Prof. dr hab. inż. Konrad Wojciechowski
Politechnika Śląska
Wydział Automatyki Elektroniki i Informatyki
Instytut Informatyki

Gliwice 11.11.2014

Recenzja rozprawy doktorskiej

Tytuł rozprawy:

Bi-clustering – algorithms and applications

Autor rozprawy:

mgr inż. Paweł Foszner

Promotor rozprawy:

Prof. dr hab. inż. Andrzej Polański

Tematyka pracy

Praca jest napisana w języku angielskim, liczy 132 strony, 10 rozdziałów, zawiera bibliografię złożoną z 54 pozycji, listę symboli i skrótów, wykaz tabel i rysunków, dodatek zawierający wyniki rysunkowe i tabelaryczne pokazujące/porównujące zastosowanie badanych algorytmów bi-klasteryzacji do symulacyjnych zbiorów danych.

Oceniana praca doktorska poświęcona jest problemom bi-klasteryzacji danych. Wybraną przez doktoranta tematykę pracy należy uznać za współczesną i jednocześnie interesującą. Bi-klasteryzacja, czyli jednoczesne grupowanie wzdłuż wierszy i kolumn macierzy danych obserwacyjnych lub pomiarowych jest kierunkiem badań naukowych o szerokim wachlarzu zastosowań, w analizie tekstów, analizie obrazów, bio-informatyce, klasyfikacji danych itd. Liczne zastosowania dostarczają impulsów do rozwoju nowych algorytmów. Istnieje szeroka literatura dotycząca algorytmów bi-klasteryzacji, a także wydawane są pozycje monograficzne podsumowujące publikowane wyniki. Większości z prowadzonych w zakresie bi-klasteryzacji badań naukowych nie można uznać za zamknięte. Istnieje szereg kierunków badań, w ramach których prowadzi się intensywne prace. Łatwo sprawdzić w bazach danych publikacji naukowych, że liczba doniesień naukowych poświęconych różnym aspektom bi-klasteryzacji stale się zwiększa. Zwłaszcza intensywnie rozwijane są różne aspekty zastosowań algorytmów bi-klasteryzacji.

Należy także zwrócić uwagę, że bi-klasteryzacja jest bardzo szerokim zagadnieniem. W rozwoju metod bi-klasteryzacji można wyróżnić szereg kierunków takich jak różne warianty sformułowania problemu bi-klasteryzacji, algorytmy rozwiązywania zadań bi-klasteryzacji, porównywanie wyników uzyskiwanych z zastosowaniem różnych algorytmów bi-klasteryzacji, rozwijanie sposobów prezentacji lub wizualizacji wyników algorytmów bi-klasteryzacji, uogólnienia zadania bi-klasteryzacji na struktury danych o wyższych wymiarach, zastosowania metod bi-klasteryzacji do różnych problemów badawczych. Dla każdego z tych kierunków rozwoju ukazują się prace specjalistyczne. W aspekcie tych uwag można poddać krytyce tytuł rozprawy, który jest zbyt ogólny. Po pracy zatytułowanej „bi-klasteryzacja – algorytmy i zastosowania" można by spodziewać się przedstawienia całości zagadnień związanych z bi-klasteryzacją. Nie ma jednak możliwości objęcia w sensowny sposób tak szerokiej tematyki badawczej w pojedynczej pracy doktorskiej. Mimo dość obszernego spektrum problematyki naukowej pokrytego przez doktoranta, z pewnością oceniana praca doktorska nie obejmuje tak

szerokiego zakresu tematycznego, jaki wynikałby z jej tytułu. W rzeczywistości rozważania przestawione dalej w pracy skupiają się na bardziej szczegółowej problematyce. Byłoby w związku z tym lepiej, gdyby tytuł odzwierciedlał bardziej konkretnie najważniejsze oryginalne zagadnienia badawcze podjęte w pracy, porównywanie wyników różnych algorytmów bi-klasteryzacji oraz rozwijanie metod „konsensusowych" bi-klasteryzacji.

