RDITT-wpł. 16.05.2023
M. Flury

Dr hab. inż. Zbigniew Świder, prof. PRz
Katedra Informatyki i Automatyki
Politechnika Rzeszowska im. Ignacego Łukasiewicza
al. Powstańców Warszawy 12
35-959 Rzeszów

Rzeszów, 9.05.2023

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy:	Data clustering with mixtures of multidimensional distributions
Autor rozprawy:	mgr Mateusz Kania
Promotor rozprawy:	prof. zw. dr hab. inż. Andrzej Polański
Dziedzina:	nauki inżynieryjno-techniczne
Dyscyplina:	informatyka techniczna i telekomunikacja

Niniejsza recenzja została przygotowana na zlecenie Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej.

1. Cel i zakres rozprawy

Rozprawa doktorska mgr Mateusza Kani dotyczy opracowania i porównania wybranych modeli i związanych z nimi algorytmów nienadzorowanego grupowania opartych na modelu i na odległości oraz ich implementacji wraz ze szczegółowymi badaniami ich wydajności na różnych zbiorach danych przy użyciu różnych metryk. Praca kładzie nacisk na wykorzystanie algorytmów opartych na modelach wykorzystujących wielowymiarowe mieszaniny rozkładów.

W rozprawie postawiono trzy główne hipotezy:

1. Nienadzorowane metody grupowania oparte na mieszaninach dystrybucji osiągają optymalną wydajność, gdy statystyki danych są zgodne z rzeczywistymi rozkładami.
2. Nienadzorowane grupowanie oparte na mieszaninach dystrybucji jest konkurencyjne w porównaniu z metodami opartymi na odległości.
3. Możliwość zastosowania grupowania opartego na mieszaninach rozkładów do praktycznych problemów polega na opracowaniu implementacji algorytmicznej wyspecjalizowanej dla dużych rozmiarów zbiorów danych.

W ramach pracy zaimplementowano dwa algorytmy oparte na wielowymiarowych modelach mieszanin, Gaussian Mixture EM i Multinomial Mixture EM, i porównano je z czterema algorytmami opartymi na odległości - aglomeracyjnym grupowaniem hierarchicznym, k-means, k-medoids i rozmytym c-means. Algorytmy zostały zastosowane zarówno do symulowanych, jak i rzeczywistych zbiorów danych, a wybrane metryki zostały wykorzystane do ilościowej oceny wyników grupowania, w tym skorygowany indeks Randa, współczynnik prostego dopasowania, ważony indeks Jaccarda, zrównoważona dokładność oraz metryki oparte na sprzężonym rozkładzie beta-binomialnym. Wyniki badań symulacyjnych Autora pokazały, że dla rozpatrywanych danych algorytmy oparte na modelu, w szczególności Gaussian Mixture EM i Multinomial Mixture EM, przewyższają w wielu przypadkach algorytmy oparte na odległości.

2. Struktura i zawartość rozprawy

Recenzowana praca doktorska obejmuje formalnie 4 główne rozdziały, poprzedzone wstępem oraz zakończone podsumowaniem. Zasadnicza część rozprawy liczy łącznie 121 stron, a także zawiera bibliografię liczącą 39 pozycji oraz spis rysunków i tabel.

Praca rozpoczyna się wstępem, w którym przedstawiono motywację i tezy postawione w rozprawie. Autor krótko przypomina cel nienadzorowanego uczenia maszynowego jako techniki stosowanej w analizie danych, w tym wielowymiarowych i złożonych zbiorów danych.

W rozdziale 2 dokonano przeglądu algorytmów bazujące na modelach, a w szczególności odmian algorytmu EM (*Expectation-Maximization*) jako efektywnej techniki obliczeniowej stosowanej do estymacji parametrów modelu statystycznego w obecności brakujących lub niekompletnych danych. Rozważono dwie wersje EM: wielowymiarową mieszaninę Gaussowską EM (*Multivariate Gaussian Mixture EM*) oraz mieszaninę wielomianową EM (*Multinomial Mixture EM*). Szczegółowo przedstawiono ich zasadę działania, różnice i dziedziny zastosowań, a także ich przykładowe implementacje (w formie uproszczonego pseudokodu).

