Dr hab. inż. Janusz Jeżewski, prof. ITAM
Instytut Techniki i Aparatury Medycznej ITAM
janusz.jezewski@itam.zabrze.pl

Zabrze, 22 maja 2018

RECENZJA

rozprawy doktorskiej magistra inżyniera Tomasza Orczyka

pt.: „Klasyfikacja danych niekompletnych w oparciu o komitet klasyfikatorów"

Promotor: dr hab. Piotr Porwik

Promotor pomocniczy: dr hab. inż. Bartłomiej Płaczek

Opiniowana rozprawa doktorska dotyczy problematyki klasyfikacji danych, a w szczególności opracowania efektywnej metody analizy danych rzeczywistych, które cechuje niekompletność oraz brak zrównoważenia. Z tego rodzaju danymi, które są szczególnie trudne do automatycznej oceny, mamy zwykle do czynienia w przypadku analizy zagadnień biomedycznych. Wysoka skuteczność automatycznej oceny danych biomedycznych decyduje o możliwości wykorzystania klasyfikatora dla potrzeb wspomagania lekarza przy podejmowaniu obiektywnych decyzji diagnostycznych.

Postawione zagadnienie jest ważne zarówno z punktu widzenia badawczego, jak i niezwykle przydatne z punktu widzenia potrzeby wspomagania decyzji w ramach praktyki klinicznej. Tworzenie prawidłowej metodologii automatycznej klasyfikacji danych niekompletnych i/lub niezrównoważonych, ze szczególnym uwzględnieniem skutecznej analizy danych rzeczywistych, stanowi czynnik bardzo istotnie wzbogacający stan wiedzy w zakresie budowy systemów rozpoznawania obiektów czy wspomagania procesu podejmowania decyzji.

W celu rozwiązania problemu skutecznej oceny danych niekompletnych wykorzystano opracowaną w ramach rozprawy metodę opartą na homogenicznym komitecie (zespole) równolegle działających klasyfikatorów typu *k*-najbliższych sąsiadów (*k*-NN), analizujących rozłączne, jednowymiarowe podzbiory przestrzeni cech klasyfikowanego obiektu. Uszczegółowienie metody sprowadzało się do wprowadzenia współczynników wagowych pozwalających na poprawną klasyfikację zbiorów danych niezrównoważonych, w których liczba obiektów poszczególnych klas nie jest jednakowa. Opracowany klasyfikator nie wymaga wstępnego przygotowania danych, a w szczególności uzupełniania brakujących

wartości, czy też równoważenia klas w zbiorze referencyjnym. Zastosowane przez Doktoranta do osiągnięcia celu postawionego w rozprawie, metodologia i narzędzia programistyczne, są zgodne z najnowszymi trendami panującymi w dziedzinie współczesnych metod inteligencji obliczeniowej.

Postawiona w rozprawie teza „Możliwe jest utrzymanie dokładności klasyfikacji na danych niepełnych poprzez wyłonienie komitetu klasyfikatorów działających w oparciu o wstępnie wyselekcjonowane cechy", jest oryginalna i istotna z naukowego punktu widzenia, a opracowane dla wykazania (uprawdopodobnienia) tej tezy metody i algorytmy z pewnością przyczynią się do poprawy efektywności automatycznej oceny danych typu rzeczywistego.

Przedstawiona do recenzji rozprawa obejmuje 113 stron druku. Zawiera wstęp, siedem rozdziałów i podsumowanie. Cały wywód rozprawy doktorskiej przeprowadzono w oparciu o wybrane i właściwie cytowane 74 pozycje piśmiennictwa. Należy podkreślić fakt, że w cytowanej bibliografii znajdują się 3 publikacje naukowe współautorstwa Doktoranta, dotyczące tematyki rozprawy. Świadczy to o tym, że prezentowana rozprawa jest naturalną kontynuacją i konsekwencją jego ukierunkowanych zainteresowań.