Tezy pracy

W ocenianej pracy zostały sformułowane dwie oryginalne tezy badawcze (str. 12). Pierwsza dotyczy opracowanej przez doktoranta metody porównywania wyników bi-klasteryzacji, bazującej na uogólnionym (przez doktoranta) algorytmie Munkresa. Druga teza dotyczy opracowanego przez doktoranta algorytmu „konsensusowego", służącego do łączenia ze sobą wyników różnych algorytmów bi-klasteryzacji.

Sformułowane przez doktoranta tezy są oryginalne. Związane są ponadto z oryginalnymi badaniami naukowymi i publikacjami doktoranta. W aspekcie sformułowanych tez badawczych należy stwierdzić, że doktorant bardzo dobrze ukierunkował swoje badania naukowe podsumowane rozprawą doktorską. Istnieje obecnie bardzo dużo różnych podejść do rozwiązywania problemu bi-klasteryzacji opisywanych w literaturze. Natomiast metody porównywania wyników algorytmów bi-klasteryzacji opisywane w literaturze do tej skupiały się głównie na metodach oceny budowanych dla sztucznie generowanych danych, gdy znane były prawdziwe struktury bi-klastrów. Dla danych, w których prawdziwa struktura bi-klastrów nie jest znana istniejące metody nie mogą być zastosowane. Doktorant dobrze zidentyfikował problem naukowy wynikający z tej sytuacji i udało mu się sformułować w konsekwencji oryginalną tezę badawczą. Kontynuując i rozwijając metody porównywania wyników bi-klasteryzacji doktorant stworzył nowy algorytm do agregacji wyników różnych metod bi-klasteryzacji i wykazanie jego użyteczności sformułował jako drugą tezę swojego doktoratu.

Zawartość rozprawy

Rozdział pierwszy pracy jest pomyślany jako wstęp. Jest bardzo zwięzły. Jego pierwsze paragrafy podają ogólną motywację do podejmowania zadania bi-klasteryzacji, natomiast ostatni paragraf jest bardzo dobrze sformułowanym wykazem/przeglądem najważniejszych wyników naukowych w zakresie bi-klasteryzacji. Rozdział jest ilustrowany dwoma rysunkami, których jakość niestety pozostawia wiele do życzenia. Brak opisów osi na rysunku 1 (lub ewentualnie w tekście nawiązującym do tego rysunku) i zbyt małe litery na rysunku 2.

Rozdział drugi zawiera omówienie celów pracy. Cele pracy (str. 11) są sformułowane w sposób logiczny, stanowią plan uzyskania nowych wyników porównywania i agregowania w zadaniach bi-klasteryzacji poprzedzony przeglądem algorytmów bi-klasteryzacji. Są one dodatkowo zilustrowane rysunkami (3 i 4). Rysunki te zawierają schematy blokowe procedur stosowanych jako składowe algorytmów bi-klasteryzacji. Niestety jakość rysunków w tym rozdziale znów jest niezadowalająca, napisy na rysunku 3 są bardzo słabo czytelne.

Rozdział trzeci to dwie tezy pracy, już wcześniej omawiane w niniejszej recenzji.

Rozdział czwarty jest charakterystyką oryginalnych elementów rozprawy. W rozdziale tym wyniki przedstawione dalej w pracy są odniesione do publikacji doktoranta.

