Wydział Automatyki, Elektroniki i Informatyki
Katedra Inżynierii i Analizy Eksploracyjnej Danych

Prof. dr hab. inż.
Joanna Polańska
Kierownik Katedry

Prof. dr hab. inż. Joanna Polańska
Katedra Inżynierii i Analizy Eksploracyjnej Danych
Politechniki Śląskiej

Recenzja rozprawy doktorskiej

Autor: **Mgr Grzegorz Skoraczyński**

Tytuł: **Algorytmy i modele obliczeniowe w analizie chemicznej**

Promotor: prof. dr hab. Błażej Miasojedow

Ogólna charakterystyka rozprawy

Przedłożona do recenzji rozprawa doktorska jest napisana w języku angielskim, liczy 94 strony, obejmuje 6 rozdziałów zgrupowanych dodatkowo w 3 części. Do treści dodano listę wszystkich rysunków, listę tabel oraz wykaz skrótów. Spis literatury obejmuje 173 pozycje. Do pracy dołączone są dwa obszerne streszczenia, w języku angielskim oraz w języku polskim.

Obszar badawczy pracy obejmuje tematy związane z zastosowaniem metod modelowania matematycznego oraz technik obliczeniowych do wspomagania badań naukowych w chemii. W pracy przedstawiono osiągnięcia Doktoranta związane z dwoma tematami z tego obszaru badawczego. Pierwszy temat to uliniowienie widm chromatografii cieczowej ze spektrometrią mas. Drugi to algorytmiczna predykcja syntezowalności złożonych związków chemicznych z wykorzystaniem idei retrosyntezy.

Pierwsza część pracy obejmuje ogólne opisy dalej rozważanych problemów, a także listę publikacji Doktoranta związanych z pracą razem z opisem jego wkładu współautorskiego w tych publikacjach.

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice
+48 32 237 2144 / +48 601 891 245
joanna.polanska@polsl.pl

Druga część pracy poświęcona jest tematowi uliniowienia widm chromatografii cieczowej ze spektrometrią mas. Obejmuje ona rozdziały 2, 3 i 4. Rozdział drugi poświęcony jest wskaźnikom podobieństwa widm chromatograficznych oraz wprowadzeniu nowej koncepcji podobieństwa bazującej na odległości Wassersteina. W rozdziale poddaje się krytyce korelacyjny (kosinusowy) wskaźnik podobieństwa widm. Następnie przedstawia się koncepcję wskaźnika Wassersteina definiowanego jako minimalny koszt transportu jonów pomiędzy dwoma porównywanymi widmami. Minimalny koszt transportu (czyli odległość Wassersteina) można policzyć rozwiązując problem programowania liniowego, gdzie minimalizowany wskaźnik jest definiowany jako suma kosztów przesunięć jonów a ograniczenia wynikają z rozkładów jonów w dwóch porównywanych widmach. W podrozdziale 2.2 Doktorant opisuje zastosowanie odległości Wassersteina do rozwiązywania problemu dekonwolucji widm chromatograficznych, to znaczy do oceny proporcji występowania składników chemicznych w mieszaninie na podstawie widma chromatograficznego tej mieszaniny. W tym aspekcie cytowane są publikacje, w których Doktorant jest współautorem, opisujące algorytm „masserstein" dekonwolucji składowych mieszanin. Konstrukcja tego algorytmu bazuje na zastosowaniu odległości Wassersteina oraz na sformułowaniu problemu minimalizacji tej odległości, podług wektora proporcji składowych mieszaniny, jako problemu programowania liniowego. W podrozdziale 2.3 Doktorant opisuje technikę regularyzacji problemu programowania liniowego, związanego z minimalizacją odległości Wassersteina, zrealizowaną przez dodanie do niego dodatkowego składnika kary (entropic penalty). Przez to uzyskuje się poprawę tempa zbieżności algorytmu minimalizacji. Dalej, na podstawie analizy funkcjonowania algorytmu dla rzeczywistych danych Doktorant stwierdza, że wyniki uliniowienia algorytmem Wasserstersteina są wrażliwe na pojawiający się w widmach szum. Mają tendencję do przypisywania cech wynikających z szumu do sygnału. Aby usunąć lub ograniczyć tę wadę proponuje, za cytowanymi pozycjami literaturowymi kolejną modyfikację funkcji odległości polegającą na dodaniu do niej dodatkowego składnika uniemożliwiającego transportowanie zbyt odległych od siebie pików pomiędzy widmami. Doktorant dyskutuje różne możliwości konstrukcji, jako dobre rozwiązanie przyjmuje modyfikację bazującą na odległości Kullbacka – Leiblera.

