dr hab. Jan Bazan, prof. UR
Uniwersytet Rzeszowski,
Wydział Matematyczno-Przyrodniczy
ul. Pigonia 1, 35-310 Rzeszów
bazan@ur.edu.pl

30 stycznia 2017 r.

Recenzja rozprawy doktorskiej

mgr. Mirona Bartosza Kursa

zatytułowanej:

Stabilne i wydajne metody selekcji cech z wykorzystaniem systemów uczących się

Promotor: *prof. dr hab. Marek Niezgódka*

1. Cel i zakres rozprawy

Praca dotyczy bardzo dobrze znanego z literatury problemu selekcji cech, który wciąż jest jednym z największych wyzwań w takich dziedzinach jak np. wnioskowania statystyczne, eksploracja danych, odkrywanie wiedzy z danych i wiele innych. Problem ten zwykle wiąże się z konstruowaniem metod selekcji cech, które są bardzo pomocne w wielu zastosowaniach np. do usuwaniu zbędnych lub szkodliwych cech oraz zmniejszanie fizycznego rozmiaru danych, co daje możliwość użycia bardziej złożonych obliczeniowo metod do ich analizy. Chodzi tutaj np. o usprawnienie działania algorytmów tworzenia klasyfikatorów, grupowania obiektów lub odkrywania zależności w danych np. w postaci reguł asocjacyjnych. Jednakże po wykorzystywaniu niewłaściwie dobranych metod selekcji cech często istnieje niebezpieczeństwo, że algorytmy analizy danych mogą odkryć fałszywą wiedzę dotyczącą danych, jak choćby błędne wzorce lub losowe zależności, które są silniejsze od faktycznie występujących. Ponadto, selekcja cech może usunąć przydatne informacje z danych lub wytłumić zdarzenia rzadkie i ekstremalne, ale istotne z punktu widzenia analizowanych zjawisk i procesów. Tymczasem mamy prawo oczekiwać, że po zastosowaniu metody redukcji cech, dane powinny być dalej reprezentatywne w kontekście badanych zjawisk, zachowując możliwie dużą część istotnej na początku informacji. Dlatego wszelkie badania mające na celu opracowanie nowych, efektywniejszych metod selekcji cech są bardzo ważne dla dalszego rozwoju systemów eksploracji danych.

Z ogólnego punktu widzenia, istniejące metody selekcji cech dzielą się na dwie klasy metod. Do pierwszej z tych klas należą metody, które poszukują możliwie małego podzbioru cech zapewniającego możliwie dobrą dokładność jakiejś metody modelowania (w tej recenzji będę je nazywał metodami selekcji **minimalnych-optymalnych** cech). Natomiast, do drugiej z tych klas należą metody, które poszukują podzbioru wszystkich cech, które niosą istotną informację i przez to są potencjalnie użyteczne dla dowolnej metody modelowania (będę je nazywał metodami selekcji **wszystkich-istotnych** cech).

Głównym celem rozprawy jest zatem wykazanie, że pomimo swoich zalet i popularności, metody selekcji minimalnych-optymalnych cech mogą być w wielu przypadkach danych mniej efektywne od metod klasy wszystkich-istotnych. Tyczy się to szczególnie sytuacji, gdy liczba cech do redukcji jest duża.

W związku z tym, w rozprawie zaproponowano kilka metod selekcji cech klasy wszystkich-istotnych oraz przebadano je na rzeczywistych zbiorach danych.

Opisana wyżej problematyka jest niebanalna. Dlatego uważam, że podjęta tematyka może z powodzeniem stanowić przedmiot rozprawy doktorskiej.

2. Zawartość rozprawy

Rozprawa napisana jest w nowym stylu, który polega na zaprezentowaniu ogólnych elementów rozprawy (streszczenie, wstęp i dalsze perspektywy), które zostały uzupełnione tekstem 4 publikacji opublikowanych w czasopismach z ministerialnej listy A, stanowiących główne rozdziały rozprawy. Warto podkreślić, że taki sposób przedkładania rozprawy doktorskiej jest dopuszczalny przez obowiązujące obecnie przepisy w tym zakresie.

