prof. dr hab. inż. Marzena Kryszkiewicz
Politechnika Warszawska
Instytut Informatyki
ul. Nowowiejska 15/19
00-665 Warszawa
mkr@ii.pw.edu.pl

Warszawa, 31.05.2022

Recenzja rozprawy doktorskiej

mgr. Grzegorza Góry

zatytułowanej

Combining instance-based learning and rule-based methods for imbalanced data

1. Zagadnienie naukowe rozpatrzone w pracy

Opiniowana rozprawa jest poświęcona zagadnieniu klasyfikacji wykorzystującej połączenie metody opartej na instancjach z metodą opartą na regułach z uwzględnieniem specyfiki niezbalansowanych danych. Jako, że zbiory danych są często niezbalansowane, niejednokrotnie silnie, z czym można zetknąć się np. w danych medycznych, postawiony w rozprawie problem ma istotne znaczenie praktyczne. Jakkolwiek istnieją prace dotyczące metod klasyfikacji niezbalansowanych danych oraz klasyfikacji mieszanej, łączącej podejście oparte na instancjach z regułowym, istnieje potrzeba ich dalszego ich rozwoju, który prowadziłby do wypracowania jeszcze skuteczniejszych metod klasyfikacji niezbalansowanych danych, które byłyby wydajne i wygodne w użyciu.

Celem rozprawy było opracowanie wysokiej jakości metod klasyfikacji, stanowiących połączenie metod klasyfikacji z użyciem najbliższych sąsiadów z tymi opartymi na regułach dla niezbalansowanych zbiorów danych o atrybutach numerycznych i nominalnych. Uzyskane rezultaty obejmują algorytmy RIONA i jego uproszczenie ONN (przedstawione w rozdziale 3) oraz RIONIDA i jego uproszczenie ONIDA (przedstawione w rozdziale 4), a także ich analizę teoretyczną (przedstawioną w rozdziałach 3-4 oraz dodatkach) i eksperymentalną (przedstawioną w rozdziale 5). Algorytmy RIONA i ONN są autorstwa mgr. Grzegorza Góry oraz dr. Arkadiusza Wojny. Algorytmy te zostały przewidziane do klasyfikacji w przypadku zbalansowanych zbiorów danych. Algorytmy RIONIDA i ONIDA zostały samodzielnie opracowane i przeanalizowane przez mgr. Grzegorza Górę. Algorytmy te zostały zaprojektowane pod kątem danych niezbalansowanych do binarnej klasyfikacji.

Rozprawa ma charakter teoretyczno-doświadczalny.

2. Wiedza Autora

Przeprowadzona w rozprawie w rozdziale 2 analiza literatury światowej oraz w rozdziale 5 metoda planowania eksperymentów badawczych potwierdza głęboką wiedzę Autora w zakresie uczenia maszynowego, ze szczególnym naciskiem na metody klasyfikacji oparte na instancjach oraz regułowe metody klasyfikacji, a także zagadnienia związane z oceną jakości klasyfikacji i specyfiką danych zbalansowanych i niezbalansowanych. Bibliografia liczy 240 prac.

3. Wkład Autora

Samodzielnym i oryginalnym dorobkiem Autora rozprawy jest opracowanie algorytmu RIONIDA do leniwej, binarnej klasyfikacji danych niezbalansowanych. Postawiony problem stanowi znacząco większe wyzwanie niż opracowanie skutecznego algorytmu klasyfikacji dla przypadku zbalansowanych klas decyzyjnych. RIONIDA stanowi nietrywialną adaptację zaproponowanego we wcześniejszych publikacjach Autora rozprawy i dr. Arkadiusza Wojny algorytmu RIONA, który był opracowany pod kątem klasyfikacji dla przypadku co najmniej 2 klas decyzyjnych, ale zbalansowanych.

