dr hab. inż. Robert Burduk
Politechnika Wrocławska
Wydział Elektroniki
Ul. Wybrzeże Stanisław Wyspiańskiego 27
50-370 Wrocław

Wrocław, dnia 2016-11-30

RECENZJA

rozprawy doktorskiej mgra inż. Marcina Giberta
na temat: **„Integracja metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji"**

Problem badawczy i jego znaczenie

Zakres rozprawy dotyczy metod eksploracji danych i mieści się w dziedzinie nauk technicznych, dyscyplina nauki *Informatyka*. Recenzowana praca związana jest z problematyką klasyfikacji nadzorowanej, w której dostępne dane mogą być typu numerycznego jak i tekstowego. Doktorant dla tak zdefiniowanego problemu badawczego proponuje oryginalne podejście integrujące metody eksploracji danych używane w przypadku dysponowania danymi tekstowymi oraz numerycznymi. Wykorzystane w pracy metody „bazowe" są dobrze znane w literaturze i mają wiele cech, dzięki którym są często stosowane w zadaniach praktycznych.

W rozprawie sformułowano hipotezę badawczą, która zawiera stwierdzenie, że integracja metod dotyczących klasyfikacji danych tekstowych oraz numerycznych zwiększa tzw. nośność informacyjną. Postawiona hipoteza została uszczegółowiona za pomocą dodatkowych złożeń oraz zweryfikowana na drodze badań eksperymentalnych. W eksperymentach użyto trzech zbiorów danych związanych z realnymi problemami podejmowania decyzji.

Tematyka rozprawy oraz postawiony w niej problem badawczy dotyczą zagadnień eksploracji danych, która ma swoją ugruntowaną pozycję w dyscyplinie naukowej *Informatyka*. W związku z aktualnością problemu badawczego, w którym mamy do czynienia z różnymi źródłami danych, a w konsekwencji danymi o różnych typach, opracowane rozwiązanie posiada duży potencjał zastosowania w zagadnieniach przetwarzania dużej ilości danych (Big Data).

Struktura pracy oraz wiedza Autora

Recenzowana praca została napisana w języku polskim i liczy 164 strony maszynopisu. Składa się z siedmiu rozdziałów, spisu rysunków, tabel i symboli oraz literatury. Odliczając numerowane rozdziały „Wprowadzenie" oraz „Podsumowanie" recenzowana dysertacja składa się z pięciu merytorycznych rozdziałów. Rozdziały o nr 2 oraz 3 opisują stan wiedzy dotyczący odpowiednio klasyfikacji danych tekstowych oraz numerycznych. Przedstawiona w nich treść wskazuje, że Autor posiada wiedzę teoretyczną, która dotyczy omawianej w pracy problematyki i mieści się w nurcie badań związanych z eksploracją danych. W tekście można znaleźć wiele prostych przykładów ilustrujących omawianą problematykę.

Oryginalne rozwiązanie problemu badawczego zostało zaprezentowane w rozdziale 4. Znajduje się tam opis procedury potrzebnej do integracji danych tekstowych oraz numerycznych. Autor zaproponowaną metodę podzielił na cztery logiczne części: eksploracja wstępna oraz właściwa danych tekstowych, reprezentacja danych numerycznych oraz ich eksploracja.

Rozdział 5 zawiera opis oraz wyniki badań eksperymentalnych, które zostały wykonane na trzech rzeczywistych zbiorach danych. W celu udowodnienia postawionej w pracy hipotezy badawczej Autor porównał proponowane przez siebie rozwiązanie z trzema innymi modelami. Jeden z nich wykorzystuje wyłącznie dane numeryczne, drugi z nich wyłącznie dane tekstowe, a trzeci nazywany jest w pracy modelem integrującym wyniki dwóch wcześniej wymienionych. Opis badań eksperymentalnych zawiera wiele przykładów przedstawiających stosowane metody przetwarzania danych, co znacznie ułatwia lekturę pracy.