Rozdział 1 to wprowadzenie obejmujące ogólne przedstawienie motywacji, problemu badawczego, celu i tezy rozprawy oraz wkładu autora w aktualny stan wiedzy. W rozdziale 2 zamieszczono opis pierwszej głównej algorytmicznej metody selekcji cech przedkładanej rozprawy, którą nazwano metodą Boruta. Natomiast w rozdziale 3, opisano metodę rFerns selekcji cech, która jest oryginalną modyfikacją znanej z literatury metody paproci losowych. Rozdział 4 jest poświęcony weryfikacji wprowadzonych w poprzednich dwóch rozdziałach metod do analizy danych pochodzących z mikromacierzy DNA. Natomiast w rozdziale 5 opisano zastosowanie prezentowanych w rozprawie metod do rozpoznawania brzmienia instrumentów w utworach muzycznych, przy czym dodano pewną modyfikację metody rFerns pozwalającą na wykorzystanie jej do konstrukcji klasyfikatora wieloetykietowego.

Warto dodać, że zarówno w rozdziałach 2, 3, 4 i 5 przedstawiono wyniki eksperymentów wykonanych na zbiorami danych.

Na koniec, w rozdziale 6 bardzo krótko podsumowano rozprawę, wspominając liczne cytowania prezentowanych podejść, przede wszystkim w zastosowaniach do analizy różnego rodzaju danych danych biologicznych.

3. Poprawność i oryginalność postawionej tezy (wkład autora)

Na podstawie swoich doświadczeń w zakresie metod selekcji cech, Autor sformułował tezę rozprawy, którą można przedstawić w następujący sposób.

Metody selekcji cech z klasy minimalnych-optymalnych, pomimo swoich zalet i popularności, w wielu przypadkach zbiorów danych są mniej efektywne od metod klasy wszystkich-istotnych, gdyż ograniczają wyjaśniającą rolę samej selekcji cech i często wprowadzają istotne ryzyko utraty odporności analizy na szum i fałszywe przypadkowe zależności. Tyczy się to szczególnie sytuacji, gdy liczba cech do redukcji jest duża.

Zdaniem Recenzenta, postawienie takiej tezy jest uzasadnione, choćby z powodu wątpliwości jakie pojawiają się u badaczy podczas analizy rzeczywistych danych. Na przykład, niektórzy lekarze klinicyści, współpracujący z informatykami podczas budowy systemów inteligentnych twierdzą, że

metody klasy minimalnych-optymalnych, które często przy tworzeniu modeli obliczeniowych odrzucają większość część cech, działają w sprzeczności z ich codzienną praktyką polegającą na wykorzystywaniu pewnej redundancji cech, pozwalającą, ich zdaniem, na podejmowanie pewniejszych decyzji medycznych. Oczywiście przesłanki, jakimi kierował się Autor rozprawy uzasadniając powyższą tezę niebezpieczeństwem odkrycia fałszywej wiedzy i usunięcia z danych wiedzy o zdarzeniach rzadkich i ekstremalnych, także są przekonywujące.

Tak więc, w związku z powyższymi argumentami uwaga badaczy w sposób naturalny kieruje się na metody klasy wszystkich-istotnych cech.

Dla wykazania sformułowanej wyżej tezy wykonano następujące prace, których wyniki są wkładem Autora rozprawy w rozwój dyscypliny naukowej.

1. Zaproponowano dwie metody (algorytmy) selekcji cech należących do klasy wszystkich-istotnych cech.
2. Zaimplementowano te metody we własnej bibliotece oprogramowania w środowisku języka R.
3. Wykonano eksperymenty na danych rzeczywistych wykorzystując własne implementacje oraz konkurencyjne implementacje innych metod, zinterpretowano wyniki eksperymentów oraz przedstawiono wnioski.

Poniżej omawiam bardziej szczegółowo wyniki wyżej wymienionych prac.

Ad 1. Zaproponowano dwa następujące algorytmy związane z selekcją cech metodami wszystkich-istotnych cech.

a) Pierwszy z algorytmów, zwany metodą Boruta, polega na dodawaniu dodatkowych, mało istotnych cech (tzw. cieni) do istniejącego zbioru cech w danych systemie informacyjnym. Następnie są one wykorzystywane jako odniesienie dla oceny istotności oryginalnych atrybutów w kontekście pełnej struktury analizowanych danych, przy czym atrybuty są oceniane w oparciu o miarę przydatności obliczoną na podstawie lasów losowych wyuczanych podczas treningu na danych. Na podstawie tej miary metoda Boruta prowadzi selekcję iteracyjnie, usuwając z systemu informacyjnego stopniowo cechy uznane za nieistotne, przy czym dzięki dodanym cieniom i dobranej mierze przydatności atrybutów, selekcja jest bardziej stabilna i niepodatna na przeuczenie czy też szumy w danych.