W przypadku obu tych algorytmów klasyfikacja polega na wyznaczeniu k+ najbliższych sąsiadów klasyfikowanego obiektu (tzn. k najbliższych sąsiadów i tych, których odległość od klasyfikowanego obiektu jest taka sama jak odległość k-tego najbliższego sąsiada), a następnie na ich weryfikacji przy użyciu tylu reguł decyzyjnych, ilu jest sąsiadów, i ostatecznie na podjęciu decyzji o przydziale do klasy decyzyjnej na podstawie tych sąsiadów, którzy zostali zweryfikowani pozytywnie. Każda z tworzonych reguł jest budowana na podstawie wartości atrybutów warunkowych klasyfikowanego obiektu i jednego z k+ najbliższych sąsiadów. Część warunkowa utworzonej reguły pokrywa zarówno obiekt klasyfikowany jak i sąsiada użytego do jej wytworzenia. Część decyzyjna reguły odpowiada wartości decyzyjnej użytego sąsiada. Jeśli część warunkowa reguły pokrywa także jeszcze innego sąsiada o innej wartości decyzyjnej niż wartość decyzyjna reguły, to reguła jest uznana za niespójną, a sąsiad, który był użyty do budowy tej reguły jest zweryfikowany negatywnie i nie jest później wykorzystywany do podjęcia decyzji o przydziale klasyfikowanego obiektu do klasy decyzyjnej. Niech R oznacza zbiór k+ najbliższych sąsiadów klasyfikowanego obiektu, którzy zostali zweryfikowani pozytywnie. Podejmowanie decyzji przez RIONA na podstawie zbioru sąsiadów R jest przeprowadzane w sposób standardowy dla metody kNN, przydzielając klasyfikowany obiekt do klasy decyzyjnej najsilniej reprezentowanej w R. W przypadku algorytmu RIONIDA klasyfikowany obiekt jest przydzielany do klasy mniejszościowej, jeśli względna liczba zweryfikowanych pozytywnie sąsiadów z klasy mniejszościowej wynosi co najmniej p, gdzie p jest parametrem.

Podstawowym parametrem algorytmu klasyfikacji w metodzie RIONA jest parametr – k, wykorzystywany do wyszukiwania k+ najbliższych sąsiadów klasyfikowanego obiektu. Algorytm klasyfikacji w metodzie RIONIDA oprócz k ma jeszcze 2 inne podstawowe parametry: wspomniany wyżej parametr p – wykorzystywany do podejmowania decyzji o przypisaniu klasyfikowanego obiektu do klasy decyzyjnej oraz parametr skalowania s – wykorzystywany do ograniczania zakresu części warunkowych reguł w porównaniu z warunkami tworzonymi przez algorytm RIONA.

Wprowadzenie możliwości skalowania zakresu części warunkowych reguł uważam za istotne i bardzo ciekawe osiągnięcie Autora rozprawy. Warto odnotować, że im mniejszy zakres

części warunkowej reguły, tym mniejsze pokrycie części warunkowej reguły, a tym samym mniejsze prawdopodobieństwo, że pokryci sąsiedzi będą należeć, do różnych klas decyzyjnych, czyli mniejsze prawdopodobieństwo, że reguła zostanie uznana za niespójną, a sąsiad, który został użyty do jej budowy reguły, zostanie zweryfikowany negatywnie i nie będzie brany pod uwagę przy podejmowaniu decyzji o przydziale klasyfikowanego obiektu do klasy decyzyjnej. Dla s = 1, zakres części warunkowych reguł jest maksymalny i jest tworzony jak w przypadku algorytmu RIONA. Zatem dla s = 1 najwięcej sąsiadów zostanie zweryfikowanych negatywnie, w związku z czym decyzja o przydziale klasyfikowanego obiektu będzie podejmowana z użyciem najmniejszej liczby sąsiadów zweryfikowanych pozytywnie. Dla $s \in [0, 1)$ zakres części warunkowych reguł jest tym mniejszy i mniej zależny od wartości atrybutów warunkowych sąsiadów używanych do ich budowy, im mniejsza jest wartość s. Jeśli s = 0, zakres części warunkowych reguł jest budowany wyłącznie na podstawie wartości atrybutów warunkowych klasyfikowanego obiektu. W przypadku gdy s < 0, dla każdego atrybutu warunkowego *a* budowany jest warunek $a \in \emptyset$. W rezultacie, jeśli s < 0, części warunkowe tworzonych reguł nie pokrywają żadnego k+ najbliższego sąsiada i, w konsekwencji, wszyscy ci sąsiedzi są uwzględniani przy podejmowaniu decyzji o przydziale klasyfikowanego obiektu do klasy decyzyjnej. Parametr s zatem wpływa w nietrywialny sposób, przy zastosowaniu podejścia regułowego, na wyznaczenie tych spośród k+ najbliższych sąsiadów klasyfikowanego obiektu, na podstawie których zostanie przydzielony do klasy decyzyjnej.