W rozdziale 3 dokonano przeglądu algorytmów bazujące na dystansie, a w szczególności grupowanie hierarchiczne (aglomeracyjne i rozdzielcze) oraz metody *k-means*, *k-medoid* oraz *k-fuzzy*. Podkreślono istotną różnicą pomiędzy grupowaniem hierarchicznym, a pozostałymi trzema algorytmami oraz to, że ich wspólną cechą jest wykorzystanie metryki odległości do pomiaru podobieństwa między punktami danych.

W rozdziale 4 przedstawiono poszczególne etapy gromadzenia i przygotowania danych, w tym tworzenie symulowanych wielowymiarowych danych normalnych i wielomianowych, przygotowanie danych rzeczywistych (publicznie dostępnych), filtrację i dekompozycję wariancji, a także skalowanie oraz grupowanie danych.

W przedostatnim rozdziale przedstawiono wyniki analizy danych, zarówno stworzonych sztucznie (symulowanych) wygenerowanych dla wielowymiarowych mieszanin normalnych oraz mieszanin wielomianowych (przy użyciu wektorów losowych o zadanych parametrach), jak i pozyskanych z publicznie dostępnych zbiorów danych, takich jak danych medycznych o mutacjach somatycznych DNA, ekspresji genów, arytmii serca, aktywności sportowej oraz danych z teleskopu Kepplera.

W rozdziale 6 przypomniano cel rozprawy, podsumowano uzyskane wyniki oraz pokazano ich związek z przedstawionymi wcześniej tezami. Zwrócono uwagę na wykorzystane metryki, a więc skorygowany indeks Randa, prosty współczynnik dopasowania, ważony indeks Jaccarda, zrównoważoną dokładność oraz metryki oparte na sprzężonym rozkładzie beta-binomialnym, co pozwoliło na ilościowe określenie jakości wyników grupowania oraz porównanie wydajności różnych algorytmów na różnych zbiorach danych.

3. Najważniejsze osiągnięcia rozprawy

Biorąc pod uwagę zawartość pracy oraz pozytywną ocenę jej zawartości merytorycznej, za główne osiągnięcia Autora należy uznać przeprowadzenie obszernych badań, w tym zarówno na danych symulowanych z wykorzystaniem tysięcy wielomianowych mieszanin rozkładów gaussowskich i wielomianowych, jak również na zestawach rzeczywistych zbiorów danych z różnych publicznie dostępnych źródeł, w tym danych genomicznych i medycznych.

Wyniki badań symulacyjnych Autora pokazują, że dla przebadanych zbiorów danych algorytmy oparte na modelach, w szczególności Gaussian Mixture EM i Multinomial Mixture EM, w wielu przypadkach przewyższają algorytmy oparte na odległości. Opracowane algorytmy (opracowane przez Autora) są dostępne w języku R (repozytorium github).

Najważniejszymi elementami rozprawy, decydującymi o jej wartości naukowej i badawczej, są:

- Sformułowanie algorytmów dla mieszanin wielowymiarowych rozkładów Gaussa oraz rozkładów wielomianowych.
- Opracowanie narzędzi programowych (w środowisku języka R) implementujących nienadzorowane algorytmy grupowania oraz optymalizacja implementacji tak, aby umożliwiała grupowanie dużych zbiorów danych (rzędu setek tysięcy cech/obserwacji).
- Implementacja wybranych algorytmów grupowania opartych na odległości (na podstawie kodu źródłowego dostępnego w literaturze).
- Opracowanie narzędzi programowych implementujących zbiór wskaźników jakości grupowania w środowisku językowym R.
- Opracowanie narzędzi programowych do symulacji wielowymiarowych danych dla rozkładów Gaussa lub rozkładów wielomianowych.
- Utworzenie zbiorów zawierających rzeczywiste dane służących do badania porównawczego, z możliwością zmiany struktury i rozmiaru.
- Przeprowadzenie badania porównawczego dla wszystkich analizowanych algorytmów grupowania, zarówno dla rzeczywistego, jak i symulowanego zbioru danych.

Należy zauważyć, że Autor podjął się realizacji bardzo ciekawego oraz istotnego z punktu widzenia praktycznych zastosowań tematu badawczego. Poszczególne wyniki badań zostały opublikowane w kilku współautorskich pracach w języku angielskim, co świadczy pozytywnie o dużej wiedzy Autora rozprawy w zakresie poruszanej tematyki badawczej.