b) Drugi z algorytmów, zwany metodą rFerns jest oryginalna modyfikacją znanej z literatury metody paproci losowych. Został zaproponowany jako "głęboko stochastyczny" system uczący się wprowadzony jako wydajna obliczeniowo alternatywa dla metody Boruta, która nie radziła sobie na dużych danych. W swojej istocie, metoda rFerns polega na konstruowaniu k tzw. paproci losowych, które są rodzajem drzewa binarnego o ustalonej wysokości, tworzonego dla losowo wybranego zbioru atrybutów i obiektów, z wykorzystaniem losowego doboru kryterium podziału węzła, tj. losowego wyboru atrybutu do podziału oraz progu wartości tego atrybutu. Liście takiej paproci pozwalają na wyliczenie pewnej miary przynależności obiektu testowego do określonej klasy decyzyjnej, po czym miary te są agregowane, dzięki czemu tego rodzaju paproć losowa jest klasyfikatorem ogólnego przeznaczenia i jest zdolna do szacowania ważności cech.

Ad 2 Autor zaimplementował proponowane w rozprawie algorytmy we własnej bibliotece oprogramowania w środowisku języka R.

Ad 3. W celu oceny proponowanych rozwiązań pod względem ich efektywności w porównaniu do konkurencyjnych podejść, wykonano eksperymenty na zbiorach danych wykorzystując własne implementacje oraz konkurencyjne implementacje innych metod. Wyniki tych eksperymentów opisano w rozdziałach 2, 3, 4 i 5 rozprawy.

a) W rozdziale 2 opisano eksperymenty pokazujące przydatność metody Boruta do budowy efektywnych klasyfikatorów. Okazało się, że zastosowanie tej metody do selekcji cech powoduje, że jakość klasyfikacji wytworzonego klasyfikatora dla analizowanych danych znacząco wzrosła, przy czym stosowano klasyfikator tworzony metodą lasu losowego.

b) W rozdziale 3 opisano eksperymenty wykonane na różnych zbiorach danych, których celem było porównanie efektywności proponowanej w rozprawie metody rFerns z klasyczną metodą lasów losowych. Okazało się, że obydwie metody dały podobną jakość klasyfikacji. Jednak metoda rFerns działała szybciej do metody lasu losowego, co sprawia, że metodę rFerns można traktować jako lepszą propozycję, gdyż jest ona mniej wymagająca obliczeniowo.

c) W rozdziale 4 analizowano znane z literatury zbiory danych pochodzących z mikromacierzy DNA , czyli dane w których występuje bardzo dużo cech przy stosunkowo niewielkiej liczbie obiektów. Przy tej okazji dokonana została ocena metody Boruta, metody rFerns i innych metod w aspekcie stabilności selekcji cech, przy czym do tego celu użyto autorską metodę oceny opartą o tzw. samozgodność wyników szacowania przy pomocy metody bootstrap. Wyniki tych analiz empirycznie potwierdzają że przyjęcie celu selekcji tak jak w metodach minimalnych-optymalnych może doprowadzić do znaczącej utraty stabilności selekcji. Dowodzą także, że popularna w podobnych pracach ocena jakości selekcji przez błąd predykcji modelu wytrenowanego tylko na wyselekcjonowanych atrybutach może prowadzić do fałszywych wniosków. Uzyskane wyniki potwierdzają zatem słuszność stosowania metod klasy wszystkich-istotnych.

d) Wreszcie w rozdziale 5 analizowano zastosowanie prezentowanych w rozprawie metod do rozpoznawania brzmienia instrumentów w utworach muzycznych, przy czym głównym celem było porównanie dla tych danych proponowanej w rozprawie metody rFerns z klasyczną metodą lasów losowych. Okazało się, że także i dla tych danych obydwie metody dają podobną jakość klasyfikacji, a metoda rFerns była mniej wymagająca czasowo od metody lasu losowego. Dodatkowo, zaproponowano pewną modyfikację metodę rFerns, a dokładnie modyfikację sposobu wyliczenia miary przynależności obiektu testowego do określonej klasy decyzyjnej w oparciu o liście paproci. Modyfikacja ta spowodowała, że każda paproć uzyskana w tej metodzie stała się klasyfikatorem wieloetykietowym, co jeszcze przyspieszyło działanie metody.

4. Wiedza i umiejętności Autora do poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników

W rozprawie Autor zamieścił przegląd aktualnego stanu wiedzy w zakresie klasyfikacji metod selekcji cech. Opisy te zostały wykonane z odwołaniami do literatury, co pozwoliło umiejscowić w literaturze przedmiotu prezentowane badania. Bez wątpienia świadczą one o dużej wiedzy kandydata. Ponadto, należy tutaj mocno podkreślić, że napisanie rozprawy wymagało wcześniejszego skonstruowania, w tym zaprogramowania, środowiska eksperymentalnego.