W rozdziale piątym zostało sformułowane zadanie bi-klasteryzacji. Za odpowiednimi pozycjami literaturowymi podane są definicje kilku rodzajów bi-klastrów w dwu wymiarowych danych obserwacyjnych oraz dokonany jest przegląd wskaźników jakości bi-klasteryzacji. Wskaźniki jakości są

omówione na bazie publikacji z literatury. Przez wskaźnik jakości bi-klasteryzacji rozumie się wskaźnik, który mierzy lub ocenia, w jakiejś przestrzeni, odległość pomiędzy ocenianą strukturą bi-klastrów w danych a prawdziwą strukturą. Dla uzyskania niektórych z przedstawianych wskaźników (MSR) konieczna jest znajomość prawdziwej struktury bi-klastrów. Inne wskaźniki (ACV, ASR) mogą być wyliczone bez znajomości prawdziwej struktury. Na bazie omawianych wskaźników, w ostatnim punkcie rozdziału omawia się reguły zatrzymania iteracji w algorytmach poszukiwania struktury bi-klastrów w danych.

W rozdziale szóstym autor przedstawia przeanalizowane i zaimplementowane przez siebie w ramach pracy algorytmy bi-klasteryzacji danych, algorytmy bazujące na „nieujemnej faktoryzacji", macierzy danych (NMF), algorytm analizy semantycznej z ukrytymi zmiennymi (PLSA), algorytm bazujący na kompozycji (mieszaninie) analizatorów czynnikowych (FABIA), algorytm bazujący na przeszukiwaniu grafu dwudzielczego (QUBIC), prosty algorytm (CWTC) sekwencyjnego łączenia wierszy i kolumn, algorytm blokowy Hartigana, algorytm usuwania kolumn i wierszy Chenga i Churcha, algorytm heurystycznego przeszukiwania grafu (SAMBA) oraz algorytm „plaid" Lazzeroniego i Owena.

W rozdziale siódmym autor przedstawia teorię porównywania wyników uzyskiwanych przez różne algorytmy bi-klasteryzacji. Widać od razu, że porównywanie wyników uzyskiwanych przez różne algorytmy bi-klasteryzacji jest zadaniem nietrywialnym. Wynikiem każdego algorytmu jest wiele bi-klastrów. Każdy z nich ma złożona strukturę. Różne algorytmy bi-klasteryzacji mogą dać w wyniku różne liczby bi-klastrów. Aby porównać wyniki różnych algorytmów należy poszukiwać odpowiedniości pomiędzy uzyskanymi zbiorami w odniesieniu do danych. Autor najpierw wprowadza/omawia (dobrze znany) wskaźnik Jaccarda jako miarę podobieństwa pomiędzy dwoma bi-klastrami. Wprowadza także pojęcie istotności(dokładności)/odzysku(kompletności) (relevance/recovery) dla odniesienia wyniku bi-klasteryzacji do danych. Za publikacją Hochreitera i współautorów wprowadza także indeks „konsensusowy", który jest rozwinięciem wskaźnika Jaccarda. W drugim punkcie rozdziału autor omawia dobrze znany algorytm Munkresa dopasowania pomiędzy dwoma zbiorami (wektorami) elementów, których podobieństwo „parami" można zmierzyć. Stwierdza (za literaturą) możliwość rozwiązania zadania dopasowania metodą programowania całkowitoliczbowego (zero-jedynkowego) oraz wykazuje dużą przewagę algorytmu Munkresa nad metodą programowania całkowitoliczbowego. Punkt trzeci rozdziału siódmego jest oryginalnym osiągnięciem doktoranta. Autor formułuje uogólniony problem dopasowania, który dotyczy wielu zbiorów (wektorów) elementów, omawia znane algorytmy rozwiązania tego problemu (publikowane w latach 1968-2002) oraz przedstawia własny, dość prosty algorytm rozwiązania uogólnionego problemu dopasowania. Algorytm jest przedstawiony na stronach 53-57, ma dość prostą i jasno przedstawioną ideę. Mankamentem przedstawionych wyników jest tylko wymienienie prac literaturowych dotyczących uogólnionego problemu dopasowania. Bardzo przydatne i rozsądne byłoby znacznie dokładniejsze porównanie własnego algorytmu z literaturą.

