Warszawa, 29 maja, 2015

dr hab. Dominik Ślęzak
Instytut Matematyki Uniwersytetu Warszawskiego
ul. Banacha 2, 02-097 Warszawa
slezak@mimuw.edu.pl

RECENZJA ROZPRAWY DOKTORSKIEJ

mgr Adama Skowrona

pt. „Sequential covering regression rule induction and optimization of regression rule-based data models"

Wstęp

Badania opisane w rozprawie obejmują problematykę automatycznego uczenia się modeli decyzyjnych na podstawie dostępnych danych treningowych, w tym metody wyznaczania reguł o przesłankach będących koniunkcjami warunków na dostępnych wartościach atrybutów oraz konkluzjach przyjmujących postać liczbową, związaną z charakterystyką wyszczególnionego numerycznego atrybutu decyzyjnego.

W literaturze, jak również w komercyjnych systemach eksploracji danych, dużą popularnością cieszą się tak zwane pokryciowe algorytmy indukcji reguł, które, zazwyczaj w sposób iteracyjny, starają się znaleźć zespół reguł, takich że (prawie) każdy obiekt treningowy spełnia przesłankę (przynajmniej) jednej z reguł. Powstało do tej pory bardzo wiele opracowań i algorytmów iteracyjnej indukcji reguł, koncentrujących się na szeregu kryteriów optymalizacyjnych łączących w sobie prostotę i skuteczność modeli regułowych.

W przeważającej większości, powyższe metody zostały jednakże rozwinięte tylko na potrzeby problemu klasyfikacji, gdzie wartości atrybutu decyzyjnego przyjmują postać kategorycznych klas decyzyjnych. Z jednej strony jest to o tyle zrozumiałe, że budowa klasyfikatorów, w tym klasyfikatorów regułowych, jest powszechnie uważana za jedno z koronnych zastosowań systemów uczących się. Z drugiej strony, biorąc pod uwagę praktyczny potencjał modeli regułowych, warto rozważyć zaadaptowanie ich na potrzeby zadań regresji, gdzie atrybuty decyzyjne przyjmują wartości numeryczne i gdzie jakość znajdowanych w danych reguł określa się poprzez agregacyjną analizę tych wartości podczas konstrukcji ich przesłanek.

Niniejsza rozprawa doktorska stara się wypełnić tę oczywistą lukę. Autor słusznie zauważa, iż szerokie zastosowanie algorytmów pokryciowych do rozwiązywania problemów klasyfikacji w nikłym stopniu przełożyło się jak dotąd na ich analogiczne zastosowania w rozwiązywaniu problemów regresyjnych. Dotyczy to w szczególności wstępujących i zstępujących strategii pokryciowej indukcji reguł, zarówno pod kątem optymalizacji i wykorzystania pojedynczych reguł, jak i całych ich zbiorów. Dostosowanie tych strategii do realiów konstrukcji modeli regresyjnych można uznawać za główny cel rozprawy.

Biorąc pod uwagę aktualny stan badań w omawianym zakresie, uważam, iż powyższy cel i otrzymane wyniki są pod względem zarówno naukowym, jak i praktycznym w pełni godne rozprawy doktorskiej.

Zawartość rozprawy

Rozprawa składa się z siedmiu rozdziałów i wykazu literatury. Jest ona napisana w języku angielskim.

W rozdziale pierwszym, autor przedstawia główne założenia i kluczowe aspekty nowatorskie rozprawy.

W rozdziale drugim, autor omawia podstawy modeli regułowych, podejścia znane z literatury, metody oceny reguł, jak również techniki wykorzystania wyuczonych reguł, wymagające głosowania, tudzież rozwiązywania konfliktów pomiędzy regułami. Jedna z tych technik ma istotne walory nowatorskie, zaś stojące za nią intuicje znajdują pozytywne odzwierciedlenie w dalszych wynikach eksperymentalnych.

W rozdziale trzecim, autor przedstawia uprzednio już wspomniane strategie wstępujące oraz zstępujące, wprowadza nowatorskie podejście do ustalania numerycznych konkluzji konstruowanych reguł, a także omawia wzorowaną na statystyce metodologię niezbędną dla oceny jakości reguł podczas ich budowy, parametryzowalną ze względu na bardziej optymistyczny bądź pesymistyczny stopień ufności reguł.

W rozdziale czwartym, autor opisuje szereg heurystycznych algorytmów optymalizujących reguły w trakcie oraz już po ukończeniu procesu ich konstrukcji. Tak jak w przypadku technik omawianych w rozdziale drugim, także tutaj autor zwraca uwagę na różnice pomiędzy zadaniami klasyfikacji i regresji. Ponadto, także tutaj warto zaznaczyć, iż dwa z podanych algorytmów filtracji reguł to nowe propozycje.

Kolejne dwa rozdziały przedstawiają wyniki eksperymentalne, otrzymane na ogólnie znanych danych porównawczych, jak i dla dwóch rzeczywistych zbiorów danych związanych z problematyką górnictwa. Wyniki są dogłębnie analizowane, co prowadzi do szeregu niezwykle cennych wniosków praktycznych. Oczywiście ważne jest tu również porównanie nowych algorytmów do istniejących wcześniej metod.

