Tytuł, stopień, imię i nazwisko
dr hab. inż. Bogdan Kwolek, prof. n. AGH

data 27.06.2014

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: Bi-clustering – algorithms and applications

Autor rozprawy: mgr inż. Paweł Foszner

Promotor rozprawy: prof. dr hab. inż. Andrzej Polański

1. Obszar problemowy rozprawy

Przedstawiona mi do recenzji rozprawa doktorska dotyczy analizy skupień danych. Prace badawcze dotyczyły metodologii łączenia dwugrup z wykorzystaniem węgierskiej metody dla zagadnień przydziału, projektowania algorytmów wyższego poziomu do integracji wyników uzyskanych z różnych algorytmów dwugrupowania danych oraz wskaźników jakości umożliwiających ocenę skuteczności algorytmów dwugrupowania danych (oceny podobieństw między skupiskami danych). Wynikiem końcowym rozprawy są udoskonalone metody, które przebadano i porównano eksperymentalnie z pokrewnymi metodami w oparciu o oprogramowanie udostępnione przez Autora w Internecie. Praca ma charakter badawczo-doświadczalny. Autor zaproponował oryginalne rozwiązania, które zwiększają efektywność istniejących rozwiązań oraz zbudował narzędzia do badań eksperymentalnych. Jej głównym celem naukowym było wprowadzenie udoskonaleń do istniejących algorytmów, pokazanie, że są one bardziej skuteczne w porównaniu z istniejącymi algorytmami oraz opracowanie rozwiązań umożliwiających ocenę skuteczności różnych algorytmów. Wyniki badań eksperymentalnych są przekonujące i inspirujące. Badania doświadczalne jednoznacznie wskazują na użyteczność zaproponowanych rozwiązań.

Cele pracy zostały jasno określone w rozdziale 2. Tezy pracy przedstawiono w rozdziale 3. Postawione przez Doktoranta tezy są oryginalne i istotne z naukowego punktu widzenia, zaś opracowane w ramach rozprawy metodologie i algorytmy mające potwierdzić ich słuszność są oryginalne i istotne dla praktyki. Oryginalny wkład Autora został zaprezentowany w rozdziale 4. Zaproponowane rozwiązania odniesiono do rozwiązań znanych z literatury. Tematykę rozprawy należy uznać za interesującą, wpisującą się w nurt najnowszych badań w dziedzinie analizy skupień danych. Wynikiem końcowym pracy są udoskonalone algorytmy, zaproponowane miary podobieństwa oraz oprogramowanie do badań eksperymentalnych. Tematyka rozprawy jest w pełni uzasadniona, interesująca i aktualna.

2. Zawartość rozprawy

Rozprawa napisana jest w języku angielskim. Praca liczy 100 stron (z dodatkami 134 strony) i składa się z 10 rozdziałów oraz bibliografii. W rozprawie odwołano się do 54 pozycji bibliograficznych. Następstwo rozdziałów jest właściwe. Rozdział 1 stanowi wprowadzenie do problematyki rozprawy. W następnych 3 rozdziałach przedstawiono motywację podjęcia badań, omówiono cele pracy, sformułowano tezy pracy oraz zestawiono główne osiągnięcia. Zamieszczone w rozdziale 4 zestawienie oryginalnych osiągnięć jest rzetelne i umożliwia określenie oryginalnego wkładu Autora. Wnikliwa analiza przeprowadzonych wcześniej prac w zakresie analizy danych mikromacierzowych stała się punktem wyjściowym do podjęcia badań. Na podstawie analizy źródeł Autor wyciąga poprawne wnioski oraz prezentuje wkład rozprawy w rozwój dyscypliny. Zasadnicze problemy