W czwartym podpunkcie rozdziału siódmego autor przedstawia drugi, ze swych oryginalnych wyników, zastosowanie opracowanego algorytmu rozwiązywania uogólnionego problemu dopasowania do stworzenia „konsensusowego" algorytmu bi-klasteryzacji. Opracowany algorytm jest opisany na stronach 61-63.

W rozdziale ósmym doktorant przedstawia najpierw znane z literatury algorytmy wizualizacji struktury bi-klastrów w danych. Przedstawione algorytmy zilustrowane są rysunkami 25, 26, 27. Zostały one wykonane, dla przykładowych danych, z zastosowaniem oprogramowania pochodzącego z cytowanych prac. Przedstawione rysunki są jednak zbyt mało konkretnie opisane w tekście. Nie wiadomo ile bi-klastrów miał analizowany zbiór danych, jak były rozmieszczone, jak mają się do nich struktury przedstawione na rysunkach. W drugim punkcie rozdziału ósmego przedstawia się metody

oceny składu genów w sygnaturach genowych z zastosowaniem terminów ontologii genowych. Metody te mają związek z tematyką rozprawy o tyle, że mogą być użyte jako „pośrednia" droga do oceny jakości bi-klasteryzacji w danych mikromacierzowych. Ocena może polegać na testowaniu hipotez o nadreprezentacji terminów ontologii w sygnaturach genowych wynikających z uzyskanej struktury bi-klastrów. W punkcie trzecim przedstawiona jest metoda wizualizacji graficznej podobieństwa pomiędzy wynikami bi-klasteryzacji przez uogólnione diagramy Venna (za źródłem literaturowym). Mimo związków tematycznych pomiędzy trzema punktami rozdziału ósmego, jego konstrukcja robi jednak niespójne wrażenie. Byłoby lepiej jakoś najpierw usystematyzować metody prezentacji wyników bi-klasteryzacji a potem je dokładniej przedstawić. Każda z metod charakteryzuje się innym podejściem, odniesieniem do zbioru danych, do związków (przekryć) pomiędzy bi-klastrami, lub do pewnych interpretacji danych (przez ontologie). Prawdopodobnie można także zaproponować inne podejścia, zwłaszcza jeśli chodzi o interpretacje danych.

Rozdział dziewiąty poświęcony jest omówieniu wyników obliczeniowych. Rozdział rozpoczyna się przedstawieniem środowiska obliczeniowego opracowanego przez doktoranta do bi-klasteryzacji danych i porównywania wyników różnych algorytmów bi-klasteryzacji. Wydaje się, że opracowane oprogramowanie jest interesującym osiągnięciem doktoranta. Powinno być zatem szerzej omówione we wstępie do pracy. Opracowane oprogramowanie jest dość kompletne, jest wyposażone w bazę danych (symulowanych lub rzeczywistych), zawiera zaimplementowane przez doktoranta algorytmy bi-klasteryzacji, umożliwia wykorzystanie zasobów obliczeń równoległych o ile istnieją w systemie komputerowym, umożliwia także przedstawienie wyników obliczeń w sposób zarówno liczbowy jak i graficzny. W punkcie czwartym rozdziału dziewiątego autor przedstawia wyniki porównania opracowanego przez siebie algorytmu „konsensusowego" łączenia różnych algorytmów z wynikami zastosowania wszystkich zaimplementowanych w systemie algorytmów. Wyniki te przedstawione są dla danych rzeczywistych w punkcie 9.4.3 oraz dla danych symulowanych w bardzo obszernym dodatku. Z przedstawionych wyników widać prawie zawsze poprawę uzyskaną przez zastosowanie nowej metody.

W rozdziale dziesiątym autor przedstawia podsumowanie całej pracy.

