prof. dr hab. n. farm. Łukasz Komsta
Uniwersytet Medyczny w Lublinie
Wydział Farmaceutyczny z Oddziałem Analityki Medycznej
Katedra i Zakład Chemii Leków
ul. Jaczewskiego 4, 20-090 Lublin, tel. 81 4487387, fax 81 4487381

Recenzja pracy doktorskiej mgr Klaudii Drab

„Metody grupowania danych i ich wybrane modyfikacje dedykowane eksploracji danych eksperymentalnych"

Metody chemometryczne mają swoje stałe miejce w analizie chemicznej, a rozwój komputerów i oprogramowania czyni je dostępnymi dla przeciętnego analityka. Znajdują zastosowanie na każdym etapie projektu badawczego: od planowania eksperymentu po pomoc w interpretacji wyników i wyciąganiu wniosków.

Techniki grupowania danych znajdują zastosowanie w analizie danych bez nadzoru i umożliwiają automatyczną klasyfikację (podział) obiektów na grupy wraz z interpretacją czynników odpowiedzialnych za przydział do konkretnej grupy. Ich rozwój jest procesem ciągłym, a przedłożona do oceny dysertacja doktorska stanowi tego doskonały przykład. Została wykonana w Instytucie Chemii Wydziału Matematyki, Fizyki i Chemii Uniwersytetu Śląskiego pod kierunkiem prof. dra hab. Michała Daszykowskiego, w zespole naukowym mającym swoje stałe miejsce na arenie światowej chemometrii.

Licząca 146 stron praca rozpoczyna się od streszczenia, zestawienia użytej notacji matematycznej oraz wykazu skrótów i akronimów. Zwięzłe wprowadzenie osadza czytelnika w tematyce badawczej przed poznaniem celu pracy, który Autorka sformułowała w siedmiu punktach. Do najważniejszych wątków badawczych zaplanowanych przez Doktorantkę należy zaliczyć propozycje modyfikacji istniejących algorytmów grupowania danych, opracowanie nowych miar podobieństwa do oceny porównawczej dwuwymiarowych danych sygnałowych oraz implementację takiej modyfikacji istniejących algorytmów grupowania danych, aby można było uwzględnić znaną niepewność pomiarową. Już na początku recenzji pragnę zauważyć, że cel jest ambitny, a recenzowana praca proponuje nowe algorytmy chemometryczne, radzące sobie z ważnymi problemami.

Po sformułowaniu celu pracy Doktorantka wprowadza czytelnika w metody instrumentalne, opisuje podstawowe założenia detekcji spektrofotometrycznej, charakteryzuje otrzymane w ten sposób dane, przedstawia cechy metod separacyjnych z punktu widzenia zbierania danych, omawia wpływ sprzężenia kilku metod na otrzymane zestawy danych doświadczalnych oraz prezentuje ograniczenia metod instrumentalnych. Oddzielną część stanowi wprowadzenie czytelnika w strukturę danych wielowymiarowych (*multilinear*) oraz stosowane techniki wstępnej obróbki danych. Przedstawiona jest również teoria cech, jakie muszą wykazywać miary podobieństwa i odmienności stosowane w chemometrycznej ocenie wyników. W dalszej kolejności Autorka przedstawia najczęściej stosowane miary odmienności - odległość Euklidesa oraz Mahalanobisa, jak również współczynnik korelacji Pearsona.

Dwie strony pracy poświęcone są klasyfikacji metod chemometrycznych, głównie celem określenia

Podpisane cyfrowo przez Lukasz Komsta
Date: 2016.03.17 18:27:40 CET
Location: Lublin

jaki rodzaj metod stanowią te, które Doktorantka modyfikowała w ramach badań własnych. Dalej Autorka opisuje założenia metod projekcji, analizę czynników głównych oraz metody grupowania danych. Te ostatnie dzieli na metody hierarchiczne (HCA) oraz niehierarchiczne (k-średnich, gazu neuronowego, ekspandującego gazu neuronowego, ekspandującego k-średnich); kończąc na opisie metod bazujących na gęstości danych doświadczalnych (DBSCAN, OPTICS). Praca zawiera również obszerny opis metod współgrupowania danych (CC, k-spectral, regresji macierzy rzadkiej) oraz wprowadzenie do problemu selekcji zmiennych. Część teoretyczną kończy zestawienie zastosowań metod eksploracji danych.