Jak już wspominałem wyżej, pozycja rozprawy w stosunku do obecnego stanu wiedzy polega na tym, że rozprawa polemizuje z aktualnymi stereotypami dotyczącymi jakości cech wyznaczonych metodami z klasy minimalnych-optymalnych, wskazując na wady tych metod i proponuje zamiast nich użycie metod klasy wszystkich-istotnych cech, które pozwalają w wielu przypadkach na uniknięcie tych wad. Zaproponowane nowe metody z pewnością będą użyteczne w selekcji cech dla rzeczywistych danych, co potwierdzają liczne cytowania literaturowe prezentowanych w rozprawie metod.

Odnosząc się do umiejętności Autora rozprawy w zakresie poprawnego i przekonywującego przedstawiania uzyskanych wyników należy stwierdzić, że w tym zakresie Autor rozprawy wykazał się w wystarczającym stopniu. Wprawdzie wiele zagadnień zostało opisane dość lakonicznie, ale po dokładnym wczytaniu się w tekst rozprawy można je zrozumieć.

4. Uwagi na temat rozprawy

Z formalnego i matematycznego punktu widzenia rozprawa nie budzi zastrzeżeń recenzenta.

Podczas lektury rozprawy nasuwają się jednak pewne drobne uwagi. Dobrze by było, aby Kandydat odniósł się do nich podczas obrony rozprawy.

1) Rozprawa praktycznie nie zawiera żadnych informacji o zastosowaniach prezentowanych metod przez innych badaczy oraz konstrukcji konkretnych systemów komputerowych, wykorzystujących proponowane metody w praktyce. Tymczasem metody te mają potencjalnie liczne zastosowania, a publikacje będące podstawą rozprawy były już wielokrotnie cytowane. Na przykład publikacja będąca podstawą drugiego rozdziału rozprawy była już cytowana w Google Scholar 171 razy, a w Web of Science 83 razy. Natomiast publikacja będąca podstawą czwartego rozdziału rozprawy była cytowana w Google Scholar 27 razy, a w Web of Science 13 razy. Brak w rozprawie informacji o tych zastosowaniach wynika zapewne z faktu, że publikacje były przecież cytowane po ich wydaniu, a może nawet po przygotowaniu rozprawy. Tak czy inaczej, zdaniem Recenzenta, zastosowania rozprawy związane z tymi cytowaniami powinny być chociaż skrótowo przedstawione podczas obrony.

2) Proponowane w rozprawie metody praktycznie nie wykorzystują wiedzy dziedzinowej dostarczonej przez ekspertów. Tymczasem, w wielu współczesnych zastosowaniach związanych z analizą danych taka możliwość pojawia się, co często powoduje istotny wzrost jakości i efektywności tych systemów. Powstaje pytanie: czy w prezentowanych podejściach jest sens użyć wiedzy dziedzinowej, a jeśli tak, to w jaki sposób mogłoby to być robione oraz w jakimś stopniu mogłoby to być pomocne w zwiększeniu efektywności proponowanych metod? Choć zagadnienie to nie jest bezpośrednio związane z treścią rozprawy, to warto porozmawiać o nim podczas obrony.

3) W rozprawie przedstawiono dyskusję na temat złożoności obliczeniowej czasowej prezentowanych metod algorytmicznych z której wynika, że duża złożoność czasowa algorytmu Boruta spowodowała potrzebę szukania jego mocno stochastycznej wersji nazwanej rFerns. Oznacza to, że algorytm Boruta nie jest skalowalny i może nie działać efektywnie dla większych danych. Nasuwa się pytanie czy jest możliwe opracowanie rozproszonej wersji tego algorytmu, w celu uczynienia go skalowalnym (np. za pomocą

klastra obliczeniowego)? To zagadnienie także powinno być wyjaśnione podczas obrony rozprawy.

5. Podsumowanie

Uzyskane wyniki są interesujące zarówno z teoretycznego, jak i praktycznego punktu widzenia. Dlatego niezależnie od wspomnianych wyżej drobnych mankamentów, uważam pracę za wartościową. Autor wykazał dobre opanowanie wielu różnorodnych technik matematycznych i informatycznych. Sposób wykorzystania tych technik wskazuje na opanowanie przez Niego warsztatu naukowego.

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) [1] moja ocena rozprawy pod względem trzech podstawowych kryteriów 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?

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

C. Czy kandydat ma umiejętność samodzielnego prowadzenia pracy naukowej?

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

W związku z powyższym, wnioskuję o dopuszczenie rozprawy doktorskiej do publicznej obrony.

Jan E. Baran

Podpis

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