Do szczególnie wartościowych i oryginalnych elementów rozprawy należy zaliczyć:

1. Opracowanie struktury komitetu klasyfikatorów typu *k*-NN (SF*k*-NN, ang. Separate Features k-NN), działających w oparciu o jednowymiarowe, rozdzielne podzbiory cech analizowanego obiektu, charakteryzującego się wysoką dokładnością klasyfikacji danych niepełnych. Wśród zalet metody należy wyróżnić możliwość wykonywania obliczeń równoległych, co jest wynikiem przyjęcia niezależnej analizy cech, i co umożliwia skrócenie wymaganego czasu obliczeń. Dla potrzeb skutecznej oceny danych niezrównoważonych opracowano ważoną wersję algorytmu SF*k*-NN (SF*k*-NN/C), w której współczynniki wag mogą być wyznaczane globalnie (dla całego zbioru danych) lub lokalnie (dla poszczególnych cech obiektów).
2. Oszacowanie wielkości strat informacji w zbiorze danych o wartościach brakujących (tzw. wartościach *null*), w zależności od rozmiaru wektora cech opisującego analizowane obiekty. Dodatkowo Autor opracował procedurę generowania danych niepełnych dla podprzestrzeni cech o różnych rozmiarach, której celem było określenie wpływu wielkości strat informacji na uzyskiwaną skuteczność komitetów klasyfikatorów, budowanych w oparciu o metodę *k*-NN. Pozwoliło to ocenić wpływ rozmiaru podprzestrzeni cech, na liczbę pełnych wektorów referencyjnych dostępnych do budowy

klasyfikatorów wchodzących w skład komitetu. To z kolei posłużyło do analizy wpływu brakujących wartości cech na uzyskiwaną przez komitet dokładność klasyfikacji.

3. Ocenę wpływu użyteczności algorytmów selekcji cech na skuteczność klasyfikacji danych niekompletnych za pomocą zaproponowanego przez Autora klasyfikatora SF*k*-NN/C. W pierwszym etapie przeprowadzonych eksperymentów, z rozważanych baz danych usunięto w sposób losowy 25% pojedynczych wartości różnych cech (wprowadzając 25% wartości *null*). W etapie drugim, tak przygotowane dane poddano analizie korzystając z wybranych metod selekcji cech: procedur rankingowych, takich jak Relief, CSF, korelacji Pearsona itp.; indywidualnego rankingu cech oraz rankingu cech dla klas na podstawie wartości wskaźników jakości klasyfikacji F-Measure SR(F) i G-Measure SR(G); a także metod opakowanych zarówno dla przeszukiwania w przód FFS, jak i w tył BFS. Dla oszacowania dokładności klasyfikacji z użyciem suboptymalnych zestawów cech, zastosowano walidację krzyżową typu *leave one out*. Ustalenie zestawu cech dystynktywnych nie tylko pozwoliło ograniczyć złożoność algorytmu klasyfikującego (zredukować czas obliczeń), ale również podnieść dokładność samej klasyfikacji za pomocą procedury SF*k* NN/C.
4. Przeprowadzenie oceny wpływu procesu doboru cech na skuteczność działania klasyfikatora SF*k* NN/C w oparciu o bazę danych z wynikami badań pacjentów zakażonych wirusem HCV. Baza ta zawierała wyłącznie niezrównoważone dane rzeczywiste, gdzie wektory danych opisujące poszczególne przypadki bardzo często były niepełne. Brakujące wartości wynikały z przyczyn naturalnych, w przeciwieństwie do sztucznie wprowadzanych luk w pozostałych bazach testowych. Wyniki eksperymentów przeprowadzonych dla wstępnie wyselekcjonowanego zestawu cech wskazały na wyższą skuteczność proponowanej w rozprawie metody, w odniesieniu do przyjętego zestawu metod referencyjnych. Wykazano tym samym możliwość praktycznego wykorzystania klasyfikatora SF*k*-NN/C dla potrzeb wspomagania diagnostyki zakażenia wirusem HCV.

Podczas lektury recenzowanej rozprawy znalazłem kilka zagadnień niewyjaśnionych lub dyskusyjnych, które zostały zestawione w następujących punktach:

1. W przedstawionym przez Autora przykładzie klasyfikacji danych niezrównoważonych (str. 3), podstawowym problemem nie jest zjawisko różnej liczebności analizowanych klas, ale głównie brak liniowej rozdzielności obiektów przedstawionych na rysunku 1.1.a. Można zauważyć, że obiekty należące do klasy „1" (□) „zajmują" całą rozważaną

przestrzeń. Niezależnie więc od liczby obiektów należących do klasy drugiej (Δ), bazując wyłącznie na kryterium położenia (odległości) w oryginalnej przestrzeni cech, nie będzie można oddzielić od siebie obiektów tych dwóch klas.