Ostatni merytoryczny rozdział, rozdział 6, zawiera dyskusję wyników. Celem tej dyskusji jest udowodnienie postawionej w rozdziale 1 hipotezy badawczej. Należy powiedzieć, że przy pewnych uwagach przedstawionych w dalszej części recenzji, postawioną hipotezę należy uznać za udowodnioną.

Spis literatury liczy 110 pozycji. Cytowane prace dobrane są prawidłowo i odnoszą się do omawianych problemów. Drobną uwagą jest stosunkowo mała liczba pozycji literaturowych opublikowanych w renomowanych czasopismach naukowych.

Wkład Autora – oryginalne osiągnięcia

Wkład Autora w rozwój metod eksploracji danych, dedykowanych dla zbiorów z numerycznym i tekstowym typem danych, polega na:

1. opracowaniu nowego algorytmu eksploracji danych, który pozwala na klasyfikację zbioru posiadającego dane typu numerycznego jak i tekstowego,
2. eksperymentalnej weryfikacji opracowanego algorytmu dla trzech różnych praktycznych problemów decyzyjnych, takich jak:
   - wyszukiwanie rentownych zamówień publicznych,
   - inwestowanie na Giełdzie Papierów Wartościowych,
   - wyszukiwanie atrakcyjnych ofert pracy,
3. eksperymentalnemu porównaniu zaproponowanego algorytmu z innym algorytmami, które wykorzystują dostępne dane numeryczne lub tekstowe.

Recenzowana praca ma charakter koncepcyjno-eksperymentalny. Autor zaproponował rozwiązanie problemu klasyfikacyjnego, w którym dostępny zbiór danych posiada dwa różne typy danych, tzn. numeryczny oraz tekstowy. Następnie, na drodze eksperymentalnej, została potwierdzona postawiona w pracy hipoteza badawcza. Uzyskane przez Autora rezultaty wskazują, ze uzasadniona i sensowna jest integracja metod eksploracji danych tekstowych oraz numerycznych. Kierunek badań naukowych omawiany w dysertacji jest niewątpliwie ważny zarówno z praktycznego jak i poznawczego punktu widzenia.

Wykorzystane w badaniach eksperymentalnych zagadnienia należy uznać za bardzo interesujące z praktycznego punktu widzenia. Dotyczą one rzeczywistych problemów decyzyjnych, które w swojej działalności podejmują firmy szukające informacji o przetargach, rekrutujące pracowników lub też uczestniczące w obrocie papierami wartościowymi. Rezultaty pracy Doktoranta mogą mieć zatem bardzo duży oddźwięk praktyczny.

Prace Autora znane są międzynarodowej społeczności naukowej, czego wyrazem są dwie samodzielne publikacje prezentowane na międzynarodowych konferencjach, których materiały ukazały się seriach wydawanych przez wydawnictwo Springer. Dodatkowo Autor dysertacji wymieniany jest jako współautor dwóch publikacji, które znajdują się w recenzowanych czasopismach z tzw. listy B MNiSW. Należy jednak zauważyć, że jedna z publikacji (Gibert M., Improving Information-Carrying Data Capacity in Text Mining. Lecture Notes in Computer Science vol. 9330, 648-657, 2015) nie znalazła się w spisie literatury recenzowanej rozprawy. Dodatkowo pozycja 88 wg. bazy „BazTech" powinna być zatytułowana „Classification of text documents by using expanded terms in Latent Semantic Analysis". Dorobek publikacyjny Autora dysertacji jest zatem wystarczający do ubiegania się o stopień doktora.

Uwagi krytyczne i dyskusje

Hipoteza badawcza zawiera sformułowanie „wielokryterialnym procesie podejmowania decyzji". W całej dysertacji Autor tylko jeszcze raz użył pojęcia „wielokryterialny" w sformułowanych zaraz po tezie założeniach dotyczących pracy. W założeniach tych słusznie zaznaczono, że w szczególnym przypadku możemy mówić o decyzji jednokryterialnej. Autor powinien bardziej szczegółowo uzasadnić użyte do sformułowania hipotezy pojęcie wielokryterialności.