4. Poprawność pracy i uwagi krytyczne

Poprawność treści rozprawy nie wzbudza istotnych zastrzeżeń, a stwierdzenia w niej zawarte mogą być podstawą do dalszych badań, co wynika z zawartych w pracy podstaw teoretycznych popartych wynikami szczegółowych badań eksperymentalnych.

Jednocześnie Autor nie ustrzegł się pewnych drobnych niedociągnięć, a wśród uwag o charakterze krytycznym, a po trosze i dyskusyjnym, można wymienić następujące:

1. Aby porównać wydajność, Autor zaimplementował kilka wybranych algorytmów i opracował odpowiednie wskaźniki. Jak wygląda tu kwestia efektywności – czy czasy obliczeń (na tych samych zbiorach danych) są zbliżone, czy też niektóre algorytmy wymagają znacząco większych zasobów sprzętowych i długiego czasu obliczeń, a jednocześnie nie dają wyraźnie lepszych wyników w klasyfikacji?

2. Jedną z tez pracy było: „*nienadzorowane grupowanie oparte na mieszaninach dystrybucji jest konkurencyjne w porównaniu z metodami opartymi na odległości*". Czy, zdaniem Autora, sprawdzi się to dla każdej bazy danych tego typu, czy też istnieją zbiory danych, dla których ta teza niekoniecznie będzie prawdziwa?

3. Przeglądając wyniki w tabelach korelacji pomiędzy metrykami (rys. 5.5 i następne) można zauważyć, że korelacje dla danych symulowanych są zbliżone do danych rzeczywistych dla ekspresji genów TCGA, aktywności sportowej oraz arytmii serca (osiągając wynik 0.9...1.0), natomiast znacznie odbiegają od wyników dla mutacji somatycznych czy archiwum muzyki (osiągając znacznie mniejsze wartości). Zatem, czy możliwe jest takie wygenerowanie danych symulowanych, aby osiągnąć wyniki zbliżone np. do danych dla mutacji somatycznych? Jak wtedy należałoby ustawić parametry symulacji? Czy Autor wykonał takie symulacje?

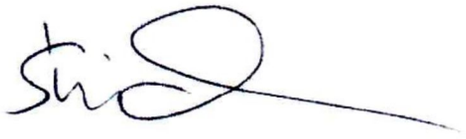
4. Drobne uwagi szczegółowe (najważniejsze):
   - Str. 16 – wzory (2.31), (2.40) wychodzą poza marginesy tekstu
   - Str. 22 – tekst w jednej linii przekracza margines z prawej strony
   - Str. 51, 52 – brak numeracji wzorów na stronach (następne są już numerowane)
   - Str. 61, 66 – etykiety na rysunku nieznacznie wychodzą poza prawy margines tekstu
   - Str. 64, 69 – rysunki (5.5), (5.10) nieznacznie wychodzą poza prawy margines tekstu
   - Str. 80, 87, 94, 101, 115 – opisy dla rys. (5.17), ... (5.47) wychodzą poza margines
   - Trochę mało pozycji literaturowych i są one zbyt rzadko cytowane w tekście pracy

5. Podsumowanie

Przytoczone wyżej uwagi dyskusyjne nie umniejszają zasług Autora ani nie kwestionują przedstawionych osiągnięć, a opisywana w pracy problematyka dotyczy aktualnych i interesujących zagadnień naukowych. Recenzowana praca zasługuje na pozytywną ocenę merytoryczną i wnosi istotny oraz oryginalny wkład w dziedzinę informatyki technicznej. Postawione cele i zadania pracy zostały zrealizowane, a jej tematyka wpisuje się we współczesny nurt badań w tym zakresie.

Stwierdzam zatem z pełnym przekonaniem, że opiniowana rozprawa Pana mgr Mateusza Kani pt. „*Data clustering with mixtures of multidimensional distributions*" zawiera samodzielne rozwiązanie ważnego i istotnego problemu naukowego, jednocześnie spełniając wszystkie wymagania przewidziane dla rozpraw doktorskich w aktualnie obowiązującej Ustawie o Tytule Naukowym i Stopniach Naukowych.

W związku z tym stawiam wniosek o **dopuszczenie rozprawy doktorskiej do publicznej obrony**.



Dr hab. inż. Zbigniew Świder