dr hab. inż. Robert Burduk, Prof. PWr
Politechnika Wrocławska
Wydział Elektroniki
Ul. Wybrzeże Stanisława Wyspiańskiego 27
50-370 Wrocław

Wrocław, dnia 22.05.2018 r.

RECENZJA

rozprawy doktorskiej mgra inż. Tomasza Orczyka
na temat: **„Klasyfikacja danych niekompletnych w oparciu o komitet klasyfikatorów"**

Problem badawczy i jego znaczenie

Zakres rozprawy dotyczy szerokiej i dynamicznie rozwijającej się dziedziny uczenia maszynowego, do której zaliczany jest problem klasyfikacji nadzorowanej. Zagadnienie klasyfikacji (nadzorowanej lub nienadzorowanej) stanowi przedmiot badań naukowych mieszczący się w dyscyplinie informatyka. W recenzowanej rozprawie Doktorant koncentruje się na zagadnieniach dotyczących odporności metod klasyfikacji na zakłócenia, które wynikają z występowania tzw. danych brakujących zarówno w zbiorze uczącym jak i testowym.

W rozprawie sformułowano tezę badawczą, która zakłada utrzymanie dokładności klasyfikacji w przypadku występowania brakujących danych w zbiorze uczącym i/lub testowym. Zdefiniowany problem badawczy został uszczegółowiony za pomocą postawionych celów badań naukowych, z których za najważniejsze należy uznać: oszacowanie wpływu brakujących danych na jakość klasyfikacji, dobór klasyfikatorów bazowych do komitetu klasyfikatorów, selekcję cech związaną z poszczególnymi etykietami klas oraz eksperymentalną weryfikację opracowanych metod klasyfikacji na rzeczywistych danych medycznych uzyskanych dzięki współpracy ze Śląskim Uniwersytetem Medycznym w Katowicach. Dodatkowo w badaniach eksperymentalnych wykorzystano kilka zbiorów danych pochodzących z publicznie dostępnego repozytorium UCI.

Tematyka rozprawy jest interesująca, w pełni uzasadniona i aktualna z punktu widzenia komputerowych metod klasyfikacji wykorzystujących zespół klasyfikatorów jak i dane niepełne oraz niezrównoważone.

Struktura pracy oraz wiedza Autora

Recenzowana praca została napisana w języku polskim i liczy 113 stron maszynopisu. Składa się z siedmiu rozdziałów merytorycznych, wstępu, podsumowania, bibliografii, spisu rysunków oraz tabel a także wykazu symboli. Rozdziały o nr 2–6 opisują stan wiedzy dotyczący metod klasyfikacji, miar jakości klasyfikacji, metod klasyfikacji danych niepełnych oraz metod selekcji cech. Przedstawiona w nich treść wskazuje, że Autor rozprawy posiada wiedzę teoretyczną, która dotyczy omawianej w pracy problematyki i mieści się w nurcie badań związanych z uczeniem maszynowym.

Oryginalne rozwiązanie problemu badawczego przedstawiono w rozdziale 7, który został podzielony na podrozdziały. Każdy z podrozdziałów przedstawia oryginalne propozycje

Autora dysertacji dotyczące problematyki klasyfikacji nadzorowanej z niepełnymi danymi. Rozdział 8 zawiera opis oraz wyniki badań eksperymentalnych, które zostały wykonane na kilku bazach danych pochodzących z repozytorium UCI oraz na bazie danych dotyczącej oceny stopnia włóknienia wątroby u pacjentów z wirusowym zapaleniem wątroby typu C. W rozdziale tym znajduje się również dyskusja otrzymanych wyników. W podsumowaniu pracy zawarte są propozycje dalszych prac naukowych, które należy uznać za ciekawe pomysły łączenia metod zaproponowanych w dysertacji. Kryterium wskazującym, które metody łączyć jest liczba brakujących danych w zbiorze uczącym lub testowym.