Metoda DBSCAN jest bardzo atrakcyjnym narzędziem analizy kompleksowych danych. Jednakże jedną z jej istotnych wad jest niejednoznaczna klasyfikacja obiektów znajdujących się na granicy dwóch leżących blisko siebie skupień. Ten znany chemometrykom fakt był motywacją do opracowania przez Doktorantkę modyfikacji algorytmu, poprawiającej klasyfikację obiektów. Idea została zaimplementowana przez Autorkę oraz przetestowana na przykładowych symulowanych danych. Analiza zachowania zmodyfikowanego algorytmu potwierdziła znaczącą poprawę otrzymanego grupowania.

Drugim wątkiem podjętych badań było wprowadzenie nowej miary podobieństwa do porównywania dwuwymiarowych chromatograficznych odcisków palca. W klasycznym podejściu analizy porównawczej chromatogramy muszą być nałożone na siebie w procesie wstępnej obróbki, aby zniwelować wpływ losowych przesunięć na osi czasu. Jedyne znane podejście omijające ten etap przygotowania, bazujące na tzw. macierzy Grama, jest rzadko stosowane ze względu na utratę sporej ilości informacji podczas obliczania tej macierzy. Pomysł Doktorantki jest nowatorski ze względu na analizę korelacji spektralnej, a nie korelacji sygnałów w funkcji czasu. Zaproponowany współczynnik, nazwany s_{ij}, pozwala na kompleksowe porównanie dwuwymiarowych sygnałów bez względu na ewentualne przesunięcia względem siebie. Dodatkowe badania na danych symulowanych potwierdzają możliwość prostego wykrycia nierozdzielonych pików, a ten problem jest również cały czas aktualny w chemometrii. Uzupełnieniem jest zastosowanie tej miary do porównywania chromatogramów próbek leku Viagra®(autentycznego i sfałszowanego) oraz segmentacji obrazów hiperspektralnych zarejestrowanych w zakresie światła widzialnego.

Kolejnym wątkiem pracy jest zastosowanie metod współgrupowania danych do oceny typowych problemów chemicznych. Metody te są rzadko stosowane w chemometrii i ta część pracy, mimo iż dotycząca istniejących algorytmów, jest wartościowa od strony aplikacyjnej. W tej części Autorka wykorzystała dane próbek oliwy z oliwek oraz opium.

Ostatnią i ważną częścią przeprowadzonych badań była implementacja metod grupowania, w których wykorzystuje się znane wartości niepewności pomiarowej. Ze względu na to, iż niepewność pomiarowa danych wielorakich jest również zmienną wieloraką, a jej rozkład jest skomplikowany i charakteryzowany przez całą macierz wariancji-kowariancji, uwzględnienie tej niepewności w grupowaniu może znacznie poprawić otrzymane wyniki. Autorka skupiła się na implementacji literaturowej propozycji takiego podejścia, dołączając do pracy kod źródłowy swojej implementacji w środowisku MATLAB.

Bibliografia dysertacji liczy 122 bardzo starannie wyselekcjonowane i wartościowe pozycje, poprawnie umiejscawiające badania w kontekście aktualnych trendów chemometrycznych.

Przygotowanie tabletki do analizy chromatograficznej opiera się najczęściej na prostej ekstrakcji np. metanolem. Taki proces w większości przypadków pozwala na oddzielenie substancji aktywnej od wszelkich substancji pomocniczych. W związku z tym chromatogram leku Viagra powinien za-

wierać wyłącznie jeden pik sildenafilu i ewentualnie piki zanieczyszczeń lub produktów rozkładu. Jeśli założymy, że fałszowanie leku nie polega na braku tej substancji (czyli pik sildenafilu pozostaje prawie bez zmian), to różnica może być zlokalizowana tylko w składzie substancji pomocniczych, które w ogóle nie są widoczne na chromatogramie. Większość opublikowanych metod rozpoznawania sfałszowanych leków z sildenafilem bazuje na analizie spektralnej całej tabletki, np. NIR, Ramana czy też dyfraktometrii rentgenowskiej. Ten fakt nasuwa mi jako recenzentowi pytanie, czym dokładnie różniły się chromatogramy sfałszowanych leków i jakie było przygotowanie próbki do chromatografii.