2. Wśród celów pracy zdefiniowanych przez Autora (str. 8) cel szósty pokrywa się z celem siódmym.
3. Przykład problematyki uzupełniania danych (str. 4) jest trafny, ale podobnie jak z przykładem zagadnienia analizy niezrównoważonych danych, nie do końca słuszny. Główny błąd polega na tym, że nie zaznaczono wszystkich wartości uzupełnionych, a jedynie te uzupełnione błędnie. Wartości prezentowane pogrubioną czcionką łatwo bowiem mogły zostać uznane za błędy grube i wyeliminowane z dalszej analizy, szczególnie jeżeli znane są fizjologiczne zakresy możliwych wartości uzupełnianej cechy (np. liczba krwinek białych). Uzupełnianie danych zawsze powinno odbywać się z odpowiednią krytyczną oceną wartości generowanych automatycznie przez zastosowany algorytm uzupełniania.
4. W opisie metod opakowanych (str. 25-30) brakuje jakichkolwiek odniesień do bibliografii, co może sugerować, że ich pomysłodawcą jest Autor rozprawy.
5. Ze wzoru 6.11 (str. 36) wynika, że wyższe wartości wag przypisane zostaną wektorom ze zbioru referencyjnego o większej odległości od analizowanego obiektu. Podstawiając $d_q = d_{min}$ otrzymujemy $w_q = 1/e \approx 0{,}37$, natomiast dla $d_q = d_{max}$ otrzymujemy $w_q = e^0 = 1$. Wskazuje to na wyższy wpływ na decyzję klasyfikatora, obiektów referencyjnych leżących dalej od obiektu klasyfikowanego, czyli odmiennie do opisywanych wcześniej przez Autora metod ważonych. Wskazane byłoby umieszczenie odpowiedniego komentarza przy opisie metody.
6. Na rysunku 6.1, str. 37 (a także na Rys. 7.1, str. 43 oraz Rys. 8.2, str. 53) Autor niepotrzebnie umieścił symbol sumy algebraicznej dla reprezentacji procesu wyznaczania ostatecznej decyzji komitetu klasyfikatorów. Zastosowanie innego oznaczenia lub odpowiedni opis ułatwiłby zrozumienie rysunku.
7. W Rozdziale 2 (str. 9) Autor słusznie zauważa, że w klasyfikatorach leniwych nie występuje etap „uczenia". Wprowadza jednak pojęcie trenowania „na (F - u) wymiarowym zbiorze uczącym" (str. 40). Rozumiem, że chodzi o wyznaczenie wektorów referencyjnych dla poszczególnych klasyfikatorów składowych komitetu. Tym niemniej zamienne stosowanie pojęć takich jak: „zbiór uczący", „zbiór referencyjny" czy „zbiór treningowy", bardzo utrudnia śledzenie prezentowanych w pracy rozważań.
8. Str. 42, 8 linia od dołu. Co oznacza „projekcja f-tej cechy obiektu klasyfikowanego"?