W rozdziale 3 pracy Doktorant opisuje opracowany przy swoim współudziale, opublikowany algorytm „alignstein" pozwalający na uliniowianie widm chromatografii cieczowej. Algorytm ten bazuje na definicji odległości Wassersteina, z modyfikacjami opisanymi w rozdziale 2. Jednak dodatkowo, konstrukcja algorytmu „alignstein" jest wzbogacona o heurystyki z dodatkowymi elementami i funkcjonalnościami, które powodują, że pozwala lepiej uliniawiać / dopasowywać chromatogramy. Idee konstrukcji algorytmu „alignstein" przedstawione są na rysunku 3.2. Obejmuje ona trzy główne bloki, preprocessing, grupowanie w celu wyznaczenia cech oraz uliniowianie / dopasowanie cech widmowych. Funkcjonalność bloku preprocesingu bazuje na wykorzystaniu publicznie dostępnego środowiska openMS, sprawdzonego i skutecznego w detekcji cech widmowych. Funkcjonalność bloku grupowania uzyskiwana jest przez

HR EXCELLENCE IN RESEARCH
Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice
+48 32 237 1775 / +48 601 891 245
joanna.polanska@polsl.pl

zastosowanie algorytmów hierarchicznych oraz bazujących na ideach k–średnich. Wreszcie blok uliniowienie / dopasowanie wykorzystuje znany z literatury algorytm bilansowania przepływu w grafach bazujący znów na metodzie programowania liniowego. Jedną z zalet tak zbudowanego algorytmu „alignstein" jest możliwość dopasowania do siebie przestawionych (zamienionych) kolejnością pików widmowych, tak jak to przedstawiono na rysunku 3.1. Funkcjonalność ta jest uzyskana dzięki bogatszej strukturze opisu widm, gdzie pik widmowy nie jest jedynie opisany przez swoje położenie na osi RT, ale dodatkowo charakteryzuje się cechami uzyskanymi w drugim bloku algorytmu.

W rozdziale 4 przedstawiono wyniki dotyczące oceny skuteczności algorytmu „alignstein" oraz jego porównania z innymi narzędziami uliniawiania widm chromatograficznych. Do oceny jakości opracowanego algorytmu wykorzystano publicznie dostępne benchmarkowe dane proteomiczne oraz metabolomiczne, dla których znane były prawdziwe składy analizowanych substancji. Do porównań użyto siedem innych opublikowanych algorytmów uliniawiania. Porównania skuteczności wszystkich tych narzędzi przedstawiono w tabelach 4.1 oraz 4.2, użyto przy tym 3 wskaźników jakości dopasowania (P – precision, R – recall, F – F-score). Przedstawione wyniki wykazują konkurencyjność opracowanego algorytmu w stosunku do już dostępnych aplikacji literaturowych. W podpunkcie 4.2 opisano też wyniki zastosowania algorytmu „alignstein" do wyników analiz proteomicznych związanych z badaniem wrażliwości małży morskich (Mytilus galloprovincialis) na trujące substancje fulren (C60) oraz benzopiren (BaP). Wreszcie w podpunkcie 4.3 przedstawiono obliczeniowe eksperymenty weryfikujące zdolność algorytmu „alignstein" do realizowania dopasowywania pików które mają zmienioną kolejność w porównywanych widmach. Eksperymenty te bazują na sztucznie utworzonych przez doktoranta danych, w których dokonano „ręcznego" przetasowania kolejność pików w chromatogramach. W ostatnim podpunkcie rozdziału, 4.4 Doktorant dokonuje podsumowania zalet i wad programu „alignstein". Jako główne zalety wymienia dokładność i odporność na zakłócenia. Jako główną wadę programu „alignstein" Doktorant przedstawia jego ograniczenia w wykrywaniu dopasowania bardzo odległych od siebie pików.