wiążące się z porównywaniem podmacierzy w algorytmach dwugrupowania danych omówiono w rozdziale piątym. W rozdziale szóstym zamieszczono przegląd algorytmów dwugrupowania danych. W rozdziale siódmym zaprezentowano opracowane metody łączenia dwugrup z wykorzystaniem węgierskiej metody dla zagadnień przydziału, metody projektowania algorytmów wyższego poziomu do łączenia wyników uzyskanych przez różne algorytmy dwugrupowania danych oraz omówiono proponowane wskaźniki jakości do oceny skuteczności algorytmów dwugrupowania danych. W rozdziale ósmym uwagę skupiono na metodach graficznych prezentacji wyników. Dziewiąty rozdział poświęcono omówieniu opracowanego oprogramowania oraz uzyskanych wyników badań eksperymentalnych. Zasadnicza część rozprawy kończy się krótkim podsumowaniem uzyskanych wyników. W dodatku rozprawy zaprezentowano zbiorcze zestawienie uzyskanych wyników badań eksperymentalnych, które pokazują, że na danych syntetycznych proponowany algorytm wyższego poziomu uzyskuje najlepsze wyniki, mając na względzie miary jakości zdefiniowane w rozdziale 7.1.

Praca jest zredagowana logicznie. Autor przedstawia koncepcję rozprawy, formułuje cel główny i cele cząstkowe. Omawiane problemy są przedstawione poprawnie pod względem merytorycznym. Rozdziały pozostają w odpowiednim związku przyczynowo-skutkowym. Jej lektura pozwala prześledzić realizację założonego celu. Zaproponowane rozwiązania są należycie udokumentowane. Wnioski zaprezentowane w podsumowaniu rozprawy mają silne oparcie w proponowanych rozwiązaniach i uzyskanych wynikach badań eksperymentalnych.

3. Znaczenie uzyskanych wyników

Do wartościowych i oryginalnych elementów pracy należy zaliczyć opracowanie metody łączenia dwugrup z wykorzystaniem węgierskiej metody dla zagadnień przydziału, metody projektowania algorytmów wyższego poziomu do łączenia wyników uzyskanych przez różne algorytmy dwugrupowania danych oraz proponowane wskaźniki do oceny podobieństw pomiędzy skupiskami danych. Zaproponowany algorytm, który w pracy nazywany jest *consensus algorithm*, jest ogólnym rozwiązaniem o znaczącej odporności na rozmaite struktury danych, które mogą pojawiać się podczas analizy skupień danych mikromacierzowych. Celem wykazania, że algorytm jest lepszy od obecnie dostępnych metod wygenerowano zestawy danych syntetycznych, które odzwierciedlają zasadnicze struktury występujące w trakcie analiz mikromacierzowych, a następnie pokazano, że rozpatrywane wskaźniki jakości *relevance*, *recovery* i *consensus score*, zob. 7.1.2 i 7.1.3, przyjmują lepsze wartości. Dla algorytmów niedeterministycznych wyniki uśredniono ze 100 powtórzeń. Oprócz eksperymentów ilustrujących zachowanie algorytmu na poszczególnych strukturach danych przeprowadzono badania eksperymentalne na danych rzeczywistych dla których wykazano empirycznie, że zaproponowany algorytm uzyskuje lepsze wyniki dla dwóch wybranych zestawów danych rzeczywistych. Oprogramowanie opracowane w ramach pracy udostępniane jest w Internecie. Wyniki badań empirycznych uzyskane w oparciu o proponowany algorytm wskazują, że integracja dwugrup uzyskanych przez metody składające się na algorytm wyższego poziomu dla różnych wartości parametrów prowadzi do lepszych wyników końcowych i tym samym lepszych wartości wskaźników ACV (Average Correlation Value). Integracja dwugrup odbywa się z wykorzystaniem przypisań wyznaczonych przez algorytm węgierski oraz wskaźników wyrażających podobieństwa między danymi należącymi do analizowanych skupisk. Integracja odbywa się w oparciu o zaproponowaną miarę podobieństwa (rozdział 7.1) oraz zmodyfikowany algorytm węgierski (rozdział 7.3). Algorytm jest oryginalny, zaś wyniki badań eksperymentalnych są spójne, przekonujące i należycie udokumentowane. Na oryginalny dorobek Autora składa się także analiza porównawcza wyników uzyskiwanych przez rozpatrywane metody.

Praca wzbogaca naszą wiedzę, jest inspirująca i stanowi bardzo dobry materiał do dalszych prac. Rozwiązania zaproponowane w pracy są całościowe i mają nowatorski charakter. Autorskie propozycje rozwiązań wskazują na dobrą znajomość zagadnień związanych z przedmiotem rozprawy oraz na dobre wyczucie istoty prac o charakterze badawczym. Podjęty temat ma istotne walory praktyczne jak i poznawcze.

