

dr hab. TOMASZ PUZYN
profesor UG

Gdańsk, 30 marca 2016 r.

RECENZJA

rozprawy doktorskiej mgr Klaudii Drab

pt. „Metody grupowania danych i ich wybrane modyfikacje dedykowane eksploracji danych eksperymentalnych”

Przedstawiona mi do recenzji rozprawa doktorska poświęcona jest zaawansowanym technikom analizy struktury wewnętrznej danych chemicznych należącym do grupy tzw. „metod uczenia bez nadzoru”. Wykonana została w Instytucie Chemii na Wydziale Matematyki, Fizyki i Chemii Uniwersytetu Śląskiego pod kierunkiem dr. hab. Michała Daszykowskiego, prof. UŚ. Warto w tym miejscu podkreślić, że liczne osiągnięcia naukowe śląskiej szkoły chemometrii są wysoko cenione w kraju i za granicą. Dlatego z dużym zainteresowaniem przystąpiłem do lektury rozprawy.

Praca liczy 150 stron, zilustrowana jest 55 rysunkami oraz odwołuje się do 122 pozycji literaturowych. Tytuł rozprawy odpowiada zawartym w niej treściom. Zwraca natomiast uwagę niepoprawne użycie w tytule słowa „dedykowane”, stanowiącego kalkę z języka angielskiego. Praca ma układ klasyczny. Po stronie tytułowej i zwyczajowych podziękowaniach następuje streszczenie. Dalej Autorka zamieszcza bardzo przydatne objaśnienia dotyczące notacji oraz wykaz akronimów używanych w pracy. Kolejnym elementem jest spis treści ukazujący strukturę rozprawy. W jej skład wchodzi następujące rozdziały: „Wprowadzenie” (3 str.); „Cele pracy” (1 str.); „Zaawansowane metody instrumentalne” (10 str.); „Struktura danych eksperymentalnych” (3 str.); „Wstępne przygotowanie danych do dalszej analizy” (1 str.); „Określanie podobieństwa występującego w danych eksperymentalnych” (15 str.); „Klasyfikacja metod chemometrycznych” (2 str.); „Metody eksploracji danych” (5 str.); „Metody grupowania danych” (20 str.); „Metody współgrupowania danych” (9 str.); „Obszary zastosowań metod eksploracji danych” (2 str.); „Badania własne” (51 str.); „Podsumowanie” (3 str.); „Załączniki” (12 str.); „Bibliografia” (12 str.) oraz *Curriculum Vitae* Doktorantki (3 str.). Struktura rozprawy jest poprawna, natomiast - zdaniem recenzenta - znacznie zyskałaby na przejrzystości, gdyby Autorka połączyła

względnie krótkie rozdziały od 3. do 11. w jeden dłuższy – zatytułowany „Część teoretyczna” lub podobnie.

W rozdziale zatytułowanym „Wprowadzenie” Autorka krótko i trafnie omawia znaczenie metod chemii analitycznej oraz chemometrii w rozwoju współczesnej chemii, geologii, biologii i medycyny. Istotnym problemem, który pomaga rozwiązać chemometria, jest występowanie znacznych ograniczeń związanych z interpretacją danych wielowymiarowych. Nie zgadzam się natomiast ze stwierdzeniem Doktorantki, że wynikają one „z braku możliwości wizualizacji danych o wymiarowości większej niż trzy”. Czy Autorka potrafiłaby podać przykłady powszechnie wykorzystywanych typów wykresów umożliwiających wizualizację danych analitycznych w przestrzeni więcej niż trzech cech?