Trzecia część pracy, dotycząca algorytmicznej predykcji syntezowalności złożonych związków chemicznych obejmuje rozdziały 5 i 6. Rozdział 5 obejmuje opisy kilku koncepcji tworzenia funkcji pozwalających przewidywać syntezowalność związków chemicznych. Pierwsza z proponowanych koncepcji polega na zamodelowaniu rozkładu prawdopodobieństwa liczby motywów w złożonych cząsteczkach (dostępnych w bazie danych) w postaci formuły 5.1 i ocenie parametrów tego modelu. We wzorze 5.1 występuje kwadratowa (symetryczna) macierz współczynników interakcji (theta). Doktorant zakłada, że zastosowanie pełnej macierzy interakcji doprowadzi do nadparametryzacji i ogranicza ich liczbę przez zastosowanie znanej w literaturze funkcji kary „lasso". Dla zrealizowania powyższych koncepcji Doktorant wykorzystuje dość złożoną konstrukcję kroków optymalizacji stochastycznej, metodę gradientu proksymalnego, algorytm Metropolisa – Hastingsa oraz próbkowanie Gibbsa. Uzyskany model matematyczny wykazuje dużą wrażliwość na parametr funkcji kary „lasso", a także ma inne wady

związane z faktem, że istotne parametry interakcji mogą być przesłaniane przez inne o wysokich częstościach.

W związku z wadami powyżej opisanego podejścia, w podpunkcie 5.2 Doktorant proponuje inne rozwiązanie, polegające na użyciu bogatszego zestawu cech opisujących kompozycję motywów (zawartość motywów w związku, masa cząsteczkowa, deskryptor przestrzennej struktury związku), a także zastosowanie techniki nadzorowanego uczenia do rozwiązania zadania klasyfikacji. Opisuje także metodę zbudowania zbioru przykładów pozytywnych i negatywnych dla uczenia (rysunek 5.6) oraz proces uczenia. Tak zbudowany algorytm Doktorant nazywa skrótem MF-Score i porównuje go z dwoma algorytmami literaturowymi SAscore oraz SCscore. Na rysunku 5.8 demonstruje istnienie korelacji pomiędzy wskaźnikami (wynikami algorytmów).

W następnym podpunkcie, 5.3, Doktorant proponuje kolejne ulepszenie algorytmu, polegające na zastosowaniu innego (nowego) systemu kodowania cząstek, ulepszenia konstrukcji przykładów negatywnych w powiązaniu z użyciem metody klasyfikacji wektorów podpierających. Uzyskany algorytm prowadzi do wyliczania wskaźnika nazwanego przez Doktoranta wskaźnikiem QC-MF-Score.

W rozdziale 6 Doktorant przedstawia interesujący projekt porównania kilku systemów przewidywania syntezowalności związków chemicznych pochodzących z literatury z opracowanymi przez siebie wskaźnikami MF-Score oraz QC-MF-Score. Jako dane odniesienia Doktorant generuje związki z wykorzystaniem programu AiZynthFinder. Pomimo faktu, że zaproponowane przez Doktoranta systemy MF-Score oraz QC-MF-Score uzyskują słabe wyniki w porównaniach (rysunek 6.3) rozdział ten prezentuje wiele interesujących wniosków i wyników dotyczących możliwości przewidywania syntezowalności związków oraz projektowania technik / ścieżek syntezy.

Ocena pracy

Moja ocena pracy jest zdecydowanie pozytywna. Doktorant bardzo dobrze porusza się w szerokim i różnorodnym obszarze modelowania matematycznego, stosowania bardzo różnorodnych metod matematycznych i statystycznych. Wykazuje się kompetencją i bardzo dobrym oczytaniem.

Obok warsztatu modelowania matematycznego, Doktorant wykazuje się także wiedzą w zakresie technik chromatografii, a także w zakresie elementów syntezy chemicznej.

Zaletą pracy jest dyskutowanie różnych możliwych wariantów rozwiązywania stawianych problemów i konsekwencji ich zastosowania. Ma to bardzo dobry wpływ na logiczną konstrukcję pracy.