4. Uwagi o charakterze polemicznym

W pracy nie dostrzegłem istotnych błędów o charakterze merytorycznym. Przy starannym przejrzeniu rozprawy dostrzegłem kilka błędów o charakterze edycyjnym – niemniej liczba potknięć nie odbiega od normy. Drobne potknięcia znajdują się m.in. w zwrotach: "is to computed", "have been proposed similarity measure", "Its two way clustering method which perform simple clustering", "A very wide range of algorithms are algorithms", "Likelihood function introduce a model", "is a bi-clustering technique propose by", "Algorithm start", "This process can be also consider as", "this thesis propose a solution", "for witch", "Next step is to using those term build network", "Kullback-Liebler", "et al", "previews figure", "To confirm the described above thesis, were created synthetic data". W kilku miejscach pracy odwołano się do symbolu AVC zamiast do zdefiniowanego na str. 22 akronimu ACV (Average Correlation Value), niepoprawnie odwołano się do numeru rozdziału, np. na str. 43, numeru rysunku, np. na str. 56, 67. Pracy nie zaszkodziłoby ponumerowanie wzorów. Notacja we wzorach prezentowanych na str. 22 i 23 nie jest wystarczająco spójna. Przykładowo, w pierwszym wzorze w podrozdziale 5.2.3 powinno być: suma_ $i \in I$ zamiast suma_i=1^n, zob. także symbole użyte w mianownikach. Brak staranności w zapisie dotyczy także wzoru podanego w podrozdziale 6.1 oraz podrozdziale 6.1.1. Opis w podrozdziale 9.4.2.2 nie jest wystarczająco jasny. Przykładowo, po zdaniu "After this, consensus result is creating as follows:" (powinno być: is created as follows:) następuje nieprecyzyjny opis. W kontekście opisu przedstawionego na str. 60, w którym wspomina się o sortowaniu wartości wyrażających podobieństwo między grupami danych (w pracy użyto zwrotu: "sort experiments by this measure"), oraz w kontekście opisu algorytmu zamieszczonego w podrozdziale 7.4, nasuwa się pytanie, czy sortowanie jest wykorzystywane – a jeśli tak, to jaki algorytm wykorzystywano. Pytania dotyczą także sposobu doboru progów T_1, T_2, T_3, a także wpływu tych wartości progowych na uzyskiwane wyniki.

Powyższe uwagi mają specyficzny charakter, gdyż nie umniejszają w najmniejszym stopniu wartościowych osiągnięć, a ukierunkowane są na uczynienie z dobrej rozprawy jeszcze lepszej.

5. Ocena końcowa

Cele pracy zostały jasno sformułowane, a ich realizacja wymagała obszernych badań. Cele pracy zostały osiągnięte. Tezy pracy zostały wykazane empirycznie. Reasumując stwierdzam, że mgr inż. Paweł Foszner wykazał się odpowiednią wiedzą z zakresu analizy skupień danych, a także dobrym opanowaniem i posługiwaniem się warsztatem badawczym. Przedstawiona mi do recenzji rozprawa doktorska zawiera poprawnie sformułowany i rozwiązany problem badawczy. Zawarte w niej wyniki oraz rozwiązania są oryginalne i zostały zaprezentowane w logicznym układzie i całościowym ujęciu. Opracowane rozwiązania stanowią bardzo dobry materiał do dalszych prac badawczych. Wkład Autora w rozwój wiedzy w dyscyplinie związanej z tematyką pracy został należycie udokumentowany. Wyniki prac badawczych opublikowane zostały w szeregu artykułach naukowych, m.in. w Trans. Computational Collective Intelligence wyd. Springer, European Conf. on Math. and Theoretical Biology i wielu wartościowych periodykach/materiałach o zasięgu światowym. Proponowane rozwiązania są wartościowe i istotne ze względu na możliwości aplikacyjne.

Uważam, że recenzowana praca doktorska Pana Pawła Fosznera w pełni spełnia wszystkie zwyczajowe i ustawowe wymagania stawiane pracom doktorskim i wnioskuję o jej dopuszczenie do publicznej obrony.

Zygmunt Kwiatek