Spis literatury liczy 74 pozycje. Cytowane prace dobrane są prawidłowo i odnoszą się do omawianych problemów. Drobną uwagą jest stosunkowo mała liczba pozycji literaturowych opublikowanych w ostatnich pięciu latach.

Wkład Autora — oryginalne osiągnięcia

Wkład Autora w rozwój metod uczenia maszynowego, dedykowanych dla problemów klasyfikacyjnych, w których występują braki danych oraz problem niezrównoważenia klas polega na:

1. opracowaniu algorytmu usuwania danych niepełnych ze zbioru uczącego przy wykorzystaniu informacji o brakach danych w zbiorze testowym,
2. opracowaniu algorytmu usuwania danych niepełnych ze zbioru uczącego w celu stworzenia homogenicznego komitetu klasyfikatorów bazowych wykorzystujących pojedyncze cechy,
3. opracowaniu metody wyznaczania współczynników wagowych, mających na celu minimalizację wpływu niezrównoważenia klas po procesie usuwania danych niepełnych,
4. oszacowaniu liczby traconych wektorów referencyjnych oraz liczby zdegradowanych wektorów uczących w zbiorach danych, w których występują wartości nieokreślone (oszacowanie bez formalnych dowodów),
5. eksperymentalnej weryfikacji opracowanych metod i algorytmów dla kilku baz danych pochodzących z publicznie dostępnego repozytorium UCI,
6. eksperymentalnej weryfikacji opracowanych metod i algorytmów dla rzeczywistych danych medycznych dotyczących oceny stopnia włóknienia wątroby u pacjentów z wirusowym zapaleniem wątroby typu C.

Recenzowana praca ma charakter koncepcyjno-eksperymentalny. Autor zaproponował rozwiązanie problemu klasyfikacyjnego, w którym występują braki danych. Uzyskane przez Autora rezultaty potwierdzają postawioną na wstępnie pracy tezę badawczą, że możliwe jest utrzymanie jakości klasyfikacji mimo występowania danych niepełnych. W tym celu zaproponowano klasyfikator złożony z bazowych homogenicznych klasyfikatorów k-NN wykorzystujących pojedyncze cechy oraz ewentualnie współczynniki wagowe zależne od prawdopodobieństwa *a priori* klas. Kierunek badań naukowych omawiany w dysertacji jest niewątpliwie ważny zarówno z praktycznego jak i poznawczego punktu widzenia.

Prace Autora, dotyczące tematyki poruszanej w dysertacji, są znane międzynarodowej społeczności naukowej, czego wyrazem są dwie publikacje posiadające współczynnik wpływu IF. Dodatkowo należy podkreślić, że czternaście prac współautorskich indeksowanych jest w bazie WoS.

R Rutkowski

Uwagi krytyczne i dyskusje

Przykład opisany w rozdziale nr 1, którego wizualizacja znajduje się na Rys. 1.1 nie jest jednoznaczny z punku widzenia przyjętych kryteriów oceny zadania rozpoznawania. Przedstawia on faktycznie problem, w którym klasy w przypadku *a* są niezrównoważone, a w przypadku *b* są zrównoważone. Równocześnie jednak przykład *a* przedstawia problem dwuklasowy, dwuwymiarowy liniowo nie separowalny, natomiast przykład *b* problem liniowo separowalny. Autor dysertacji w opisie zagadnienia niezrównoważenia klas pominą problem ich liniowej separowalności.

W zagadnieniu łączenia wielu klasyfikatorów bazowych w jeden komitet klasyfikatorów wyróżnia się dwie podstawowe przestrzenie, w których może być dokonywana integracja klasyfikatorów bazowych. Jest to tzw. przestrzeń odpowiedzi lub przestrzeń funkcji dyskryminacyjnych. W pracy Autor stosuje metodę głosowania (integracja w przestrzeni odpowiedzi) w przypadku klasyfikatora RS*k*-NN lub metodę sumacyjną (integracja w przestrzeni funkcji dyskryminacyjnych) w przypadku klasyfikatorów SF*k*-NN oraz SF*k*-N/C. Uwzględniając założenie o nieparzystej liczbie cech, metoda głosowania większościowego mogłaby być zastosowana w algorytmach SF*k*-NN oraz SF*k*-NN/C.