Przedstawione na str. 13 założenie dotyczące realizacji pracy i odnoszące się do postawionej hipotezy badawczej sprowadza badanie tzw. nośności informacyjnej do badania jednego z kryteriów mierzącego jakość klasyfikacji. Uwzględniając również wspomnianą wcześniej uwagę hipoteza badawcza mogłaby być inaczej sformułowana. Skutkiem tego byłaby mniejsza liczba założeń przyjętych w pracy.

Ocena jakości klasyfikacji w całej pracy dokonywana jest przez Autora przy wykorzystaniu wskaźników jakimi są błąd oraz jakość klasyfikacji. Są to wskaźniki komplementarne. Oznacza to, że znając wartość jednego z nich w trywialny sposób uzyskujemy wartość drugiego wskaźnika, ponieważ suma ich wartości równa się jedności. Autor przedstawiając wyniki eksperymentów umieszcza na rysunkach oraz w tabelach zawsze wartości tych dwóch wskaźników. Taka prezentacja zmniejsza czytelności, w szczególności tabel. Najlepszym tego przykładem są tabele 68–74, w których występują nadmiarowe kolumny, ponieważ z równań 148 i 149 otrzymuje się takie same wartości.

Przedstawione przez Autora trzy przypadki problemów decyzyjnych są zadaniami, w których występują dwie etykiety klasy. Autor opisuje różne miary jakości klasyfikacji, które dotyczą klasyfikacji binarnej (str. 9 oraz 62–64) i nie są tak zagregowanymi wskaźnikami jak błąd oraz jakość klasyfikacji. Zastosowanie wspomnianych wskaźników przyczyniło by się do bardziej wnikliwej analizy otrzymanych wyników

Wariant „D" opisywany jest przez Autora jako „zintegrowany wynik eksploracji z wariantów B i C". Opis integracji jest bardzo lakoniczny i powtarza się trzykrotnie (str. 101, 118 oraz 136), co wynika z opisów badań eksperymentalnych dla trzech różnych przykładów procesów podejmowania decyzji. W opisie tym pojawia się stwierdzenie „w tym wariancie polega na wyborze korzystniejszego". Stwierdzenie to sugeruje prostą selekcję najkorzystniejszego wyniku (z dwóch możliwych), niżeli integrację wyników z dwóch różnych wariantów.

W ocenie jakości modeli klasyfikacyjnych stosuję się podział dostępnych danych na zbiór uczący, walidacyjny oraz testujący. Opis badań eksperymentalnych nie zawiera jednoznacznego wskazania jak wymagane do oceny jakości klasyfikacji zbiory zostały stworzone. W każdym z omawianych przykładów znajduje się akapit mówiący o „na zbiorze 200 testowych przypadków ... z wykorzystaniem 11 przypadków treningowych" . Czy należy rozumieć, że zbiór uczący (treningowy) zawiera 11 a testujący 200 obiektów? Dalszy opis sugeruje, że przypadki (obiekty) pochodzą z jednej klasy.

W punktach 5.2–5.4 Autor posługuje się pojęciem „poziom zwrotu", które nie zostało w należyty sposób wyjaśnione. Pojęcie to pojawia się po raz pierwszy w tabeli 15 i może sugerować, że należy je łączyć ze stosowanym przez Autora pojęciem „11 przypadków treningowych".

Analiza statystyczna wyników eksperymentów została wykonana za pomocą testu McNemary. W pracy występują więcej niż dwa modele decyzyjne, tak więc prezentacja otrzymanych rezultatów powinna zawierać wyniki wspomnianego testu w układzie „każdy z każdym". Dodatkowo Autor przedstawił tylko wybrane wyniki bez ich uzasadnienia. Powiazanie tej uwagi z brakiem jednoznacznego opisu dotyczącego podziału danych na zbiory: uczący, walidacyjny oraz testujący powoduje, że analiza wyników uzyskanych i przedstawionych przez Autora jest utrudniona.

W treści punktu 2.2 (str. 32) znalazła się wzmianka o istotności doboru parametru „k" występującego w metodzie k-NN na wynik klasyfikacji. Autor niestety nie przedstawił wyników badań, uwzględniających różne wartości tego parametru.