Dodatkowo z obowiązku recenzenta chciałbym przedstawić kilka drobnych uwag wynikających głównie z pomyłek redakcyjnych:

1. Gdy elektron przechodzi na orbital o wyższej energii, orbital ten wcale nie musi być pusty. Ponadto w większości przypadków powrót do stanu podstawowego związany jest z wydzieleniem ciepła, a nie promieniowania (str. 14).
2. Odległość Euklidesa nie zmienia się tylko w przypadku transformacji afinicznej (str. 32 i 120). Każda inna transformacja (np. logarytmiczna) będzie zmieniać tę odległość.
3. Według Doktorantki (str. 33) punkty oddalone od środka o stałą wartość odległości Mahalanobisa tworzą w przestrzeni hipersferę, a w przestrzeni dwuwymiarowej elipsę. Moim zdaniem w pełnej przestrzeni nie jest to hipersfera, lecz hiperelipsoida, ponieważ odległość Mahalanobisa jest odległością ważoną z uwzględnieniem wariancji i kowariancji.
4. Współczynnik korelacji notuje się małą literą r (str. 34). Doktorantka słusznie zwraca uwagę na wrażliwość tego współczynnika na obserwacje odstające, jednak istnieje kilka modyfikacji czyniących ten współczynnik stabilnym (*robust*).
5. Ortogonalność wektorów, na które dokonuje się projekcji, nie ma nic wspólnego z maksymalizacją wariancji danych (str. 38). Oryginalny układ współrzędnych też jest ortogonalny i można go obrócić na nieskończenie wiele sposobów bez likwidacji ortogonalności. Podobnie ortogonalność ta nie ma nic wspólnego z ewentualną utratą informacji (str. 39).
6. Dlaczego wyodrębnianie grup z danych nie wykazujących naturalnego podziału jest efektem „poszukiwania lokalnego" (str. 48)?
7. Czym różni się „warunek refleksji" od „warunku silnej refleksji" (str. 29)?
8. Co ma wspólnego sferyczny kształt grup obiektów z odległością Euklidesa w grupowaniu (str. 76)?
9. Jeśli informacja w danych ukryta jest w podprzestrzeniach (str. 67), to informację taką najlepiej znajdują metody projekcji, a nie wyboru zmiennych. Z drugiej strony metody wyboru zmiennych są też szczególnym przypadkiem projekcji na podprzestrzeń zdefiniowaną wyłącznie przez osie wybranych zmiennych.
10. Doktorantka pisze, że *„ze względu na nadanie tej samej wagi wszystkim zmiennym macierzy, autoskalowania nie stosuje się w przypadku sygnałów zawierających szum"* (str. 27). Warto byłoby rozwinąć to stwierdzenie.
11. Wzór nowej miary podobieństwa (str. 82) jest napisany w notacji środowiska MATLAB. Czytelnik oczekujący notacji matematycznej może mieć problemy z jego zrozumieniem.

Słabszą stroną pracy jest dość sztywny język. Polska terminologia analityczna jest przedmiotem nieustannych sporów i dyskusji, jednakże zdecydowanie użyłbym słowa „składnik" zamiast „kom-

ponent", rzeczywiste wartości (zamiast „realne"), wartości dodatnie (zamiast „pozytywne"), przekątna (zamiast „diagonala"). Otwartą dyskusją jest znalezienie lepszych polskich odpowiedników dla takich terminów jak „grupy bananowe", czy „grupy zawarte w sobie". Jako farmaceutę razi mnie też nieprofesjonalny termin „lekarstwo". W kilku momentach tekst jest bardzo skomplikowany (np. *„stanowiących niezbędny element ich rozwoju w rozpowszechnieniu ich zastosowania jako narzędzia poznania danych rozmaitego pochodzenia"*).

Niezależnie od tego praca ma wyraźne mocne strony, a przedstawione w niej badania stanowią bezsporny wkład do rozwoju chemometrii w skali światowej. Jestem przekonany, że zarówno zaproponowana nowa miara podobieństwa, jak i modyfikacja algorytmu DBSCAN wejdzie do grupy metod rutynowo stosowanych w analizie danych doświadczalnych i mocno wyznaczy nowe trendy w dalszym rozwoju algorytmów chemometrycznych.

Dostrzegam szerokie możliwości kontynuacji zaprezentowanych badań. Wprowadzona miara podobieństwa zakłada liniowość odpowiedzi detektora, która nie zawsze jest spełniona. Jest tu pole do dalszych obliczeń na symulowanych sygnałach o nieliniowej zależności, których celem miałoby być ustalenie rozkładu przyjmowanych wartości zaproponowanego kryterium podobieństwa. Jeśli utworzone macierze podobieństw zawierające współczynniki s_{ij} mają być analizowane bezpośrednio przez obliczanie wartości własnych (co sugeruje propozycja dalszych badań, str. 125), to będzie konieczne matematyczne wykazanie, że macierz złożona z takich współczynników jest zawsze dodatnio określona.

Podsumowując, praca doktorska mgr Klaudii Drab spełnia wszystkie wymagania stawiane pracom doktorskim, zawiera istotny wkład w rozwój algorytmów chemometrycznych w skali światowej i wnioskuję o dopuszczenie Doktorantki do dalszych etapów przewodu doktorskiego.

Lublin, 17 marca 2016.