9. Str. 44, 4 linia od dołu. Autor podał, że „Jako miarę odległości *d* przyjęto tutaj odległość euklidesową". W przypadku jednowymiarowych danych owa odległość euklidesowa będzie identyczna jak odległość Manhattan czy też Czebyszewa (Tabela 6.1). Z kolei brakuje informacji o rodzaju zastosowanej miary odległości w eksperymentach, gdzie analizowane są wielowymiarowe wektory cech.
10. Str. 44, 5 linia od góry. Autor powinien dodać, że opisywana niejednoznaczność dla metody SF k-NN/C nie zostanie zlikwidowana, w przypadku gdy zbiór zawiera klasy równoliczne oraz gdy zastosowana zostanie globalna funkcja wsparcia.
11. Str. 59, 2 linia od dołu. Autor używa terminu „sprawiedliwe rozwiązanie". Rozumiem intencję Autora, jednakże zależność jakości klasyfikacji od nastaw parametrów silnie zależy od konstrukcji samego algorytmu. Autor słusznie zauważa, że w przypadku metod ważonych, wpływ wartości *k* na uzyskiwane wyniki jest mniejszy, jednakże kosztem pewnego wzrostu złożoności obliczeniowej algorytmu. Stąd wydaje się, że równie a może i bardziej „sprawiedliwym rozwiązaniem", jest próba doboru wartości *k*, gwarantującej najlepszy z możliwych wyników klasyfikacji dla danej metody.
12. Rozdział 8.5. Autor stwierdza, że „bazy posiadały losowo usunięte wartości cech (*null*), wygenerowane analogicznie jak w poprzednich eksperymentach (podrozdział 8.4)", jednakże Tabele 8.28 – 8.34 przedstawiają wyniki (najprawdopodobniej) wyłącznie dla jednego przyjętego poziomu zawartości „null" w danych, ale nie wiemy dla jakiego?
13. Rozdział 9. Podsumowanie jest bardzo skromne, co przy braku dodatkowego Autoreferatu bardzo utrudnia czytającemu ocenę szczególnie wartościowych i oryginalnych elementów rozprawy.

W rozprawie znalazłem jeszcze kilka drobnych nieścisłości, należą do nich:

1. Dodatek A: Wykaz oznaczeń stosowanych w rozprawie jest bardzo ubogi. Dodatkowo przyjęty styl oznaczeń cechuje pewna nieścisłość. Autor stosuje identyczny format oznaczeń przestrzeni, zbioru oraz liczby elementów zbioru – wielka litera zapisana pochyloną czcionką. W niektórych fragmentach pracy, zbiór elementów oznaczany jest jednak przez Autora odmiennie – czcionką pogrubioną, bez pochylenia. Przykładowo wszystkie oznaczenia:
   - **R** – zbiór danych uczących (str. 2),
   - *A* – zbiór wektorów referencyjnych (str. 10),
   - **X** – zbiór danych uczących (str. 26),

- A – zbiór referencyjny (str. 32),
- X – zbór danych wejściowych (str. 36),

symbolizują zbiór Q wektorów cech wymiaru F, reprezentujących klasyfikowane obiekty. Jednocześnie w wykazie oznaczeń jest: X – przestrzeń cech (str. 107). Zastosowanie jednolitego aczkolwiek bardziej różnorodnego formatowania tych oznaczeń ułatwi czytelnikowi zrozumienie przedstawianych w pracy rozważań.

2. Nie do końca wydają się być trafione sformułowania: „dane historyczne" (np. str. 3), „losowe instancje" (str. 6), „opiera się o" (str. 40).
3. Rys. 8.4 jest zbyt mały aby mógł być czytelny.
4. W Tabeli 3.2 brakuje skrótów: SEN, SPE i ACC.
5. W Tabeli 3.1 używane są zwroty „pozytywny i negatywny", zaś na str. 93 w tym samym znaczeniu zwroty „dodatni i ujemny".
6. Na Rys. 8.3. zaprezentowano wyniki dla 5 spośród 6 baz danych, dlaczego pominięto rezultaty dotyczące bazy CRYO?

Wskazane w niniejszej recenzji: niepełny opis w pewnych fragmentach pracy, drobne błędy czy nieścisłości, oraz dyskusyjna metodyka badań w ramach niektórych eksperymentów; nie wpływają na sformułowanie końcowego wniosku opinii, a służą wyłącznie do poprawy przyszłych publikacji Doktoranta.

Podsumowując przedstawianą opinię stwierdzam, że magister Tomasz Orczyk wykazał się wiedzą oraz umiejętnościami wymaganymi do uzyskania stopnia doktora nauk technicznych, gdyż samodzielnie rozwiązał istotne zadanie naukowe z dyscypliny naukowej Informatyka. Wymienione powyżej uwagi, nie zmieniają mojej pozytywnej oceny pracy, zaś przedstawiona rozprawa doktorska spełnia wymagania przewidziane w ustawie o stopniach i tytule naukowym. W związku z tym stawiam wniosek o przyjęcie tej pracy jako rozprawy doktorskiej i o dopuszczenie jej Autora – magistra Tomasza Orczyka do jej publicznej obrony.