W dalszej części tego rozdziału mgr Drab charakteryzuje poszczególne grupy technik chemometrycznych podając przykłady ich zastosowania w analityce chemicznej. Omawiając metody modelowania, wskazuje na ich zastosowanie między innymi w badaniach właściwości antyoksydacyjnych produktów spożywczych, czy badaniach nad autentycznością artykułów spożywczych oraz leków. Ważnym zastosowaniem metod tzw. „modelowania z nadzorem” w chemii analitycznej – pominiętym przez Autorkę rozprawy – jest ilościowe badanie zależności pomiędzy strukturą chemiczną związku a jego właściwościami, w tym parametrami retencji. Metody te, znane jako QSRR (ang. *Quantitative Structure-Retention Relationships*), są wykorzystywane w chromatografii. Podejrzewam, że Doktorantka celowo pomija metody QSRR w swojej rozprawie klasyfikując je raczej jako metody chemoinformatyczne niż chemometryczne. Kiedy pod koniec ubiegłego stulecia wprowadzony został po raz pierwszy termin „chemoinformatyka”, metody chemometryczne były już powszechnie znane i stosowane. Kwestia przebiegu granicy pomiędzy chemometrią i chemoinformatyką jest zagadnieniem dyskusyjnym i zdania na ten temat są podzielone w środowisku naukowym. Bardzo chciałbym poznać zdanie Doktorantki na ten temat w części dyskusyjnej publicznej obrony.

W kolejnym rozdziale Autorka przedstawia szczegółowe cele pracy. Dotyczą one oceny przydatności wybranych technik grupowania i współgrupowania danych analitycznych, modyfikacji istniejących algorytmów grupowania, opracowania nowej miary podobieństwa pomiędzy obiektami w przestrzeni cech oraz sposobu uwzględniania niepewności pomiarowych w eksploracji danych. Cele te są ambitne i dotyczą poszukiwania rozwiązań bardzo istotnych problemów, przed którymi stoi obecnie chemometria.

W dalszych rozdziałach Doktorantka przedstawia krótką charakterystykę poszczególnych grup metod analizy instrumentalnej. Poprawnie opisuje poszczególne metody spektroskopowe oraz metody separacyjne, a także metody sprzężone, wyczerpująco przedstawiając ich zalety i ograniczenia. W tej części pracy znalazłem tylko jedną nieścisłość. Na str. 22. Autorka pisze: „(...) każdy sygnał analityczny składa się z trzech komponentów tj. szumu, linii podstawowej oraz sygnału właściwego. Pierwsze dwa są niekorzystne i utrudniają odczyt informacji opisywanej przez otrzymywane sygnały”. Trudno zgodzić się ze stwierdzeniem, że występowanie linii bazowej w widmie jest niekorzystne. Natomiast istotnym problemem mającym wpływ na wyniki analizy widma jest przesunięcie linii bazowej i związana z tym konieczność wprowadzenia odpowiedniej korekty.

W kolejnej części pracy mgr Drab przechodzi do teoretycznego opisu metod chemometrycznych. Wysoki poziom merytoryczny tej części pracy świadczy o dobrej znajomości teoretycznych podstaw metod wykorzystywanych przez Doktorantkę. Pomimo wnikliwej lektury wynikającej z obowiązku recenzenta znalazłem tutaj jedynie niewielkie potknięcia. W równaniu 12. na str. 34 (współczynnik korelacji Pearsona) w liczniku brakuje symbolu sumowania. W rozdziale 8. – omawiając poszczególne etapy procesu analitycznego – Autorka pominęła niezwykle ważną rolę metod chemometrycznych na etapie planowania eksperymentu. Omawiając podstawy teoretyczne analizy czynników głównych (PCA) Doktorantka nie wspomniała, w jaki sposób wybierana jest liczba f istotnych czynników głównych. W opisie algorytmów hierarchicznego grupowania danych (str. 44) różnica pomiędzy tzw. metodą pojedynczych i całkowitych połączeń nie została wyjaśniona wystarczająco precyzyjnie.

Następną część rozprawy stanowi prezentacja wyników badań własnych. Jako najważniejsze osiągnięcia Doktorantki przedstawione w rozprawie wymienilibym:

- Zaproponowanie modyfikacji algorytmu DBSCAN rozwiązującego problem błędnego przypisywania obiektów brzegowych do odpowiednich grup. Zaproponowana poprawka polegająca na tym, że w nowej wersji algorytmu przypisanie obiektów do grup jest determinowane przez obiekty rdzeniowe z pominięciem obiektów brzegowych, znacznie poprawiło stabilność działania algorytmu DBSCAN.
- Opracowanie nowej miary podobieństwa pomiędzy obiektami w przestrzeni cech (oznaczonej w pracy symbolem: s_{ij}) wyprowadzonej ze współczynnika korelacji r-Pearsona.
- Zastosowanie tej miary do opracowania metodyki porównywania tzw. „dwuwymiarowych odcisków palca” w chromatografii. Korzyścią z zastosowania miary s_{ij} w tym zakresie jest znaczne uproszczenie procesu wstępnego przygotowania danych. Szczególnie ważne jest tutaj pominięcie czasochłonnego

etapu nakładania sygnałów instrumentalnych na siebie. Wykorzystanie miary podobieństwa s_{ij} umożliwia również lepszy rozdział sygnałów substancji koeluujących.

- Zademonstrowanie użyteczności miary s_{ij} w procesie analizy danych zorganizowanych w postaci trójwymiarowych tensorów, np. obrazów hiperspektralnych.
- Porównanie efektywności dostępnych algorytmów w kontekście tzw. współgrupowania danych chemicznych. Współgrupowanie danych umożliwiające identyfikację podmacierzy danych charakteryzujących się podobnymi cechami znajduje szczególnie ważne zastosowanie w tzw. „naukach -omicznych” takich jak: proteomika, metabolomika, genomika i transkryptomika.
- Zaproponowanie sposobu uwzględniania niepewności pomiarowej w procesie grupowania danych za pomocą algorytmu DBSCAN, co jest szczególnie ważne w przypadku łączenia danych pochodzących z różnych źródeł.

Wymienione osiągnięcia naukowe spełniają z nadkładem zwyczajowe i formalne wymagania stawiane rozprawom doktorskim. Ta część pracy stanowi w mojej ocenie bardzo istotny wkład do współcześnie wykorzystywanych w chemometrii metod grupowania. Cennym uzupełnieniem tekstu samej rozprawy są załączone na końcu wydruki zawierające opracowane przez Autorkę nowe funkcje i algorytmy dla środowiska MATLAB (tzw. M-pliki). Mam nadzieję, że zostaną one udostępnione publicznie umożliwiając w ten sposób szerokie wykorzystanie wyników uzyskanych przez Doktorantkę.

Z roli recenzenta wynika jednak również obowiązek wskazania słabszych fragmentów pracy oraz zagadnień, które chciałbym przedyskutować z Kandydatką podczas publicznej obrony.

Po pierwsze, zabrakło mi w rozprawie bardziej szczegółowej dyskusji uzyskanych wyników. Elementy dyskusji występują na zakończenie każdego z podrozdziałów, w których Autorka prezentuje najważniejsze wyniki. Nie zawierają one jednak szerszego odniesienia do literatury. Cennym uzupełnieniem dyskusji mogłoby być również wskazanie kierunków dalszego rozwoju oraz potencjalnie nowych zastosowań opracowywanych metod uczenia bez nadzoru.

Po drugie, testowanie zmodyfikowanej wersji algorytmu DBSCAN przeprowadzone zostało w oparciu o tylko jeden zbiór danych. Nasuwa się pytanie: Czy jakkolwiek wpływ na działanie tego algorytmu mogłoby mieć inne rozłożenie grup obiektów w przestrzeni? Na przykład, jak algorytm zachowa się w przypadku nieostrego rozdzielenia danych o strukturze dwóch grup zawartych w sobie (por. rys. 10d, na str. 43)?

Po trzecie, w tekście rozprawy nie znalazłem szczegółowych informacji, w jaki sposób autorzy wyznaczyli najbardziej optymalną kombinację tzw. „parametrów wejściowych” w metodzie DBSCAN (tzn. wartości: $\text{MinPts} = 255$ oraz $r = 0,6$)? Praktyczne wskazówki w tym zakresie byłyby istotne z punktu widzenia przyszłych użytkowników tej metody.

Po czwarte, moim zdaniem Doktorantka dość niefortunnie nazwała zdefiniowaną przez siebie nową miarę odległości w przestrzeni cech. Symbol s kojarzy się bardziej z odchyleniem standardowym lub entropią, niż z odległością. Niektóre powszechnie używane miary podobieństwa mają specyficzne nazwy, które ułatwiają ich zapamiętanie (np. odległość Tanimoto czy odległość euklidesowa). Sugeruję, żeby Autorka zaproponowała i wprowadziła do użycia adekwatną nazwę dla opracowanej przez siebie wielkości.