Ocena pracy

Mocne strony pracy

Mocne strony pracy to przede wszystkim dobrze wybrane hipotezy badawcze oraz opracowanie dwóch oryginalnych algorytmów związanych z problematyką bi-klasteryzacji. Opracowane algorytmy dotyczą dobrze sformułowanych problemów naukowych. Pierwszy z nich pozwala na dokonanie porównania wyników bi-klasteryzacji w sytuacji, gdy prawdziwa struktura bi-klastrów nie jest znana. Rozszerza to dotychczas publikowane wyniki. Drugi algorytm pozwala łączyć ze sobą różne algorytmy bi-klasteryzacji. W stosunku do publikowanych rezultatów algorytm ten także stanowi nowe osiągnięcie.

W pracy przez wszechstronne badania obliczeniowe wykazano przydatność opracowanych nowych algorytmów zarówno dla danych symulacyjnych jak i dla danych rzeczywistych. Dane rzeczywiste pochodziły z różnych zakresów tematycznych, z problematyki analizy tekstu, a także z dziedziny analiz statystycznych danych mikromacierzowych. Dla obu tych zbiorów danych autorowi udało się wykazać przydatność swojej metodologii obliczeniowej. Stosował przy tym techniki oceny jakości wyników adekwatne do specyfiki danych, analizy leksykalne dla danych tekstowych oraz analizy reprezentacji terminów ontologicznych w wynikach bi-klasteryzacji danych mikromacierzowych. Dane symulacyjne są bardzo obszerne, a także wykazują przydatność metod opracowanych przez doktoranta.

Przeprowadzone obliczenia bazowały na opracowanym przez doktoranta wszechstronnym środowisku obliczeniowym dla zadań bi-klasteryzacji. Opracowanie tego środowiska obliczeniowego jest interesującym osiągnięciem doktoranta.

Mocną stroną rozprawy jest także odniesienie przedstawionych wyników do wystarczająco szerokiego wykazu cytowanej literatury.

Słabe strony rozprawy

Elementy krytyczne były już wymienione przy omawianiu zawartości rozprawy. Tu są częściowo podsumowane i dodane są także pewne uwagi krytyczne o ogólnym charakterze.

Praca jest krótka, a jej zakres tematyczny jest dość szeroki. Powoduje to, że część zagadnień jest omówionych tylko szkicowo. Przykłady: Algorytmy w rozdziale 6 są przedstawione tylko przez listy wzorów. Brakuje szerszych odniesień do tych formuł. Wyniki umieszone w dodatku są tylko zbiorem tabel, brakuje do nich szerszych komentarzy. Punkt 9.2 zawiera tylko listę programów, brak jakiegokolwiek ich omówienia.

Mimo odniesienia do wielu prac z literatury często niestety brakuje dokładniejszych porównań pomiędzy opracowanymi rozwiązaniami a opublikowanymi metodami. Najbardziej istotny z nich jest brak porównania opracowanego algorytmu wielowymiarowego dopasowania (rozdział 7) do wcześniej opublikowanych prac.

Grafika ilustrująca wyniki uzyskane w rozprawie jest dość niejednorodna. Niektóre z rysunków mają poważne usterki. Praca zawiera usterki językowe i edycyjne.

Podsumowanie oceny rozprawy

Podsumowując całość rozprawy stwierdzam, że Autor wykazał się bardzo dobrym opanowaniem warsztatu badawczego zarówno w zakresie badań podstawowych, jak praktycznej implementacji i symulacyjnym badaniu algorytmów.

Ocena końcowa

Stwierdzam, że w recenzowanej rozprawie został poprawnie sformułowany, a następnie również poprawnie rozwiązany z zastosowaniem metod naukowych, trudny problem bi-klasteryzacji. Oceniana rozprawa doktorska spełnia wymagania, jakie Ustawa o Stopniach i o Tytule Naukowym przewiduje dla rozpraw doktorskich. Wobec powyższego wnioskuję o jej przyjęcie jako rozprawy doktorskiej i dopuszczenie do dalszych etapów przewodu doktorskiego.

K Wojciechowski