Uwagi redakcyjne i formalne

- Strona 4: „przydatną i popularna", „decyzyjnego procesy", brak spacji w odnośniku „[94,s.17]", zmiana wielkości czcionki w ostatnim akapicie.
- Strona 5: Jedno ze zdań zaczyna się od słowa „Zaś".
- Strona 8 i inne: Występuje „proces podejmowania decyzji PD" oraz „proces PD". Po wprowadzeniu akronimu PD pierwsza wymieniona forma jest niepoprawna.
- Strona 9: „może być podstawową miarę procesu".
- Strona 11: „wskazują na główny problem do rozwiązania w przyszłych badaniach podkreślając".
- Strona 13: „Dlatego w do".

- Strona 17: „Drugie podejście większym stopniu".
- Strona 17: „na uczenie maszynowym".
- Strona 32: Funkcja sim() występuje po lewej i prawej stronie równania (17), po lewej powinna być ze znakiem wyróżniającym np. tzw. daszkiem.
- Strona 56: Wzory o numerach 74 i 75 są takie same.
- Strona 57: „W niektórych sytuacja ekspert".
- Strona 58: „Poprawność wyboru atrybutów, dyskretyzacji wartości tych atrybutów".
- Strona 59: „zdefiniowana", powinno być „zdefiniowaną".
- Strona 70: „celu języka", „co znacznie usprawnia proces konstruowanie".
- Strona 74: „dopierany".
- Strona 75: „jest funkcji istotności".
- Strona 76: Liczba zamiast „ilość", ta sama uwaga na str. 97.
- Strona 83: „MyNemara".
- Strona 84: „Zadanie decyzyjne polega wytypowaniu", „badanie testowo", „zgodności otrzymanych wynikach".
- Strona 86: „rzozdziału".
- Strona 88: „powiązania form fleksyjne".
- Strona 93: „Dla reguły pewnych oraz niepewnych obliczono".
- Strona 96: „W tym celu Poszczególne".
- Strona 101: „na rysunkach 22."
- Strony 101, 102, 120, 137, 138: "wartości miar jakości decyzji ERR", w tym wypadku jednej, czyli powinno być „miary".
- Strona 118: W opisie rysunku jest miara „F1 score", która nie jest przedstawiona.
- Strona 121: „sprawdzenia sądów", raczej „sprawdzenia hipotezy o ...".
- Strony 130, 134: Opis tabel wskazuje na liczbę reguł większą od jednej, w odpowiednich tabelach jest jedna reguła.
- Strona 131: „wzorem (1139)", poza zakresem numeracji istniejącej w pracy.
- Strona 139: „W efekcie analizy fleksyjne dokonano", „zbadano istotność atrybutu wynik eksploracji danych".
- Strona 140: „wynikających z ze zbioru".
- Strona 148: „Badania testowy przeprowadzone".
- Strona 152: Formatowanie pozycji [45].

- Strona 162: Drobne błędy w „Spisie symboli" np., „indeks wyrazu", powinno być „Indeks wyrazu".
- Kolejność odnośników do literatury np. str. 3 „[15, s. 21] [61, s. 4] [55, s. 398]" , str. 5 „[35] [64] [49]", str. 42 „[46][37][85]".
- Niekonsekwentne (nawet w tym samym numerowaniu lub opisie) stosowanie łącznika, półpauzy oraz myślnika (np. str. 6, 19, 51, 82).
- Autor stosuje nadmiarową ilość oznaczeń zawierających indeksy np. „mtw", „itw".
- W przypadku wielu rysunków i tabel nie zachowano stosowanych w pracy marginesów (np. str. 67, 71, 93, 134, 139).

Podsumowanie

Reasumując stwierdzam, iż mgr inż. Marcin Gibert posiada ogólną wiedzę teoretyczną w dziedzinie eksploracji danych. Recenzowana praca zawiera sformułowany i rozwiązany problem naukowy a badania eksperymentalne zostały przeprowadzone z użyciem danych rzeczywistych. Dodatkowo Autor zaprezentował umiejętności 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. Marcina Giberta do publicznej obrony.

R. Burduk