Występujące we wzorze 7.7 prawdopodobieństwo opisane jako „prawdopodobieństwo przynależności obiektu" jest prawdopodobieństwem *a priori* klasy o etykiecie *c*. W tym kontekście Autor do wyznaczenia współczynników wagowych poszczególnych etykiet klas używa prawdopodobieństw *a priori* klas.

Czwarty cel pracy został zdefiniowany jako „Opracowanie algorytmu podejmowania decyzji (fusera) komitetu klasyfikatorów". Punkt 7.3, który wg. Autora ma potwierdzać realizację tego celu zawiera propozycję uwzględnienia wag w metodzie sumacyjnej, stosowanej w łączeniu klasyfikatorów bazowych. Punkt nie zawiera zatem nowej metody fuzji, lecz sposób wyliczenia wag.

W punkcie 7.3 Autor dysertacji uzasadnia wprowadzenie modyfikacji klasyfikatora SF*k*-NN polegającej na uwzględnieniu wag, które mają „wspierać" klasę mniejszościową. W przypadku współczynnika nazywanego przez Autora lokalnym (wzór 7.9) brakuje dyskusji jego wpływu na zrównoważenie klas. Przy założeniu losowości danych brakujących, zaproponowany system wagowy może preferować w rzeczywistości klasę liczniej reprezentowaną w całym zbiorze uczącym. Taka sytuacja może wystąpić np. gdy wszystkie braki danych (dotyczące pojedynczej cechy) będą występowały tylko dla obiektów należących do jednej klasy. Wówczas klasa dominująca (globalnie, uwzględniając wszystkie cechy) może stać się klasą mniejszościową (lokalnie, uwzględniając pojedynczą cechę).

Przykład analizowany w punkcie 7.4 jest słuszny przy bardzo silnych założeniach dotyczących pojawiających się braków danych w zbiorze uczącym. Jednym z nich jest liczba brakujących danych w poszczególnych wektorach zbioru uczącego, która jest taka sama. W przypadku danych generowanych losowo łatwo spełnić przyjęte założenia, natomiast dla danych rzeczywistych mogą one stanowić istotne ograniczenie.

W punkcie 8.3 przedstawione są wyniki eksperymentów dla 5 baz danych. W opisie badań eksperymentalnych oraz w wynikach zawartych w dalszej części pracy Autor wykorzystuje 6 baz danych. Omawiany punkt zawiera stwierdzenie „danych pochodzących ze wszystkich baz", czyli z pięciu czy sześciu? Autor nie przedstawienia również wyników jakości klasyfikacji poszczególnych klasyfikatorów bazowych, analizy wpływu wartości parametru *k* (w przypadku klasyfikatora *k*-NN), czy też specyfiki działania omawianych klasyfikatorów bazowych w procesie ich uczenia z wykorzystaniem tylko jednej cechy.

Analiza statystyczna wyników eksperymentów została wykonana z wykorzystaniem testu post-hoc Bonferroniego-Dunna. W przypadku wszystkich tabel (Tab. 8.21-8.24), dla których przeprowadzono wspomniany test liczba klasyfikatorów, baz danych oraz przyjęty

Z Rewdel

poziom ufności nie ulegają zmianie. W związku z powyższym wartość różnicy krytycznej powinna być taka sama dla wszystkich wyników zawartych w wymienionych tabelach. Przestawienie graficzne umieszczone pod odpowiednimi tabelami sugeruje nierówne wartości różnicy krytycznej. Dodatkowo w treści pracy nie podano wartości numerycznej różnicy krytycznej uzyskanej dla przyjętych danych (liczba klasyfikatorów, baz danych oraz przyjęty poziom ufności).