Bardzo ważnym osiągnięciem Autora rozprawy jest samodzielne opracowanie algorytmu przyrostowego wyznaczania optymalnych ze względu na wybrane miary jakości klasyfikacji (G-mean, miara F oraz dokładność) wartości parametrów k, p i s w metodzie RIONIDA. Stanowi on istotne rozszerzenie opracowanego wcześniej przyrostowego algorytmu wyznaczania optymalnej wartości parametru k dla algorytmu RIONA, który był prezentowany we wspólnych publikacjach Autora rozprawy i dr. Arkadiusza Wojny. W porównaniu z tymi publikacjami w rozprawie wniesiono poprawkę, polegającą na tym, że optymalna wartość k jest wyznaczana na podstawie k_{max}+ najbliższych sąsiadów, a nie k_{max} najbliższych sąsiadów. Opracowane algorytmy mogą być bezpośrednio wykorzystane lub zaadaptowane do wyznaczania optymalnej wartości parametrów metod kNN i k+NN dla danych zbalansowanych i niezbalansowanych, a także dla innych niż wymienione wyżej miary jakości klasyfikacji. Stanowią też inspirację, jak w ogólności tworzyć algorytmy przyrostowego wyznaczania optymalnych wartości parametrów klasyfikatorów dla zadanych miar ich oceny.

Bardzo cennym, samodzielnym osiągnięciem Autora rozprawy jest wyprowadzenie szeregu wyników teoretycznych dotyczących algorytmów RIONA i RIONIDA, w tym parametrów algorytmów.

Należy nadmienić, że Autor rozprawy zaproponował także algorytmy ONN i ONIDA jako szczególne warianty algorytmów RIONA i RIONIDA. ONN i ONIDA reprezentują uczenie maszynowe oparte na instancjach, bez łączenia go z podejściem regułowym.

W celu oceny zaproponowanych algorytmów, Autor przeprowadził ich gruntowną i wszechstronną ocenę eksperymentalną. Eksperymenty zostały bardzo dobrze zaplanowane. W rozprawie przedstawiono m.in. wyniki eksperymentów przeprowadzonych przy użyciu RIONA i RIONIDA z zastosowaniem pseudometryki City and Simplified Value Difference oraz 8 innych klasyfikatorów (BRACID, MODLEM-C, kNN, MODLEM, J48, PART, RIPPER, RISE) na 20 zbiorach danych. RIONIDA okazała się zdecydowanie najlepszym algorytmem klasyfikacji ze względu na miarę G-mean oraz najlepszym ze względu na miarę F. Badano wpływ stosowania filtrów SMOTE i ENN, wykorzystywanych w przypadku

danych niezbalansowanych. Dla badanych zbiorów danych, wartość miary G-mean dla algorytmu RIONIDA była na ogół większa bez stosowania tych filtrów w przeciwieństwie do algorytmu RIONA, którego jakość klasyfikacji rosła, gdy stosowano filtry. Niemniej, wartość miary G-mean dla algorytmu RIONIDA była na ogół większa niż dla algorytmu RIONA, czy to zastosowanego z filtrami, czy bez. Eksperymentalnie wykazano także, że algorytm RIONIDA osiąga na ogół wyższe wartości miary F i G-mean niż algorytm BRACID, który był zaprojektowany pod kątem danych niezbalansowanych. Badano także wpływ na jakość klasyfikacji z użyciem RIONIDA różnych metod głosowania, ważenia atrybutów, wyznaczania odległości, innych wartości skalowania dla klasy mniejszościowej i większościowej.

4. Poprawność

Rozprawa jest zredagowana bardzo starannie. Jest napisana w języku angielskim, którym Autor posługuje się biegle i precyzyjnie. Układ pracy jest dobry. Prezentacja znanych i proponowanych przez Autora rozwiązań oraz wywody na ich temat (w tym dowody twierdzeń, propozycji i konkluzji) są prowadzone w sposób logiczny, przekonujący i zrozumiały (jedynym wyjątkiem budzącym moje wątpliwości jest komentarz na str. 104 odnoszący się do linii 13 algorytmu 9, co podnoszę poniżej). Zaproponowane rozwiązania zostały dogłębnie i w sposób właściwy przeanalizowane na drodze analitycznej i eksperymentalnej.

Zauważone słabości/nieścisłości:

- Na str. 11 stwierdzono, że „In the past, there have been some attempts to combine instance- and rule-based approaches, however only for balanced data (see e.g. [53, 128]).” Faktycznie w publikacji [128] z roku 2004 zaprezentowano algorytm klasyfikacji DeEps, który łączy podejście regułowe z opartym na instancjach. W moim przekonaniu jednak, sposób wyznaczania klasy decyzyjnej w DeEps uwzględnia potencjalne niezbalansowanie klas decyzyjnych. Niemniej zaproponowany w [128] sposób wyznaczania klasy decyzyjnej faktycznie jest znacznie mniej zaawansowany niż zaproponowany przez Autora w algorytmie RIONIDA.
- Brakuje wyjaśnienia w jaki sposób wyznaczana jest odległość SVDM między wartością atrybutu nominalnego *a* klasyfikowanego obiektu a wartością atrybutu *a* obiektu ze zbioru trenującego, gdy ta pierwsza różni się od wszystkich wartości atrybutu *a* w zbiorze trenującym.
- W Algorytmie 9 (RIONIDA-classify) i Algorytmie 10 (ONIDA) $p_{current}$ jest wyznaczane jako $|supportSet(d_{min})| / |neighbourSet|$. Natomiast z komentarza na stronie 104 wynika, że Autor rozprawy przyjął, że w przypadku RIONIDA-classify $|neighbourSet| = |supportSet(d_{min})| + |supportSet(d_{maj})|$. Ta równość jest prawdziwa dla algorytmu ONIDA, ale w przypadku algorytmu RIONIDA-classify nie ma tej gwarancji. Dla RIONIDA-classify zachodzi nierówność: $|neighbourSet| \geq |supportSet(d_{min})| + |supportSet(d_{maj})|$.
- W rozprawie nie przedstawiono pseudokodu algorytmu ONN. Co prawda na str. 84 jest przedstawiony jego krótki tekstowy opis, ale nie jest wystarczająco precyzyjny, żeby mieć pewność, co Autor rozprawy rozumie pod pojęciem ONN. Czy ONN to wariant algorytmu RIONA, z którego usunięto sprawdzanie spójności reguł zarówno na etapie wyznaczania optymalnej wartości k jak i na etapie klasyfikacji obiektu testowego? Czy też może ONN ma być odpowiednikiem tylko RIONA-classify?
- Analogiczne wątpliwości dotyczą relacji algorytmów RIONIDA i ONIDA. Z zamieszczonych pseudokodów wynika, że ONIDA jest wariantem algorytmu RIONIDA-classify, z którego usunięto sprawdzanie spójności reguł i przyjęto, że wartość

progowa p wynosi $|Class(d_{min})| / |trnSet|$. Czy taki wariant algorytmu RIONIDA-classify nie powinien nosić nazwy ONIDA-classify, a ONIDA być zmodyfikowanym wariantem RIONIDA?

Powyższe uchybienia/wątpliwości nie wpływają na moją bardzo wysoką ocenę osiągnięć naukowych Autora, które przedstawił w rozprawie.

5. Przydatność rozprawy

Zaproponowana przez Autora metoda leniwej klasyfikacji w postaci algorytmu RIONIDA jest skuteczna, elegancka, łatwo zrozumiała i generująca zrozumiałe dla człowieka uzasadnienia dla podejmowanych decyzji, co w wielu zastosowaniach, np. medycznych, bankowych, czy sądowych, ma ogromne znaczenie. Jest odporna na niezbalansowane dane, mimo, że nie stosuje filtrów. Ma wbudowany, przyrostowy mechanizm doboru optymalnych ze względu na stosowaną miarę jakości klasyfikacji (dokładność, miara F, G-mean) wartości parametrów: liczby sąsiadów k, progu decyzyjnego p i skalowania s zakresu warunków w regułach. W zależności od danych i wyuczonej automatycznie wartości optymalnej parametru skalowania s, RIONIDA może przyjąć postać klasyfikatora bazującego wyłącznie na k+ najbliższych sąsiadach w przypadku, gdy $s < 0$, albo na zredukowanym mniej lub bardziej w zależności od wartości $s \in [0, 1]$ podzbiorze k+ najbliższych sąsiadów, uzyskanym w wyniku ich weryfikacji przy użyciu lokalnie tworzonych reguł. RIONIDA jest niewątpliwie metodą, stanowiącą cenną inspirację do dalszych prac w zakresie uczenia maszynowego.

6. Podsumowanie

Opiniowana rozprawa zawiera oryginalne rozwiązanie postawionego problemu naukowego. Jej Autor wykazał się posiadaniem ogólnej wiedzy teoretycznej w dyscyplinie Informatyka, a także umiejętnością twórczego, samodzielnego prowadzenia pracy naukowej. Na drodze teoretycznej Autor wyprowadził szereg własności zaproponowanych rozwiązań. Ich skuteczność wykazał na podstawie szeroko przeprowadzonych eksperymentów obliczeniowych. Uzyskane przez Autora wyniki, które przedstawił w rozprawie, uważam za oryginalne, wartościowe i ciekawe poznawczo.

Stwierdzam, że recenzowana rozprawa doktorska mgr. Grzegorza Góry spełnia z nadmiarem wymagania stawiane rozprawom doktorskim przez obowiązującą ustawę i wnoszę o dopuszczenie jej do publicznej obrony. Ze względu na oryginalny wkład zaproponowanego w rozprawie rozwiązania, jakość uzyskanych wyników, spodziewany duży wpływ na rozwój uczenia maszynowego oraz możliwość szerokiego zastosowania w praktyce, wnoszę również o wyróżnienie opiniowanej rozprawy doktorskiej.