HR EXCELLENCE IN RESEARCH
Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice
+48 32 237 1775 / +48 601 891 245
joanna.polanska@polsl.pl

Wyniki opisane w pracy były publikowane jako artykuły w bardzo poważnych czasopismach naukowych. Zgodnie z opisami współudziału, Doktorant jako współautor wniósł istotny wkład w opracowanie tych publikacji.

Z opisu udziału współautorskiego, a także z treści pracy wynika, że Doktorant posiada także dobry warsztat programistyczny. Jest autorem oprogramowania, w środowisku Python realizujących złożone algorytmy wykorzystywane w ramach pracy.

Uwagi krytyczne i dyskusyjne

1. Sformułowanie „in chemical analysis" w tytule pracy wydaje się trochę niefortunne. Analiza chemiczna (chemical analysis) oznacza zwykle techniki eksperymentalne badania składu chemicznego substancji. W pracy jest to rozumiane trochę inaczej.
2. Polskie streszczenie jest bardzo złej jakości. Angielskie zdania zostały przełożone mechanicznie, być może z użyciem jakiegoś programu tłumaczącego.
3. W rozdziale 2 wprowadza się dodatkową funkcję kary w postaci entropii aby uzyskać szybszą zbieżność i lepszą stabilność numeryczną. Dobrze byłoby zdefiniować te pojęcia, a także odnieść się do rozmiarów danych chromatogramów w analizowanych przykładach. Nie jest jasne w jakim sensie rozwiązania uzyskane metodą programowania liniowego bez modyfikacji (2.2) byłyby gorsze od rozwiązań uzyskanych z użyciem dodatkowej funkcji kary.
4. W teorii optymalizacji, dla problemu programowania liniowego dobrze znane są metody punktu wewnętrznego oraz algorytm Karmakara. W aspekcie regularyzacji proponowanych i stosowanych w pracy dobrze byłoby wspomnieć o tych podejściach.
5. Na rysunku 5.1 błędnie opisana jest oś pozioma. Powinno być: „Number of motifs in molecules + 1".
6. Modelowanie rozkładów brzegowych na rysunku 5.1 przez rozkład Dzeta wydaje się być dość rozsądną koncepcją umożliwiającą uniknięcie wprowadzania nowych parametrów. Jednak powinno się motywację do tego podejścia chyba lepiej dodatkowo uzasadnić. Przy parametrach widocznych na rysunku 5.1, dla każdego ze związków test chi kwadrat najprawdopodobniej odrzuciłby hipotezę o zgodności obserwowanego rozkładu z rozkładem Dzeta.
7. Sformułowanie, że wzór (5.1) jest instancją pola Markowa wymagałoby chyba udowodnienia, że zachodzi / zachodzą własność / własności Markowa dla podzbiorów węzłów grafu.
8. Być może interesujące byłoby modelować rozkład częstości motywów w podpunkcie 5.1 stosując model mieszaniny rozkładów wielomianowych.

HR EXCELLENCE IN RESEARCH
Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice
+48 32 237 1775 / +48 601 891 245
joanna.polanska@polsl.pl

Konkluzja

Uwagi krytyczne i dyskusyjne nie podważają ogólnej pozytywnej oceny wartości pracy. Osiągnięcia i oryginalne elementy publikacji wchodzących w skład rozprawy, a także jakość całego tekstu rozprawy są na pewno wystarczające do jej ogólnej pozytywnej oceny. Uzyskane rezultaty stanowią oryginalny wkład własny Doktoranta w rozwój dyscypliny naukowej Informatyka. Uzyskane wyniki świadczą, że Pan mgr Grzegorz Skoraczyński wykazał się dogłębną znajomością najnowszych metod i algorytmów chemoinformatyki, jak również dobrym opanowaniem warsztatu badawczego i dojrzałością naukową. Stwierdzam, że rozprawa spełnia odpowiednie warunki określone w obowiązującej ustawie i wnioskuję o jej dopuszczenie do publicznej obrony.

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice
+48 32 237 1775 / +48 601 891 245
joanna.polanska@polsl.pl