasyfikacji binarnej – średni sygnał (BLCA), silny sygnał (PTLD) oraz słaby sygnał (KIRC). RAFS został porównany z innymi minimalno-optymalnymi algorytmami selekcji cech (mRMR, RFE i wstępnie z JMIM). W eksperymentach algorytm RAFS był testowany w połączeniu z następującymi klasyfikatorami – z regresją liniową, naiwnym klasyfikatorem bayesowskim i lasem losowym, w którym wewnętrznym klasyfikatorem był RFE. Jakość klasyfikacji mierzono na zbiorze testowym za pomocą statystyki AUC, zaś stabilność selekcji cech mierzono dwoma miarami, mianowicie Jaccard Score oraz Consistency Score.

Przebieg algorytmu RAFS na zbiorze BLCA dawał w porównaniu z innymi metodami wyniki klasyfikacji zarówno lepsze, jak i bardziej stabilne w rygorystycznej zewnętrznej walidacji krzyżowej. Zastosowanie algorytmu RAFS wyposażonego w miarę STIG do zestawu danych PTLD ujawniło natomiast więcej ciekawych genów niż poprzednie badania. Pomimo słabego sygnału, algorytm RAFS uruchomiony na zestawie danych KIRC był w stanie osiągnąć najwyższą obserwowalną wydajność klasyfikacji, jak również stabilność wyboru cech w rygorystycznej zewnętrznej walidacji krzyżowej.

[1] Odnosząc się do prac cytowanych w recenzowanej rozprawie doktorskiej mgr. Radosława Piliszka zachowuję numerację z bibliografii tejże rozprawy.

Ponadto, mgr. Piliszek wykorzystał w swojej rozprawie doktorskiej wprowadzoną uprzednio przez siebie i współpracowników ([131]) bibliotekę do wyboru cech MDFS (ang. *MultiDimensional Feature Selection*), która posłużyła jako silnik obliczeniowy dla nowo zaproponowanych algorytmów, wzbogacając ją między innymi o możliwość liczenia różnych miar wywodzących się z teorii informacji.

3. Poprawność

Rozprawa jest starannie napisana i zredagowana. Mgr. Piliszek prezentuje przejrzyście zarówno zagadnienia przytaczane z literatury, jak i własne wyniki. Za przykład mogą posłużyć wprowadzenie do teorii informacji przedstawione w podrozdziale 2.5, jak i prezentacja miary STIG w podrozdziale 3.3. Eksperymenty są zaprojektowane i przeprowadzone poprawnie, zaś ich rozplanowanie jest właściwe względem testowanych hipotez badawczych. Wnioski płynące z badań empirycznych są poprawnie sformułowane. Doktorant ilustruje rozważania teoretyczne i empiryczne umiejętnie dobranymi rysunkami, wykresami i tabelami, co czyni recenzowaną rozprawę przejrzystą i ułatwia jej lekturę.

Rozprawa jest napisana w zwięzły sposób, co zasadniczo stanowi jej zaletę, może za wyjątkiem rozdziału 6. Rozdział ten prezentuje wyniki przeprowadzonych analiz empirycznych poszczególnych metod. Choć wnioski formułowane w tym rozdziale są poprawne, to brakuje w nim w szczególności podrozdziału zbiorczo omawiającego wyniki zaobserwowane w poszczególnych eksperymentach, w którym zostałyby ponownie przytoczone i zinterpretowane najważniejsze rezultaty poszczególnych eksperymentów. Taki podrozdział pozwoliłby zapewne na czytelniejsze porównane wyników, co stanowiłoby dobitniejsze uzasadnienie i ilustrację wniosków przedstawionych w – także trochę zbyt zwięzłym – rozdziale 7.

Z edycyjnego punktu widzenia rozprawę mgr. Piliszka można stawiać za wzór dla rozpraw doktorskich z informatyki. W szczególności, opisy rysunków i tabel pojawiających się w rozprawie mają stosowną formę, zawierają zwiększone marginesy, zmniejszoną czcionkę i interlinię, i przez to wyraźnie odróżniają się od tekstu głównego. Ponadto na rysunkach grupujących wiele wykresów każdy z wykresów jest stosownie oznaczony, co umożliwia bezpośrednie odnoszenie się w tekście głównym do poszczególnych wykresów. Wspominam o tych być może stosunkowo mało ważnych aspektach, gdyż mankamenty związane z tymi aspektami pojawiają się dość często w rozprawach doktorskich i magisterskich.

Na koniec chciałbym podkreślić że wskazane wyżej drobne uchybienia w prezentowaniu wyników (rozdział 6) w niczym nie podważają mojej zdecydowanie pozytywnej oceny recenzowanej rozprawy. Warto również zaznaczyć, że rozprawa mgr. Piliszka charakteryzuje się swoistą kompletności jeśli chodzi o rozprawy z zakresu uczenia maszynowego, gdyż poza rozważaniami teoretycznymi i opisem badań empirycznych, Doktorant prezentuje w rozdziałach 3 i 4 aspekty implementacyjne opracowanych przez siebie metod, co znacznie ułatwi ich ewentualne zastosowania przez innych badaczy i praktyków.

4. Wiedza kandydata

Treść rozprawy wskazuje, iż mgr. Piliszek ma szeroką wiedzę i jest bardzo dobrze przygotowany do prowadzenia badań naukowych w zakresie uczenia maszynowego i eksploracji danych. Analizy prowadzone w przedłożonej rozprawie doktorskiej świadczą o dociekliwości naukowej Doktoranta i znajomości bieżącej literatury przedmiotu. W szczególności, rozważając metodę STIG w zagadnieniu mierzenia ograniczonej do zmiennej decyzyjnej odmienności cech, mgr. Piliszek po pierwsze dotarł do

Ślęzak

pracy [172] wskazującej na błąd w dowodzie w oryginalnej pracy [156] (wprowadzającej metodę STIG pod inną nazwą), skutkujący tym że metoda STIG nie spełnia nierówności trójkąta. Następnie, ponieważ praca [172] wskazywała błąd bez jego formalnego udowodnienia, Doktorant przedstawił własny dowód, osiągając poziom dociekliwości metodologicznej i sprawności dowodowej właściwy pracom doktorskim.

Należy też podkreślić, że zawartość recenzowanej rozprawy świadczy o bardzo dobrym przygotowaniu mgr. Piliszka do prowadzenia prac badawczo-rozwojowych, a także prac związanych z praktycznymi zastosowaniami metod uczenia maszynowego. Przegląd stanu wiedzy o selekcji cech z wykorzystaniem teorii informacji świadczy o erudycji i znajomości literatury. Rozważania empiryczne w rozdziałach 6 i 7, jak i algorytmy proponowane w rozdziałach 3, 4, 5 – wraz z ich eksperymentalną analizą – potwierdzają dobre przygotowanie metodologiczne Doktoranta do prowadzenia badań w zakresie eksploracji danych.

5. 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)[2] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego?

[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*

Podsumowując uważam, że opiniowana rozprawa mgr. Radosława Piliszka spełnia wymagania stawiane przez ustawę o stopniach naukowych i tytule naukowym w odniesieniu do rozpraw doktorskich i może być dopuszczona do publicznej obrony.

Ślęzak

Podpis

[2] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf