wpł. RPITT w dniu 13.06.2023
M. Skow

Dr hab. inż. Marek Kowal, prof. UZ
Instytut Sterowania i Systemów Informatycznych
Wydział Informatyki, Elektrotechniki i Automatyki
Uniwersytet Zielonogórski
email: M.Kowal@issi.uz.zgora.pl

Zielona Góra, 22.05.2023r.

Recenzja rozprawy doktorskiej

Tytuł rozprawy w języku angielskim: **Data clustering with mixtures of multidimensional distributions**

Tytuł rozprawy w języku polskim: **Grupowanie danych z wykorzystaniem mieszanin wielowymiarowych rozkładów**

Autor rozprawy: **mgr Mateusz Kania**

Promotor rozprawy: **prof. dr hab. inż. Andrzej Polański**

Dziedzina: **nauki techniczne**

Dyscyplina: **informatyka techniczna i telekomunikacja**

1. Cel, zakres i charakter rozprawy

Głównym celem recenzowanej pracy doktorskiej było opracowanie, implementacja i weryfikacja efektywności algorytmów grupowania danych w oparciu o mieszaniny rozkładów wielowymiarowych. Zaproponowano dwa podejścia do budowy modeli na potrzeby algorytmów grupowania, które bazowały odpowiednio na mieszaninie wielowymiarowych rozkładów normalnych oraz mieszaninie rozkładów wielomianowych. Opracowane algorytmy zaadaptowano do grupowania danych o bardzo dużej wymiarowości, rzędu dziesiątek tysięcy zmiennych. Istotnym celem pracy była ocena efektywność opracowanych algorytmów. W toku wielu eksperymentów algorytmy przetestowano z wykorzystaniem różnorodnych zbiorów danych. Ponadto, opracowane metody porównano z czterema odległościowymi metodami grupowania danych.

Praca swoim zakresem obejmowała przedstawienie podstaw teoretycznych z zakresu modeli opartych o mieszaniny rozkładów prawdopodobieństwa. W szczególności skupiono się na omówieniu algorytmu EM (Expectation Maximization), który był podstawą konstrukcji opracowanych w ramach pracy algorytmów grupowania. Istotnym elementem przeprowadzonych rozważań teoretycznych było dostosowanie ogólnego mechanizmu grupowania danych za pomocą mieszaniny rozkładów dla scenariuszy, w których wykorzystano mieszaninę wielowymiarowych rozkładów normalnych oraz mieszaninę rozkładów wielomianowych. Ponadto w zakres pracy weszło omówienie implementacji zaproponowanych algorytmów grupowania, które opisano za pomocą pseudokodów oraz przedstawiono wiele innych istotnych szczegółów związanych z implementacją, takich jak proces inicjalizacji algorytmów, warunek stopu, problemy natury numerycznej i metody radzenia sobie z nimi, itp. Opracowane metody zostały zaimplementowane przez

Doktoranta w języku R i udostępnione w publicznym repozytorium. W ramach pracy zaprezentowano również podstawy teoretyczne metod grupowania danych bazujące na odległościach. Przedstawiono wybrane metody grupowania hierarchicznego oraz wybrane metody typu iteracyjno-optymalizacyjnego. W zakresie pracy było również omówienie procesu generowania sztucznych danych oraz przedstawienie zbiorów danych rzeczywistych, które wykorzystano do weryfikacji efektywności zaproponowanych metod grupowania danych. W tym obszarze omówiono również proces wstępnego przetwarzania danych wejściowych i ich wizualizacji. Zaprezentowano również miary jakości grupowania danych wykorzystane do oceny efektywności testowanych algorytmów. Dodatkowo opisano biblioteki programistyczne wspomagające obliczenia numeryczne. Zakres rozprawy obejmował także przeprowadzenie kompleksowych badań porównawczych dla opracowanych metod grupowania danych z metodami grupowania odległościowego. Zaprezentowano bogate raporty i analizy dotyczące wyników przeprowadzonych badań, z których wynika, że opracowane metody stanowią dobrą alternatywę dla metod grupowania bazujących na odległościach. W krótkim podsumowaniu zaprezentowano zagregowane wyniki badań, płynące z nich wnioski oraz potencjalne kierunki dalszych badań.

Rozprawa ma charakter badawczy. W jej ramach proponuje się zmodyfikowane metody i algorytmy informatyczne do grupowania danych. Aby przedstawić proponowane rozwiązania Doktorant operuje odpowiednim formalizmem matematycznym. Jednak rozważania teoretyczne są raczej tylko elementem pomocniczym pozwalającym zrozumieć ideę proponowanych metod grupowania danych. Ze względu na specyfikę poruszanych problemów, poprawność i skuteczność zaproponowanych metod grupowania danych wykazuje się przez wykonanie bardzo licznych eksperymentów na zbiorach danych symulowanych i rzeczywistych. W rozprawie pojawia się również wątek dotyczący adaptacji zaproponowanych metod do grupowania danych o bardzo dużej wymiarowości. Dzięki czemu zaproponowane metody grupowania znajdują zastosowanie między innymi w genomice. Praca ujawnia również charakter aplikacyjny ponieważ zaproponowane metody zostały zaimplementowane w języku R i udostępnione publicznie w formie biblioteki.

Uwzględniwszy powyższe fakty uważam, że cel rozprawy Pana magistra Mateusza Kanii jest sformułowany właściwie ponieważ dotyczy aktualnego problemu naukowego, którego rozwiązanie ma istotne znaczenie dla rozwoju Informatyki Technicznej i Telekomunikacji. Zakres pracy jest prawidłowy, właściwy podjętemu problemowi naukowemu.

2. Zawartość rozprawy

Rozprawa doktorska została napisana w języku angielskim. Praca liczy 131 stron i składa się z 6 rozdziałów, spisu literatury, spisu rysunków oraz spisu tabel.

Rozdział 1 jest krótkim wprowadzeniem do tematyki rozprawy. Na wstępie tego rozdziału zaprezentowano pobieżnie ogólne zagadnienia z obszaru nienadzorowanego grupowania danych oraz sformułowano problem doboru algorytmu grupowania danych do rozwiązywanego problemu. Następnie sformułowano cel i trzy tezy pracy. W kolejnych podrozdziałach wymieniono oryginalne osiągnięcia, które pozwoliły udowodnić tezy. Ponadto wymieniono publikacje naukowe, w których doktorant zaprezentował swoje

osiągnięcia naukowe oraz wskazano repozytorium GitHub, w którym Doktorant zamieścił kod w języku R, który powstał w ramach pracy doktorskiej.

Rozdział 2 poświęcony jest na przedstawienie podstaw teoretycznych metod grupowania danych opartych na mieszaninie rozkładów oraz na wyprowadzenie, dostosowanie i opis implementacji dwóch algorytmów grupowania danych opartych na mieszaninie wielowymiarowych rozkładów normalnych oraz rozkładów wielomianowych.
Na wstępie przedstawiono podstawowe informacje na temat rozkładu normalnego, Bernoulliego, dwumianowego i wielomianowego. Następnie wprowadzono pojęcie mieszanin rozkładów oraz zaprezentowano opis i podstawowe cechy mieszanin rozkładów normalnych, dwumianowych oraz wielomianowych.
W dwóch kolejnych podrozdziałach 2.2 i 2.3 skupiono się na przedstawieniu procesu estymacji parametrów dla mieszanin rozkładów z wykorzystaniem metody największej wiarygodności. Zaprezentowano jak rozwiązuje się taki problem za pomocą metody EM (Expectation Maximization). W kolejnych podrozdziałach przedstawiono kolejne etapy konstrukcji algorytmów grupowania danych w oparciu o jednowymiarową mieszaninę rozkładów normalnych, wielowymiarową mieszaninę rozkładów normalnych oraz wielowymiarową mieszaninę rozkładów normalnych z diagonalną macierzą kowariancji z wykorzystaniem metody EM. Ostatecznie przedstawiono pseudokod ilustrujący implementację opracowanej metody grupowania danych opartej o mieszaninę wielowymiarowych rozkładów normalnych oraz doprecyzowano szczegóły implementacyjne związane z inicjalizacją parametrów, warunkiem stopu, poszczególnymi krokami metody EM. Opisano również w jaki sposób radzono sobie z problemami numerycznymi związanymi z bardzo dużymi lub bardzo małymi wartościami, które pojawiają się podczas obliczeń.
W ostatnim podrozdziale przedstawiono kolejne etapy konstrukcji algorytmu grupowania danych za pomocą mieszanki rozkładów wielomianowych. Podobnie jak w przypadku rozkładu normalnego przeprowadzono rozważania teoretyczne, które pozwoliły wyprowadzić wzory na estymację parametrów mieszanki rozkładów za pomocą metody EM. Zaprezentowano pseudokod dla implementacji zaadaptowanej metody EM oraz opisano szczegóły implementacyjne związane z inicjalizacja parametrów, warunkiem stopu oraz realizacją dwóch głównych kroków iteracyjnej metody EM.

Rozdział 3 przybliża tematykę grupowania danych z wykorzystaniem metod odległościowych. Metody te są przedstawione w ramach rozprawy ponieważ z nimi będą porównywane metody opracowane w rozdziale 2.
W pierwszym podrozdziale zdefiniowano kluczowe pojęcia związane z metodami odległościowymi takie jak odległość, skala, standaryzacja zmiennych. Zaprezentowano trzy dobrze znane miary odległości: euklidesową, Minkowskiego i Manhattan.
Kolejny podrozdział poświęcono na opisanie grupowania danych za pomocą podejścia hierarchicznego. Opisano ideę podejścia aglomeracyjnego i deglomeracyjnego. Następnie sformułowano różne miary, które wykorzystuje się do pomiaru odległości/niepodobieństwa pomiędzy klastrami danych.
W podrozdziale trzecim zaprezentowano iteracyjno-optymalizacyjne podejście do grupowania danych za pomocą metody k-średnich (k-means). Opisano ogólną ideę takiego podejścia oraz przedstawiono jego implementację za pomocą algorytmu Hartigan-a Wong-a.
Podrozdział czwarty poświęcony jest rozmytemu algorytmowi fuzzy c-means (FCM). Zaprezentowano w nim ogólną ideę tego algorytmu oraz funkcję celu, która jest

minimalizowana podczas iteracyjnej procedury aktualizującej środki klastrów i przynależność danych do klastrów.
Ostatni podrozdział przedstawia podstawy działania algorytmu k-medoidów

Rozdział 4 rozpoczyna się od zaprezentowania ogólnego schematu przetwarzania danych, który wykorzystano podczas wykonanych w ramach pracy eksperymentów badawczych. Pozostała część tego rozdziału zawiera sześć podrozdziałów opisujących w głównej mierze zbiory danych użyte do eksperymentów weryfikujących efektywność metod grupowania, metody wstępnego przetwarzania danych oraz definicję miar jakości grupowania danych.
W pierwszym podrozdziale zaprezentowano mechanizmy generacji danych syntetycznych z wykorzystaniem generatorów liczb pseudolosowych oraz scharakteryzowano sześć zbiorów z danymi rzeczywistymi. Zbiory rzeczywiste pochodziły z ogólnodostępnych repozytoriów Kaggle, UCI, NCBI oraz NCI. Były to odpowiednio zbiory z danymi dotyczącymi mutacji somatycznych DNA wśród pacjentów chorych na różne typy raka (Somatic Mutation Counts), ekspresji genów wśród pacjentów chorych na rożne typy raka (Gene Expressions), klasyfikacji gatunków organizmów na podstawie kodonów (Codons frequency), klasyfikacji arytmii na podstawie cech z elektrokardiogramów (Arrhythmia), klasyfikacji gatunku muzyki na podstawie różnych cech utworów (The Free Music Archive), klasyfikacji rodzajów aktywności sportowej na podstawie pomiarów z czujników umieszczonych na ubraniach ćwiczących osób (Sport activities) oraz danych astronomicznych, których celem była identyfikacja planet charakteryzujących się środowiskiem podobnym do ziemskiego (NASA Kepplers).
Podrozdział drugi jest bardzo krótki i tłumaczy tylko proces generowania różnych wersji zbiorów danych rzeczywistych poprzez ograniczenie liczby klastrów do 2, 3, 4, 5 i 6 klastrów. Wygenerowano dla każdej liczby klastrów po 10 różnych kombinacji klastrów w efekcie otrzymując z jednego zbioru danych 50 podzbiorów. Celem tego działania było zwiększenie istotności wyników eksperymentów, które zamiast na jednym zbiorze były uzyskiwane dla wielu różniących się od siebie zbiorów danych.
Podrozdział trzeci został poświęcony na opisanie metod wstępnego przetwarzania danych, do których zaliczono selekcję cech z wykorzystaniem kryterium BIC (Bayesian Information Criterion) i dekompozycji wariancji zmiennych oraz standaryzację danych.
Podrozdział czwarty jest bardzo krótki i ma formę tabeli, w której wymieniono algorytmy grupowania danych, które były wykorzystywane w eksperymentach prowadzonych w ramach pracy doktorskiej.
Podrozdział piąty tłumaczy w jaki sposób za pomocą algorytmu węgierskiego przypisywano prawidłowe etykiety grupom danych znajdowanym przez algorytmy grupowania. Ponadto podrozdział definiuje wiele miar jakości grupowania, które wykorzystano podczas eksperymentów do oceny poszczególnych algorytmów grupowania. Również w ramach tego podrozdziału opisano cztery metody redukcji wymiarowości (PCA, tSVD, tSNE i losowe projekcje), które później wykorzystano do wizualizacji danych wielowymiarowych.
Ostatni szósty podrozdział poświęcono na omówienie bibliotek programistycznych do algebry liniowej (BLAS, LAPACK) oraz biblioteki wspierającej obliczenia równoległe (OpenMP). Ponadto omówiono w skrócie cechy języka R.

Rozdział 5 przedstawia wyniki eksperymentów, których celem było porównanie algorytmów grupowania danych z wykorzystaniem różnych zbiorów danych. W sumie wykonano 9 głównych eksperymentów odpowiednio dla danych syntetycznych wygenerowanych z mieszaniny rozkładów normalnych, mieszaniny rozkładów wielomianowych oraz 7 zbiorów z danymi rzeczywistymi. W sumie przetestowano 6

różnych metod grupowania danych w tym 2 metody bazujące na mieszaninach rozkładów wielowymiarowych oraz 4 algorytmy odległościowe. Należy podkreślić, że eksperymenty dla poszczególnych zbiorów danych były wielokrotnie powtarzane dla różnej liczby klastrów i różnych ich kombinacji. W efekcie uzyskiwano bardzo dużo wyników a ocena poszczególnych metod grupowania bazowała na statystykach miar jakości obliczanych z wielu powtórzeń dla różnych podzbiorów testowych. Raporty z wynikami z poszczególnych eksperymentów są do siebie bardzo podobne. Zawsze na wstępie eksperymentu dla zadanego zbioru danych omówiona jest struktura zbioru danych, która jest ilustrowana wizualizacją danych po redukcji wymiarowości. Następnie w formie wykresów pudełkowych zaprezentowane są wyniki dla miary ARI (Adjusted Rand Index) w rozbiciu na metody grupowania danych, metody wstępnego przetwarzania danych oraz liczbę klastrów w zbiorze testowym. Następnie prezentowane są wykresy z rozkładami opisującymi prawdopodobieństwo prawidłowego przypisania danych do klastra dla poszczególnych metod grupowania danych oraz różnych metod wstępnego przetworzenia danych. Za pomocą wykresów kołowych zaprezentowano statystyki dla miary MBA (Median Balanced Accuracy) w rozbiciu na metody grupowania oraz metody wstępnego przetworzenia danych. Kolejne wykresy prezentują miary jakości WSMC (Weighted Simple Matching Coefficient), SMC (Simple Matching Coefficient) i WJACC (Weighted Jaccard) w postaci wykresów skrzypcowych w rozbiciu na metody grupowania oraz metody wstępnego przetwarzania danych. Na zakończenie każdego raportu prezentowana jest macierz korelacji dla miar jakości aby ocenić spójność ocen jakości grupowania. Wyniki zaprezentowane na wymienionych wykresach są za każdym razem szczegółowo omówione w tekście danego podrozdziału. Od przedstawionego schematu prezentacji wyników odbiega jedynie eksperyment przeprowadzony dla zbioru danych z obserwacji astronomicznych (NASA Kepplers) ponieważ w jego przypadku mamy do czynienia tylko z dwoma klastrami danych więc zaprezentowano wyłącznie strukturę danych oraz wykresy przedstawiające rozkłady prawdopodobieństwa prawidłowego przypisywania danych do klastra. Uzyskane wyniki wskazują, że metody grupowania danych bazujące na mieszaninie rozkładów uzyskują porównywalne wyniki do metod grupowania bazujących na odległości a dla modelu z mieszaniną rozkładów normalnych wielu przypadkach nawet lepsze.

Rozdział 6 jest podsumowaniem wyników eksperymentów zaprezentowanych w ramach rozdziału piątego. Ze względu na bardzo dużą liczbę wyników zaprezentowanych w rozdziale piątym porównanie metod grupowania danych jest utrudnione. Dlatego w rozdziale szóstym zaprezentowano trzy tabele z zagregowanymi wynikami odpowiednio dla danych symulowanych z mieszaniny rozkładów normalnych, danych symulowanych z mieszaniny rozkładów wielomianowych oraz danych rzeczywistych. W tym przypadku dla poszczególnych metod grupowania danych w rozbiciu na metody wstępnego przetwarzania danych zaprezentowano znormalizowaną liczbę poprawnych przypisań danych do klastrów. Zaprezentowane wyniki potwierdziły, że metoda grupowania bazująca na mieszaninie rozkładów normalnych (GMM) uzyskała dla danych symulowanych za pomocą mieszaniny rozkładów normalnych oraz danych rzeczywistych najlepsze wyniki.

Bibliografia zawiera 39 pozycji literaturowych. W większości są to artykuły z recenzowanych czasopism o zasięgu międzynarodowym oraz pozycje książkowe renomowanych wydawnictw. Pozycje literaturowe są aktualne i dobrze dobrane. Jednak niewielka liczba jak na rozprawę doktorską przytoczonych pozycji literaturowych budzi

pewien niedosyt. Tym bardziej, że temat poruszony w ramach rozprawy jest popularny i obecny w literaturze naukowej.

Po zapoznaniu się z treścią rozprawy stwierdzam, że mgr Mateusz Kania zaprezentował w jej ramach rozwiązanie oryginalnego problemu naukowego. Wykazał się przy tym, umiejętnością samodzielnego formułowania i rozwiązywania problemów naukowych oraz ogólną wiedzą teoretyczną i praktyczną w dyscyplinie Informatyka Techniczna i Telekomunikacja. W ramach pracy wykonano I przedstawiono bardzo obszerne badania eksperymentalne z wykorzystaniem rzeczywistych oraz symulowanych zbiorów danych. Należy przy tym podkreślić, że zaprezentowane badania są niezwykle istotne ponieważ mogą znaleźć zastosowanie w badaniach medycznych z zakresu genomiki w obszarze badań nad nowotworami.

3. Poprawność i oryginalność postawionych tez i stopień w jakim one zostały wykazane

W rozprawie sformułowano trzy tezy:

1) ***Unsupervised clustering methods based on mixtures of distributions achieve optimal performance when data statistics are consistent with actual distributions.***

2) ***Unsupervised clustering based on distributions' mixtures is competetive compared to distance-based methods.***

3) ***Applicability of clustering based on mixtures of distributions to practical problems relies on elaborating algorithmic implementation specialized for large sizes of datasets.***

Tezy zostały sformułowane poprawnie ponieważ zawierają stwierdzenia, które są weryfikowalne na bazie zaplanowanych prac badawczych i aktualnego stanu wiedzy naukowej. Jednocześnie bieżąca wiedza nie daje nam oczywistych odpowiedzi na stwierdzenia zawarte w tezach więc należy je uznać warte przeprowadzania procesu badawczego w celu ich sprawdzenia.

Aby wykazać niniejsze tezy, wykonano szereg badań eksperymentalnych i rozwiązano następujące problemy naukowe:

- Opracowano dwa algorytmy grupowania danych z wykorzystaniem modeli mieszanin rozkładów wielowymiarowych, które zaadaptowano dla problemów o ogromnej wymiarowości rzędu dziesiątek tysięcy wymiarów co pozwoliło zastosować opracowane algorytmy dla danych rzeczywistych, między innymi w obszarze bioinformatyki. Uzyskane w ramach przeprowadzonych eksperymentów obliczeniowych wyniki potwierdziły tezę nr 3.

- Przeprowadzono niezwykle obszerne badania eksperymentalne, które zweryfikowały prawdziwość tezy nr 1 oraz nr 2. Badania przeprowadzono dla danych symulowanych oraz bogatego zbioru różnego typu danych rzeczywistych. Przeprowadzono kompleksową ocenę efektywności opracowanych algorytmów i porównano je z wieloma algorytmami grupowania bazującymi na miarach odległości.

- Zaimplementowano wiele algorytmów grupowania danych oraz miar jakości do oceny i wizualizacji procesu grupowania danych w języku R. Zostały one udostępnione publicznie w formie biblioteki .

Doktorant wykazał się bogatym warsztatem badawczym, który umożliwił mu skonstruowanie i implementację zaawansowanych algorytmów, zaprojektowanie i przeprowadzenie złożonych eksperymentów badawczych oraz przeprowadzenie analizy wyników co w efekcie pozwoliło udowodnić tezy postawione w ramach rozprawy.
Na dojrzałość naukową Doktoranta wskazują również jego publikacje naukowe w czasopismach naukowych takich jak Scientific Reports (IF=4.996), Applied Sciences (IF=2.838) oraz w materiałach konferencji międzynarodowej i w dwóch rozdziałach książek.

Biorąc pod uwagę powyższe fakty uznaje, że wszystkie tezy zostały w pełni udowodnione z wykorzystaniem przyjętych w nauce metod i procedur badawczych. Przedstawione wyniki naukowe potwierdzone badaniami eksperymentalnymi są istotne dla reprezentowanej dyscypliny naukowej.

4. Uwagi i komentarze

Rozprawę doktorską jako całość oceniam dobrze. Uwagi i komentarze, które zamieszczam poniżej nie podważają pozytywnej oceny rozprawy i jej wkładu w rozwój dyscypliny naukowej.

1. W pracy zabrakło przedstawienia bieżącego stanu wiedzy w obszarze algorytmów grupowania danych. Takim wprowadzeniem Doktorant powinien wykazać, że orientuje się w swojej dyscyplinie tzn. zna bieżące kierunki badań oraz wie jakie problemy nie zostały jeszcze rozwiązane.

2. Zaprezentowana na str. 47 procedura dekompozycji wariancji jest przedstawiona zbyt pobieżnie. Należałoby również opisać bardziej szczegółowo ogólny schemat procesu selekcji cech w oparciu o BIC.

3. Niektóre z użytych algorytmów lub pojęć np. odległość euklidesową na str. 34 lub algorytm węgierski na str. 49 wyjaśniono za pomocą przykładów wziętych z życia codziennego. Taki zabieg jest często wykorzystywany w przypadku podręczników ale niekoniecznie pożądany w przypadku rozprawy doktorskiej gdzie spodziewamy się rozważań o wyższym poziomie abstrakcji. W przypadku algorytmu węgierskiego należałoby precyzyjniej opisać w jaki sposób jest on wykorzystywany do przypisania właściwych etykiet dla znalezionych przez algorytm klastrów.

4. Na stronie 52 scharakteryzowano właściwości rozkładu beta-dwumianowego jednak tylko zdawkowo wyjaśniono jak jest on wykorzystywany do pomiaru jakości wyników grupowania danych.

5. Praca zyskałaby gdyby przeprowadzono pogłębioną analizę problematycznych z punktu widzenia obliczeń kroków poszczególnych algorytmów oraz opisano jak sobie z nimi radzono w języku R. W pracy wykorzystano zbiory danych o bardzo dużej wymiarowości dlatego warto byłoby wspomnieć jak to przekładało się na złożoność obliczeniową oraz jakich nakładów czasowych wymagało grupowanie dla poszczególnych zbiorów danych.

6. Na str. 71 zbyt pobieżnie opisano proces wstępnego przekształcenia danych dotyczących mutacji somatycznych. Praca zyskałaby gdyby szerzej zaprezentowano zagdanienia związane z genomiką.

7. W tabeli 4.1 na str. 46 warto byłoby dla poszczególnych zbiorów danych podać liczbę atrybutów opisujących próbki.

8. W rozdziale 6 będącym podsumowaniem należało wskazać wyniki badań, które potwierdzają tezy pracy.

9. W opisie wzoru 2.4 na str. 7 pojawia się informacja, że p^k oraz $(1-p)^{n-k}$ to odpowiednio prawdopodobieństwo sukcesu i porażki. Jednak prawdopodobieństwo sukcesu i porażki w pojedynczej próbie jest odpowiednio dane za pomocą p oraz $(1-p)$.

10. We wzorze 2.5 na str. 7 pojawiła się zmienna x, która chyba nie powinna się tam znaleźć.

11. We wzorze 2.13 na str. 11 zostały błędnie użyte indeksy k, l, n. Podobnie jak w przypadku wzoru 2.34 błędnie zdefiniowano prawdopodobieństwo sukcesu i porażki w opisie pod wzorem.

12. We wzorze 2.55 i 2.57 na str. 26 i 27 indeksy przy sumowanych zmiennych są niepoprawne.

13. Dla wzoru 3.11 na str. 37 należy zdefiniować znaczenie zmiennej Z oraz doprecyzować, że wzór opisuje aktualizację odległości pomiędzy klastrami po procesie łączenia klastrów.

14. Na str. 38 należałoby wyjaśnić precyzyjniej dlaczego odległość Manhattan jest lepsza od odległości euklidesowej dla danych o dużej wymiarowości.

15. Na str. 26 w punkcie 2.4.2.1 pojawia się stwierdzenie, że punkty danych należą do liczb naturalnych "Data points belong to N". Jednak to chyba chodzi o wartości pojedynczych atrybutów próbek danych.

16. Na str 47 w punkcie 4.2.1 użyto do opisu procesu generacji podzbiorów z różnymi klastrami pojęcia permutacji jednak wydaje się, że w tym przypadku mamy do czynienia z kombinacjami.

5. Uwagi szczegółowe na temat błędów o charakterze redakcyjnym i edytorskim

str. 8: W opisie wzoru 2.8 przy definicji ograniczenia sumowania się do jedności współczynników mieszanin zabrakło znaku "+".

str. 13: We wzorze 2.18 do opisu odchylenia standardowego dla rozkładu jednowymiarowego użyto symbolu Σ zamiast σ a we wcześniejszym wzorze 2.17 użyto σ.

str. 21: Brak spacji w zdaniu "...observationsx_n...".

str. 28: Literówka w zdaniu "... divinding ...".

str. 33: Pomyłka w zdaniu "... object x and object y is equal to the distance between object j and object i ...".

str. 34: We wzorze 3.4 za krótki znak pierwiastka.

str. 35: Pojawia się dwukrotnie: "(pic!)".

str. 39: Błąd redakcyjny w zdaniu "... different two different ...".

str. 39: Zdanie bez zakończenia "From the results we can see that".

str. 40: Niefortunne stwierdzenie odnośnie procesu minimalizacji funkcji celu "FCM tries to minimize objective function:".

str. 41: Nie rozwinięto skrót PAM (Partitioning Around Medoids).

str. 50: We wzorze 4.6 pojawia się współczynnik *K* a w wyjaśnieniu *k*.

str. 51: Użyto dwukrotnie nazwy współczynnika "Sensitivity" zamiast w drugim wzorze użyć "Specificity".

str. 53: Literówka w słowie "tesis".

str. 54: W akapicie "SVD and tSVD" użyto dwukrotnie nazwy "tSNE" zamiast "tSVD".

str. 55: Literówka "Beta-binomial distrubtion ...." .

str. 56: Błąd redakcyjny w akapicie "LAPACK" "more comhttps://..." .

str. 57: Literówka "metioned".

str. 86: Błąd redakcyjny w zdaniu "... with 20 amino acids ...".

6. Wnioski końcowe

Recenzowana rozprawa doktorska podejmuje ważny z punktu widzenia dyscypliny Informatyka Techniczna i Telekomunikacja problem naukowy. Wyniki przedstawione w rozprawie potwierdzają tezy pracy i wnoszą oryginalny wkład do dyscypliny naukowej. Doktorant udowodnił, że posiada szeroką wiedzę z obszaru reprezentowanej dyscypliny oraz potrafi rozwiązywać trudne problemy naukowe stosując przyjęte w nauce metody poznawcze i badawcze.

Podsumowując stwierdzam, że recenzowana rozprawa doktorska spełnia wszystkie wymogi stawiane rozprawom doktorskim przez obowiązujące przepisy. W związku z powyższym wnioskuję o dopuszczenie recenzowanej rozprawy doktorskiej do publicznej obrony.

Marek Kawa