Autor dysertacji zdefiniował 7 celów badawczych. Cele o numerach 6 i 7 powinny być zdefiniowane jako jeden cel, tym bardziej, iż w zakończeniu pracy jest mowa o realizacji szóstego, a na str. 91 o realizacji siódmego celu pracy. W oby przypadkach Autor odnosi się do badań dotyczących oceny stopnia włóknienia wątroby u pacjentów z wirusowym zapaleniem wątroby typu C.

Uwagi redakcyjne i formalne

Recenzowana rozprawa napisana jest starannie pod względem językowym, stylistycznym oraz redakcyjnym. Występują nieliczne, drobne błędy, takie jak:

- strona 14: „Ostatnim z … przykładów klasyfikatorów … typowym przykładem jest…", dwukrotne użycie słowa przykład w jednym zdaniu złożonym,
- strona 51: „… Klasyfikator", powinno być „… klasyfikator",
- strona 51, 99, 103, 104: tzw. wiszące znaki,
- różne oznaczenia klasyfikatora ze współczynnikiem wagowym – SF*k*-NN(/C) bądź SF*k*-NN/C,
- algorytm *k*-NN (*k* najbliższych sąsiadów) przedstawiony jest dwukrotnie, w rozdzielne 2 oraz 6.

Autor posługuje się poprawnie nomenklaturą naukową dotyczącą metod uczenia maszynowego, jednak w pracy można znaleźć pewne nieścisłości, takie jak:

- strona 14: „decyzję klasyfikatora *a posteriori*", wyznaczyć można prawdopodobieństwo *a posteriori*, które wykorzystywane jest przez klasyfikator Bayesa,
- strona 69: „minimalną różnicę rangi", powinno być „minimalną różnicę średnich rang",
- Autor często wykorzystuje pojęcie „określoną wartość" (np. str. 41), czy też „wartości cech" (np. str. 52). W kontekście zdań, w których te pojęcia są używane, jest to skrót myślowy nierozróżniający prawidłowo jednej konkretnej wartości cechy (dopuszczalnej wartości pochodzącej z dziedziny) od dowolnej wartości niebędącej tzw. wartością nieokreśloną (*null*).
- Rys. 8.2 zawiera niejednoznaczne połączenie wykonane linią przerywaną między zbiorem uczącym (referencyjnym) a obiektem klasyfikowanym.
- Rozmieszczenie wartości pustych w zbiorze referencyjnym, czy też testowym jest niezależne od rodzaju klasyfikatora, który zostanie wykorzystany w procesie rozpoznawania. Tytuł rozdziału 7.4 sugeruje, że przedstawione w nim rozmieszczenie wartości nieokreślonych jest charakterystyczne dla klasyfikatora oznaczonego jako SF*k*-NN(/C).
- Tytuł punktu 8.1 sugeruje, że w rozdziale tym zostanie uzasadniona struktura komitetu klasyfikatorów. W literaturze wyróżnia się komitety o strukturze równoległej (wykorzystywane w recenzowanej dysertacji) oraz komitety o strukturze szeregowej. Rozważania zawarte w omawianym punkcie uzasadniają wykorzystanie pojedynczych cech przez każdy klasyfikator bazowy wchodzący w skład komitetu, a nie strukturę tego komitetu.

R. Rowicki

Podsumowanie

Reasumując stwierdzam, iż mgr inż. Tomasz Orczyk posiada ogólną wiedzę teoretyczną z dziedziny uczenia maszynowego. Recenzowana praca zawiera sformułowaną tezę pracy, która została udowodniona doświadczalnie przy wykorzystaniu danych rzeczywistych, pochodzących z ogólnie dostępnego repozytorium danych UCI oraz danych medycznych uzyskanych dzięki współpracy ze Śląskim Uniwersytetem Medycznym w Katowicach. Lektura dysertacji pozwala stwierdzić, że Autor zaprezentował na jej łamach umiejętność samodzielnego prowadzenia pracy naukowej.

Wobec powyższego, recenzowana praca spełnia wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami). Konkludując, wnoszę o przyjęcie rozprawy oraz dopuszczenie mgra. inż. Tomasza Orczyka do publicznej obrony.

R. Burduk