W rozdziale siódmym, autor podsumowuje dotychczasowe badania, a także wyznacza sobie pewne nowe kierunki na przyszłość. Rozdział ten pokazuje, iż – w moim odczuciu – cele rozprawy zostały osiągnięte.

Uwagi ogólne

Pewne drobne uwagi krytyczne przedstawiam w dalszych częściach recenzji. Natomiast w tym miejscu chciałbym przede wszystkim podkreślić elementy nowatorskie i znaczenie otrzymanych wyników. Poza głównym celem strategicznym, jakim było wypełnienie luki związanej z dość małym dotychczasowym zainteresowaniem regułowymi modelami regresyjnymi, autor zawarł w pracy kilka ważnych pomysłów dotyczących ogólnego podejścia do adaptowania stosowanych w zadaniach klasyfikacji miar jakości reguł, filtracji reguł, a także metod głosowania pomiędzy regułami, na potrzeby problematyki regresji.

Jednakże szczególną uwagę chciałbym zwrócić na metodę „fixed strategy", która pozwala w bardzo naturalny oraz skuteczny sposób predefiniować postać reguł regresyjnych indukowanych z danych i niejako formułować punkt odniesienia dla analizy ich jakości podczas optymalizacji struktury reguł. Moim zdaniem, właśnie tego typu pomysły – z jednej strony proste i intuicyjne, zaś z drugiej strony podparte odpowiednimi podstawami matematycznymi i umiejętnie wkomponowane w algorytmy – stanowią o postępie danej dziedziny. Mam zatem nadzieję, że autor będzie ten pomysł dalej rozwijał.

Uwagi redakcyjne

Układ pracy jest czytelny. W niektórych miejscach początkowych rozdziałów, zaproponowane pomysły mogłyby być przedstawione w sposób może nieco lepiej uporządkowany, pozwalający lepiej zrozumieć, które elementy są szczególnie nowatorskie. Przykładowo, rozbudowałbym sekcję 3.4, jako jeden z bardzo istotnych elementów proponowanego podejścia. Nie zmienia to jednak faktu, że pracę czyta się dobrze.

Mocną stroną rozprawy jest umiejętny sposób odwoływania się do literatury. Autor nakreśla dzięki temu obraz dziedziny i umiejscawia w niej swoje dokonania. Na tak zarysowanym tle dziedziny, można pełniej zrozumieć i docenić wagę zaproponowanych algorytmów i otrzymanych wyników eksperymentalnych.

Osobną kwestią jest język rozprawy. Tu muszę przyznać, że – dość nieoczekiwanie – napotkałem na sporo błędów gramatycznych, przekręconych słów oraz niespójności stylistycznych. Nie jest to jakaś bardzo duża liczba, ale mimo wszystko zauważalna. Nie zmienia to mojej dobrej opinii o układzie i redakcyjnej poprawności rozprawy, ale jeśli – w co wierzę – autor będzie zamierzał publikować dalsze rozszerzenia otrzymanych wyników, to na kwestie językowe należałoby zwrócić baczniejszą uwagę.

Chciałbym dorzucić jeszcze jedną uwagę odnośnie nazwy „fixed strategy". Z jednej strony, słowo „fixed" dobrze charakteryzuje główną ideę opisywanej metody, która – jak już wspominałem – wywarła na mnie bardzo pozytywne wrażenie. Jednak obawiam się, że czytelnik, szczególnie świeżo po zapoznaniu się ze strategiami „top-down" i „bottom-up", może słowo „fixed" kojarzyć podświadomie z czymś innym. Stąd proponuję, by autor na przyszłość zastanowił się jeszcze, czy nazwa „fixed strategy" jest tutaj najlepsza.

Uwagi krytyczne

W zasadzie nie mam dalszych istotnych uwag krytycznych dotyczących przedstawionego materiału.

Pragnę tylko nadmienić, że można byłoby rozważyć głębszą dyskusję dotyczącą złożoności problemów optymalizacyjnych wiążących się z wyznaczaniem regresyjnych modeli regułowych. Dość interesujące mogłoby być też zweryfikowanie, na ile zaproponowane techniki mogłyby zostać przeniesione na grunt problematyki selekcji cech i na ile odpowiednio zmodyfikowane algorytmy selekcji cech mogłyby się przyczynić do dalszego usprawnienia przedstawionej metodologii. Z drugiej jednak strony, te akurat aspekty nie stanowiły celu niniejszej rozprawy, a więc trudno uznać moje powyższe uwagi za krytykę. Proszę je raczej traktować jako dodatkowe propozycje możliwych przyszłych kierunków badawczych.

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), uważam, że rozprawa spełnia wymagania stawiane przez powyższą ustawę w odniesieniu do rozpraw doktorskich i może być dopuszczona do publicznej obrony.

Ślęzak

dr hab. Dominik Ślęzak