Po piąte, w przykładzie 7. dotyczącym porównania próbek Viagry Autorka wspomina, że przed wykonaniem analizy PCA „dane przeskalowano”. Na czym polegało to przeskalowanie? Nasunęło mi się w tym miejscu również pytanie związane z teorią procesu przygotowania danych do PCA, na które odpowiedź chciałbym usłyszeć w trakcie publicznej obrony: Autorka, opisując teorię analizy PCA (na str. 39) wspomina, że metoda ta opiera się na dekompozycji „macierzy kowariancji lub macierzy korelacji”. Czy Doktorantka wie, w jakim przypadku macierz korelacji będzie jednocześnie macierzą kowariancji? Jak będzie wpływał na to fakt autoskalowania (bądź nie) danych przed rozpoczęciem analizy?

Po szóste, na stronie 104 Autorka rozprawy nie podaje, na podstawie jakiego kryterium do dalszej analizy wybrano cztery pierwsze czynniki główne.

Po siódme, w obliczeniach (np. w przykładzie 10. i 11.) Doktorantka posługuje się różnorodnymi zbiorami danych. Jeśli dane te pochodzą z innych opracowań, należałoby zacytować źródło. Jeśli zaś są to dane niepublikowane, sugeruję w przyszłości umieszczenie oryginalnych tabel z danymi jako dodatek elektroniczny (np. na płycie CD) do pracy. Takie podejście umożliwiłoby recenzentowi dodatkowe sprawdzenie poprawności obliczeń.

Wymienione powyżej niewielkie uchybienia nie umniejszają jednak mojej bardzo wysokiej oceny merytorycznej strony pracy.

Obowiązkiem recenzenta jest również ocena formalnej strony dysertacji. Rozprawa jest napisana przejrzysto, przy użyciu poprawnego języka. Wyjątkiem jest zbyt dosłowne tłumaczenie niektórych pojęć z języka angielskiego. Przykładami – oprócz wspomnianego już „dedykowania” występującego w tytule – mogą być takie niezręczne sformułowania jak: „komponenty próbki” (str. 22) czy „grupy bananowe”

(str. 42). Tekst jest sformatowany poprawnie, co czyni pracę bardzo czytelną. Uzupełnieniem tekstu są rysunki i schematy służące wyjaśnieniu bardziej złożonych zagadnień oraz prezentacji uzyskanych wyników. Moje jedyne zastrzeżenie dotyczy wykresów wykonanych w programie MATLAB. Zbyt mała czcionka sprawia, że opisy osi oraz jednostki są mało czytelne.

Z załączonego do rozprawy wykazu dorobku naukowego dowiedziałem się, że mgr Klaudia Drab jest autorką dwóch publikacji naukowych w czasopismach z listy JCR, wygłosiła trzy komunikaty ustne oraz zaprezentowała aż czternaście plakatów na konferencjach naukowych. Biegłość w prezentowaniu wyników prac badawczych jest – oprócz prowadzenia samych badań – ważną umiejętnością pożądaną w świecie nauki i świadczy o bardzo dobrym przygotowaniu Doktorantki do pracy naukowej.

W mojej ocenie rozprawa stanowi oryginalne rozwiązanie kilku bardzo istotnych problemów naukowych oraz potwierdza ogólną wiedzę teoretyczną Kandydatki w dyscyplinie, a także umiejętność samodzielnego prowadzenia pracy naukowej. Drobne potknięcia, które wymieniłem w niniejszej recenzji nie wpływają na bardzo wysoką ocenę merytoryczną przeprowadzonych badań. Dlatego z pełnym przekonaniem stwierdzam, że rozprawa mgr Klaudii Drab spełnia wymagania określone w Art. 13. *Ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki* z dnia 14 marca 2013 r. z późniejszymi zmianami. Wnioskuje zatem o dopuszczenie Kandydatki do dalszych etapów przewodu doktorskiego.



Tomasz Puzyn