Dr hab. inż. Grzegorz Dudek, prof. PCz
Katedra Automatyki, Elektrotechniki i Optoelektroniki
Wydział Elektryczny
Politechnika Częstochowska
Al. Armii Krajowej 17
42-200 Częstochowa

Częstochowa, dn. 23 maja 2023 r.

RECENZJA

rozprawy doktorskiej mgr Mateusza Kani

pt. *Data clustering with mixtures of multidimensional distributions*

Formalną podstawą opracowania recenzji jest pismo Przewodniczącego Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej, prof. dr hab. inż. Andrzeja Polańskiego. Oceny rozprawy doktorskiej dokonano według kryteriów określonych w ustawie z 20 lipca 2018 r. *Prawo o szkolnictwie wyższym i nauce*. Promotorem rozprawy doktorskiej jest prof. dr hab. inż. Andrzej Polański.

Charakterystyka rozprawy

Rozprawa napisana jest w języku angielskim, liczy 131 stron. Składa się z sześciu rozdziałów, bibliografii zawierającej 39 pozycji oraz wykazów ilustracji i tabel.

Tezy pracy są następujące:

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 competitive 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.*

We wstępie Autor scharakteryzował tematykę rozprawy, zawarł cel i tezy pracy, opisał oryginalne elementy pracy, podał publikacje związane z tematyka rozprawy, w których brał udział jako współautor oraz zamieścił odnośnik do kodów źródłowych zaimplementowanych przez siebie algorytmów.

Rozdział 2, *Model-based algorithms*, zawiera opis różnych wariantów algorytmów mieszaniny wielowymiarowych rozkładów gaussowskich oraz mieszaniny rozkładów wielomianowych. Opisy te poprzedzono definicjami odpowiednich rozkładów prawdopodobieństw oraz charakterystyką algorytmu maksymalizacji oczekiwań (*expectation–maximization*, EM).

W rozdziale 3, *Distance-based algorithms*, opisano porównawcze algorytmy grupowania: grupowanie hierarchiczne, k-średnich w wersji ostrej i rozmytej oraz k-medoidów.

W rozdziale 4, *Study pipeline*, Autor przedstawia schemat procedury badań eksperymentalnych. Procedura ta obejmuje gromadzenie lub generowanie danych, wstępne przetwarzanie danych, grupowanie danych, ocenę grupowania, wizualizację wyników oraz optymalizację algorytmów.

Rozdział 5 przedstawia wyniki badań eksperymentalnych dla zbiorów danych symulacyjnych i rzeczywistych. Badania podsumowano w rozdziale 6.

Opinia na temat rozprawy – uwagi krytyczne i polemiczne

Tematyka rozprawy dotyczy metod grupowania danych opartych na mieszaninach rozkładów. Grupowanie danych to istotny obszar badawczy w dziedzinie analizy danych, uczenia maszynowego i sztucznej inteligencji. Grupowanie pozwala na odkrywanie ukrytych wzorców i struktur w dużych zbiorach danych, co może prowadzić do lepszej klasyfikacji, segmentacji czy rekomendacji. Ma szerokie zastosowanie w wielu dziedzinach, takich jak medycyna, ekonomia, marketing, biologia czy analiza społeczna. Badania w tym obszarze mogą przynieść istotny wkład w rozwój tych dziedzin poprzez dostarczanie lepszych narzędzi analizy danych, identyfikację istotnych podgrup czy odkrywanie nowych zależności. W szczególności, metody grupowania oparte na mieszaninach rozkładów stanowią ciekawy obszar badawczy, który pozwala na bardziej zaawansowane modelowanie struktury danych. Dlatego uważam, że tematyka doktoratu leży w niezwykle ważnym obszarze badawczym. Oferuje wiele możliwości rozwoju wiedzy i umiejętności, a także przyczynia się do rozwinięcia nauki i technologii.

Tytuł rozprawy jest odpowiednio zwarty i komunikatywny. W pełni oddaje najistotniejsze elementy treściowe rozprawy. We wstępie rozprawy podano cel i tezy. Celem badań jest opracowanie i implementacja metod grupowania danych opartych na mieszaninach rozkładów: mieszaninie wielowymiarowych rozkładów normalnych z diagonalną macierzą kowariancji oraz mieszaninie rozkładów wielomianowych. Pierwsza teza jest dla mnie niezrozumiała. Mówi ona o tym, że metody grupowania oparte na mieszaninach rozkładów osiągają optymalną wydajność, gdy statystyki danych są zgodne z rzeczywistymi rozkładami. Statystyki danych opisują ich rozkłady, więc jak może tutaj być niezgodność? Prawdopodobnie Autor miał na myśli zgodność rozkładów przyjętych w mieszaninach z rzeczywistym rozkładem danych, ale wtedy teza powinna być sformułowana inaczej. Kontrowersyjna jest również trzecia teza: możliwość zastosowania grupowania opartego na mieszaninach rozkładów do problemów praktycznych polega na opracowaniu algorytmicznej implementacji wyspecjalizowanej dla dużych zbiorów danych. Być może Autor miał na myśli to, że aplikacyjność metody zwiększy się, jeśli będzie ona przystosowana do analizy dużych zbiorów danych. Ale czy to oczywiste stwierdzenie może być tezą doktoratu? Druga teza nie budzi zastrzeżeń, można ją zweryfikować eksperymentalnie, co Autor czyni w rozdziale piątym.

W dalszej części wstępu podano oryginalne elementy rozprawy, wśród których najważniejszy z punktu widzenia naukowego dotyczy opracowania algorytmów grupowania danych opartych na mieszaninach rozkładów dla dużych zbiorów danych, z setkami tysięcy cech. Na pochwałę zasługuje udostępnienie przez Autora w publicznym repozytorium GitHub kodów źródłowych w języku R algorytmów, które powstały w ramach pracy. We wstępie zabrakło opisu układu pracy z zawartością poszczególnych rozdziałów.

W rozprawie brakuje opisu problematyki grupowania danych oraz przeglądu literatury. Należało zdefiniować problem grupowania, jego warianty w zależności od typu grupowanych danych, opisać algorytmy grupowania i podać metody oceny jakości grupowania. Przegląd literatury powinien obejmować klasyczne algorytmy grupowania, jak i wybrane nowe rozwiązania. Należało opisać ich działanie, zalety i ograniczenia, podać ich potencjalne zastosowania, omówić wyzwania i aktualne kierunki badawcze w tej dziedzinie. Przegląd literatury powinien obejmować także alternatywne do proponowanej w rozprawie metody grupowania oparte na mieszaninach rozkładów, które doczekały się implementacji w środowisku R, np. mclust, Rmixmod, mixture i EMCluster.

Rozdział drugi, dotyczący algorytmów grupowania opartych na mieszaninach rozkładów, rozpoczyna się od charakterystyki kilku typów rozkładów prawdopodobieństw, które wykorzystywane są w mieszaninach, m.in. wielowymiarowego rozkładu Gaussa w wersji ze skorelowanymi i nieskorelowanymi zmiennymi losowymi (z diagonalną macierzą kowariancji) oraz rozkładu wielomianowego. W dalszej części objaśniono jak tworzy się mieszaniny tych rozkładów oraz scharakteryzowano algorytm EM. Podrozdział 2.3 opisuje mieszaniny wielowymiarowych rozkładów Gaussa. Autor poprawnie wyprowadza wzory na parametry rozkładów, najpierw dla przypadku jednowymiarowego, następnie wielowymiarowego w wariancie z „pełną" i diagonalną macierzą kowariancji. Sposób wyprowadzania wzorów w tych trzech przypadkach jest identyczny, więc można było się ograniczyć do przypadku najbardziej ogólnego – wielowymiarowego z „pełną" macierzą kowariancji, podając końcowe wzory dla rozpatrywanych trzech przypadków. W ten sposób można było uniknąć licznych i niepotrzebnych moim zdaniem powtórzeń. Defektem tej części pracy jest zbyt drobiazgowy podział na podrozdziały i brak wizualizacji przedstawiającej ideę metody.

W podrozdziale 2.3.4 opisano implementację algorytmu mieszaniny wielowymiarowych rozkładów Gaussa w języku R. Oznaczenia występujące w tym opisie sugerują, że jest to implementacja algorytmu z diagonalną macierzą kowariancji (Autor powinien to wyraźnie zaznaczyć). Pseudokod podany w Algorytmie 2.1 wymaga objaśnienia. Autor w ogóle nie powołuje się na ten pseudokod, opisując w dalszej części kolejne elementy algorytmu. Poza tym występują niezgodności pomiędzy pseudokodem a opisem w tekście, np. według pseudokodu wariancje inicjalizuje się środkami (*means*) powiększonymi o wartość 1e-4, gdy tymczasem według opisu środki mnoży się przez wartość 1e-4; parametry alpha inicjalizuje się wartościami z zakresu 0.05-1, a w opisie jest mowa o zakresie 0.1-0.9. Inicjalizacja parametrów rozkładów opisana jest w podrozdziale 2.3.4.1 w sposób mało zrozumiały, wymaga doprecyzowania. W podrozdziale 2.3.4.2 Autor twierdzi, że obliczenia gęstości prawdopodobieństw wymagają użycia skali logarytmicznej. Wymaga to wyjaśnienia. Czytelność podrozdziału 2.3.4.3 poprawiłyby odwołania do odpowiednich wzorów zamieszczonych we wcześniejszej części pracy. Opis szczegółów technicznych dotyczących radzenia sobie z bardzo małymi wartościami numerycznymi powinien być umieszczony w przypisach lub dodatkach, a nie w głównej treści rozprawy. Kryterium stopu podane w podrozdziale 2.3.4.4 jest nieprecyzyjne. W podrozdziale 2.3.6 Autor wraca do inicjalizacji – nie rozumiem dlaczego, skoro inicjalizacja została już opisana w podrozdziale 2.3.4.1.

Podrozdział 2.4 opisuje mieszaniny rozkładów wielomianowych. Do tego podrozdziału mam podobne uwagi jak do podrozdziału 2.3: niezrozumiały miejscami opis (np. inicjalizacji, kroku E, danych wejściowych (2.55), prawdopodobieństw (2.57)) oraz brak objaśnień pseudokodu (Algorytm 2.2) i brak odniesień do niego w tekście.

Rozdział trzeci omawia algorytmy grupowania oparte na odległościach. Definiując standardyzację w podrozdziale 3.1.2.3, niepotrzebnie przytacza się powszechnie znane wzory na wartość średnią i odchylenie standardowe. W tym pierwszym wzorze popełniono zresztą błąd: zamiast x_i powinno być x_n. We wzorze (3.3), wyrażającym standardyzację, zamiast z_n powinno być z_i. Nie rozumiem dlaczego podrozdział 3.1.3 Autor zatytułował „Odległości statystyczne", sugerując, że będzie mierzył odległości pomiędzy obiektami statystycznymi, np. rozkładami prawdopodobieństwa. Tymczasem opisuje klasyczne miary odległości pomiędzy dwoma punktami. Opis warunków definiujących odległość jako metrykę, który znajduje się pod tabelą 3.1, pozostawia wiele do życzenia. Oznaczenia w nim występujące nie odpowiadają tym w tabeli. Ponadto zawiera pleonazmy w stylu „odległości są zawsze nieujemne, co oznacza, że nigdy nie mogą być wartościami ujemnymi" lub „odległość od obiektu do samego siebie jest zawsze równa 0, ponieważ obiekt jest zawsze w tej samej odległości od siebie". Definicja odległości euklidesowej w podrozdziale 3.1.3.1 nie wymaga sięgania po twierdzenie

Pitagorasa, tak jak to uczynił Autor. Opis miar odległości powinien być bardziej profesjonalny, jak przystało na rozprawę doktorska. Większego profesjonalizmu oczekiwałbym także od opisu przedstawionych w kolejnych podrozdziałach metod grupowania. Opis matematyczny jest niespójny i miejscami niejasny, używa się różnych symboli do wyrażenia tych samych wielkości, pewne symbole i skróty są nieobjaśnione. Podrozdział 3.3.2 ucina się w pół zdania, nie wyjaśniając różnic pomiędzy algorytmem Hartigana-Wonga i k-means++. Porównanie metod k-średnich i k-medoidów jest dyskusyjne. Na przykład dlaczego liczba punktów centralnych reprezentujących grupy w k-medoidach ma być dużo mniejsza niż w przypadku k-średnich? Dlaczego k-średnich ma preferować metrykę euklidesową, a k-medoidów miejską? Dlaczego Autor uważa, że metoda k-medoidów jest rozszerzeniem metody k-średnich?

Rozdział czwarty, przedstawiający procedurę badań eksperymentalnych, rozpoczyna się czytelnym schematem obrazującym kolejne etapy badań. W początkowej części rozdziału Autor opisuje tworzenie danych symulacyjnych. Brakuje w tym opisie podsumowania utworzonych zbiorów danych w czytelnej formie i ewentualnej wizualizacja wybranych zbiorów z wykorzystaniem metod redukcji wymiaru. Taka wizualizacja pozwoliłaby się zorientować czy utworzone dane formują skupiska. Nie podano liczby punktów w utworzonych zbiorach. Nie objaśniono na czym polega filtracja używana do generowania danych. W podrozdziale 4.1.2, gdzie przedstawiono dane rzeczywiste używane w eksperymentach, brakuje szerszej charakterystyki zbiorów danych. Nie podano nawet tak podstawowych informacji jak rozmiary tych zbiorów, liczba klas i proporcje klas. W podrozdziale 4.2.1 opisano przekształcenia tych zbiorów danych i utworzenie na ich bazie dodatkowych zbiorów. Niestety opis ten jest niezrozumiały. Niezrozumiałe też jest filtrowanie danych omówione w podrozdziale 4.3. Czy wariancję wyznacza się dla zmiennych (cech), czy obserwacji (symbolem n oznaczono liczbę obserwacji)? W jaki sposób do wyznaczenia końcowej liczby grup używa się kryterium BIC?

W podrozdziale 4.5 opisano miary oceny grupowania, które stosuje się w części eksperymentalnej. Wszystkie wymienione miary wymagają znajomości poprawnego grupowania (*ground truth*) do oceny wyników grupowania przeprowadzonego przez algorytm. W rzeczywistych przypadkach takie poprawne grupowanie jest nieznane. Rodzi to pytanie na ile porównanie różnych algorytmów grupowania w części eksperymentalnej pracy przy wykorzystaniu tych miar jest uzasadnione? Dla zbiorów danych rzeczywistych Autor przyjmuje, że liczba grup równa jest liczbie klas (wszystkie zbiory danych rzeczywistych zawierają etykiety klas, co jest raczej nietypowe w problemach grupowania) i dąży do takiego pogrupowania danych, aby grupy pokrywały się z obszarami klas (wtedy przyjęte miary grupowania będą miały najkorzystniejsze wartości). Ale przecież nie można zakładać, że liczba skupisk w danych będzie równa liczbie klas, i że rozkład tych skupisk będzie odpowiadał rozkładowi klas. Nie można też zakładać, że dane będą miały etykiety klas.

Do porównania grupowania przeprowadzonego przez daną metodę z grupowaniem poprawnym używa się algorytmu węgierskiego. Jednak to zagadnienie nie jest dostatecznie jasno opisane. Algorytm węgierski wymaga, aby liczba grup przyjęta przez algorytm była równa liczbie grup docelowych. A co w sytuacjach, gdy liczba grup docelowych jest nieznana? Opis ważonego wskaźnika Jaccarda przedstawiony w podrozdziale 4.5.2.2 jest niezrozumiały. Brak omówienia wzoru (4.4), który wyraża wskaźnik ARI, podstawową miarę oceny grupowania w części eksperymentalnej. W jaki sposób wyznacza się dokładność i zbilansowaną dokładność przedstawione w podrozdziale 4.5.2.3 dla przypadku wieloklasowego? Omówiono tylko dwuklasowy. W jaki sposób wykorzystuje się rozkłady beta-dwumianowe (podrozdz. 4.5.2.6) do oceny grupowania? Sugeruję Autorowi podawanie przykładów i wizualizacji do wyjaśnienia trudniejszych kwestii.

W obszernym rozdziale piątym przedstawiono wyniki badań eksperymentalnych, porównujących metody oparte na mieszaninach rozkładów z innymi metodami omówionymi w rozdziale trzecim. Na pochwałę zasługuje ciekawy sposób prezentacji wyników w postaci wykresów i diagramów oraz bogaty komentarz Autora, choć nie zawsze wyczerpujący. Dane rzeczywiste są gruntownie opisane i zwizualizowane przy użyciu różnych technik redukcji wymiarowości. W kontekście zastrzeżeń co do definicji miar oceny grupowania poczynionych powyżej, pojawiają się pytania odnośnie ogromnego zróżnicowania wyników różnych metod oraz tej samej metody dla różnych sposobów wstępnego przetwarzania danych. Widać to szczególnie na wykresach wskaźnika ARI i prawdopodobieństwa poprawnego przypisania na podstawie rozkładów beta-dwumianowych. Na przykład na rys. 5.1 wskaźnik ARI dla metody k-medoidów przyjmuje wartości bliskie zeru, natomiast dla metody GMM (mieszanin gaussowskich) osiąga dla niektórych przypadków wartości 0,5 i większe (na skali od 0 do 1). Nawet zbliżone do siebie algorytmy k-medoidów i k-średnich mają tak zróżnicowane wyniki. Ciekawe, że zróżnicowanie pomiędzy algorytmami jest znacznie mniejsze, gdy używa się innych miar do oceny grupowania, np. mediany zbalansowanej dokładności (rys. 5.3) oraz WSMC, SMC i WJACC (rys. 5.4). Jak Autor to wytłumaczy? Na rys. 5.6, dla danych kompletnych i zredukowanych, metoda MMM (mieszaniny rozkładów wielomianowych) znajduje się wśród najwyżej notowanych wskaźnikiem ARI, podczas gdy dla danych skalowanych osiąga najniższe noty. Jej ocena zmienia się drastycznie od wartości bliskich jedynce do wartości bliskich zeru. Podobne duże zmiany, ale w odwrotnym kierunku, obserwuje się dla GMM. Jest to tym bardziej ciekawe, że pozostałe metody grupowania nie wykazują takiej wrażliwości na skalowanie. Dlaczego? Podobne dysproporcje w wynikach obserwuje się dla danych rzeczywistych. Na przykład rys. 5.12 pokazuje znaczne zróżnicowanie wyników różnych metod, a rys. 5.14 niewielkie zróżnicowanie. Wymagana jest dyskusja wyników w tym aspekcie.

W części eksperymentalnej, Autor porównuje wyniki grupowania dla każdego zbioru danych, używając zwykle sześciu algorytmów grupowania spośród jedenastu, które zapowiedział w podrozdziale 4.4 (tab. 4.2). Zestaw tych sześciu algorytmów zależny jest od zbioru danych. Wymaga to wyjaśnienia. Dla danych NASA Kepplers nie przedstawiono kompletu wyników.

W rozdziale szóstym podsumowano wyniki w zwartej formie, pokazując, że proponowane w pracy algorytmy zapewniają lepsze rezultaty grupowania niż metody porównawcze zarówno dla danych syntetycznych, jak i rzeczywistych. W konkluzji Autor podsumowuje implementacje swoich algorytmów w języku R, omawia zakres badań i kończy stwierdzeniem, że proponowane metody grupowania są konkurencyjne w stosunku do metod opartych na odległościach. Zabrakło w tych wnioskach odniesienia do pierwszej i trzeciej tezy rozprawy, wskazania zalet i wad opracowanych metod grupowania oraz oceny ich złożoności obliczeniowej. Zabrakło także porównania ze znanymi implementacjami metod grupowania opartych na rozkładach. Dlaczego implementacja utworzona w ramach pracy ma być lepsza od już istniejących?

Układ rozprawy ma braki – powinien zostać uzupełniony o opis problemu grupowania i przegląd literatury. Podział treści wymaga poprawy, jest zbyt rozdrobniony na podrozdziały. Język jakim napisana jest rozprawa jest miejscami niestaranny i niezrozumiały. Bibliografia powinna być bardziej obszerna, biorąc pod uwagę dynamiczny rozwój tego obszaru badań.

Praca ma sporo braków i wad, które wskazałem powyżej i w obecnej formie nie spełnia kryteriów oceny prac doktorskich. Praca powinna zostać uzupełniona o opis problemu i przegląd literatury. Poprawy wymagają następujące elementy: tezy, układ redakcyjny i język, opisy algorytmów i mierników oceny, interpretacja wyników oraz wnioski, w których należy odnieść się do tez i elementów nowości pracy.

Uwagi językowe, edytorskie i redakcyjne

Praca zawiera dość dużą liczbę błędów językowych i edytorskich. Powinna być przez Autora starannie przejrzana i poprawiona. Nie sposób wymienić wszystkie błędy. Poniżej wymieniam tylko niektóre.

- W wielu miejscach Autor używa terminu *„unsupervised clustering"*. Typowo grupowanie jest problemem uczenia bez nadzoru, nie trzeba tego za każdym razem podkreślać.
- Lista symboli i opis wzoru (2.44): symbol „exp" objaśniony jest błędnie jako *„exponentiation"* (potegowanie). Powinno być *„exponential function"*.
- Symbole zmiennych w wielu miejscach zapisane są czcionką prostą zamiast kursywą, np. str. 7, 11, 47, ...
- Symbole wektorów i macierzy powinny być pisane czcionką prostą i pogrubioną, aby odróżnić je od skalarów. Autor generalnie nie stosuje pogrubienia, pojawia się ono sporadycznie, np. w podrozdziale 3.4. Zdarza się, że we wzorze występuje symbol bez pogrubienia, a w opisie wzoru – pogrubiony, np. (3.5)-(3.7), (3.14), (3.15).
- Pod wzorami słowo *„where"* pisane jest niepotrzebnie pogrubioną czcionką, np. (2.1).
- W wielu miejscach przed nawiasami kwadratowymi z numerami cytowanych pozycji literaturowych brakuje spacji, np. str. 6, 7, 12, ...
- Wykładnik w zapisie funkcji wykładniczej symbolem *„exp"* podaje się w nawiasie za tym symbolem, tak jak we wzorze (2.44). Błędny zapis występuje we wzorach (2.1)-(2.3).
- Powtórzenie *„of"* w opisie zmiennych pod wzorem (2.9).
- Przed wzorem (2.10) powinno być objaśnienie, co ten wzór wyraża.
- W podrozdziale 2.3.4.1 Autor używa pierwszej osoby liczby pojedynczej (np. *I created*), co jest niewłaściwe dla prac naukowych i niespójne z pozostałą częścią rozprawy.
- Zapis funkcji Log-Sum-Exp wzorem (2.44) jest niepoprawny – problem z rozmieszczeniem nawiasów. Opis pod tym wzorem definiuje x_M jako obserwację, nic natomiast nie mówi o symbolach x_1, x_2, ... występujących w tym wzorze.
- W rozdziale 2, w opisach wzorów występują liczne powtórzenia opisów tych samych symboli. Dany symbol opisuje się tylko przy pierwszym wystąpieniu.
- Pod wzorem (2.45) zabrakło dodatkowego warunku, że udziały rozkładów w mieszaninie α sumują się do jedynki.
- Wzór (2.51) nie wyraża standardyzacji, jak napisał Autor, lecz ma na celu przeskalowanie wartości liczników, tak aby ich suma dla wszystkich komponentów mieszaniny była równa jeden. Podobnie błędne użycie terminu standardyzacja występuje m.in. na str. 27, 28 i 45.
- Str. 29: brak nawiasów wokół numerów wzorów.
- Rozdział 3 i 4: oznaczenia tych samych wielkości w różnych częściach rozprawy powinny być takie same. W rozdziale 3 i 4 stosuje się inne oznaczenia dla liczby obserwacji i liczby atrybutów/cech/zmiennych niż we wcześniejszej części pracy, co wprowadza niepotrzebne zamieszanie.
- W podrozdziale 3.1.1, opisując struktury danych Autor użył niezrozumiałej dla mnie definicji: *„an objects-by-variables matrix is two-mode because rows and columns differ."* W jakim sensie wiersze i kolumny się różnią?
- W podrozdziale 3.1.3.1 *„Minkowski distance"* powtarza się zdanie *„Minkowski distance is a generalization of Euclidean and Manhattan distance"*.
- W podrozdziale 3.2 na str. 35 występują niezrozumiałe oznaczenia *„pic!"*.
- Str. 36: powinno być *„spliter"* zamiast *„splinter"*.
- We wzorze (3.11) występują nieobjaśnione symbole.

- Str. 41: skrót PAM jest nieobjaśniony.
- Jaki sens ma podrozdział 4.3.1, mówiący o tym, że dane nie podlegają filtrowaniu, skoro wyjaśniono to w akapicie poprzedzającym ten podrozdział?
- Wykresy wskaźnika ARI przedstawione w rozdziale 5 w kilku przypadkach są nieczytelne – rys. 5.12 i 5.24.

Wniosek końcowy

Opiniowana rozprawa doktorska mgr Mateusza Kani w obecnej formie nie spełnia wymogów ustawy z 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce. Praca powinna zostać uzupełniona i poprawiona w zakresie wskazanym w niniejszej recenzji.