Prof. dr hab. inż. Zbigniew Nahorski
Instytut Badań Systemowych PAN

Warszawa, 29 lipca 2019 r.

Recenzja
rozprawy doktorskiej mgr. Macieja Jankowskiego
p.t. *Dimension Reduction Methods in Text Classification with Probabilistic Graphical Models*
przygotowanej pod kierunkiem prof. dr. hab. inż. Mariana Chudego

Recenzję wykonano na zamówienie Przewodniczącego Rady Wydziału Cybernetyki Wojskowej Akademii Technicznej dr. hab. inż. Kazimierza Worwy, prof. WAT z dn. 12 czerwca 2019 r.

1. Uwagi wstępne

Rozprawa jest poświęcona redukcji wymiarowości w opisach dokumentów tekstowych w celu ich efektywnego przetwarzania, na przykład w klasyfikacji dokumentów. Ze względu na używane do tego celu narzędzia, tematyka ta wchodzi w zakres takich obszarów, jak sztuczna inteligencja czy uczenie maszynowe. Doktorant stosuje do redukcji metody wywodzące się ze statystyki.

W rozprawie nie sformułowano tez (celów) rozprawy, jednak z wprowadzenia łatwo wywnioskować, że celem jest poprawienie znanych oraz opracowanie nowych metod wyboru ukrytej reprezentacji dokumentów tekstowych, o znacznie mniejszej wymiarowości niż liczba wszystkich słów (termów) użytych w dokumencie, natomiast wystarczającej do dalszego skutecznego przetwarzania komputerowego tych dokumentów.

Tematyka związana z przetwarzaniem tekstów jest jednym z bardziej rozwijanych zagadnień we współczesnej informatyce, a metody poszukiwania ukrytej reprezentacji dokumentów są przedmiotem intensywnych badań. Na tym tle zaproponowane przez Doktoranta nowe lub ulepszone metody wprowadzają istotny postęp w stosunku do znanych wcześniej metod.

2. Zawartość rozprawy i uwagi ogólne

Rozprawa zajmuje 93 strony, a jej główna część składa się z 5 rozdziałów, przy czym pierwszy z nich zawiera wiadomości wprowadzające, a piąty podsumowanie. Podstawowa część rozprawy obejmuje rozdziały 2, 3 i 4, które są oparte na trzech publikacjach Doktoranta. Tematycznie są one jednolite i przedstawiają różne podejścia do redukcji wymiarowości. Natomiast połączenie kilku publikacji spowodowało kłopoty z ujednoliceniem oznaczeń i terminologii, co nie do końca udało się Doktorantowi rozwiązać w sposób satysfakcjonujący.
W rozdziale 1 są wprowadzone podstawowe pojęcia używane w piśmiennictwie dotyczącym przetwarzania tekstów oraz główne metody stosowane do redukcji wymiarowości, a także pomocnicze przybliżone metody stosowane w algorytmach obliczeniowych. Są to przydatne wiadomości i trochę nawet szkoda, że nie zostały one przedstawione dokładniej. W szczególności, Doktorant wprowadza tu dwa używane w dalszych rozdziałach algorytmy ustalania tematów (topics) dokumentów, czyli ukrytych zmiennych, które dobrze charakteryzują dokument i reprezentują ten dokument w dalszym przetwarzaniu, na przykład w klasyfikacji dokumentów. Jest to dosyć schematyczny algorytm Latent Semantic Analysis (LSA) oparty na faktoryzacji macierzy wyrazów (termów) dokumentu oraz znacznie bardziej rozwinięty algorytm Latent Dirichlet Allocation (LDA) oparty na analizie bayesowskiej.

Algorytmy LDA, jak i zasadnicze algorytmy od nich pochodne, wymagają do przeprowadzenia obliczeń znajomości liczności zbioru poszukiwanych tematów.

Rozdział 2, oparty na publikacji Doktoranta [29], składa się tematycznie z dwóch części. Pierwsza jest poświęcona ustalaniu optymalnej liczby tematów (topics) dokumentów w metodzie LDA, a druga użyciu do klasyfikacji zespołów klasyfikatorów, dla zestawów tematów o różnych licznościach. Doktorant przedstawił w pierwszej części kilka znanych metod ustalania liczby tematów opisanych w literaturze, a następnie zaproponował zastosowanie do tego celu nowej metody opartej na entropii. Do wstępnego wyboru liczby tematów zaproponował też użycie metody LSA, znacznie szybszej od metody LDA, i na przykładzie wykazał, że ten sposób może być przydatny do wyboru liczby tematów w metodzie LDA lub ewentualnie do wstępnego ograniczenia zakresu poszukiwań do niewielkiego zbioru liczb tematów z otoczenia liczby wybranej.

Następnie zaproponował w drugiej części, zamiast wyboru jednej ustalonej liczby tematów, zastosowanie zespołów klasyfikatorów, podejścia stosowanego z powodzeniem w innych zadaniach decyzyjnych. Na podstawie wskazań uzyskanych z różnych modeli (czyli algorytmu ustalania zbioru tematów przy ustalonej ich liczbie), klasyfikator wybiera dla każdego dokumentu modele maksymalizujące tzw. funkcję pewności (confidence function). Doktorant rozważył trzy funkcje pewności, oparte na: głosowaniu większościowym, ważonym głosowaniu większościowym oraz mieszaniu (perplexity). Obliczenia wykonane na dwóch znanych korpusach dokumentów wykazały, że metody oparte na zespołach klasyfikatorów radzą sobie lepiej niż najlepszy pojedynczy model z ustaloną liczbą tematów. W obu częściach Doktorant zaproponował nowe podejścia oparte na metodach znanych w innych zastosowaniach i zaadaptowanych przez Doktoranta do rozpatrywanego zadania. Do pełnego porównania metod przydałoby się jednak, oprócz dokładności, porównać także złożoność obliczeniową różnych metod.

W rozdziale 3, opartym na publikacji Doktoranta [28], Doktorant rozwija zastosowanie metod opartych na zespołach klasyfikatorów, rozważając modele Multi-class Supervised Latent Dirichlet Allocation (Multi-class sLDA), czyli rozszerzenie modeli LDA, w których używa się uczenia nadzorowanego na zbiorze dokumentów z przydzielonymi klasami. Pomaga to w dokładniejszym przyporządkowaniu dokumentów do klas oraz ograniczeniu się do klas, na których nam zależy. W metodzie tej w fazie uczenia poszukuje się parametrów rozkładów prawdopodobieństwa zmiennych występujących w modelach, tak aby uzyskać jak najlepszą zgodność klas uzyskanych z modelu z klasami wzorcowymi. Tak wyuczony model może posłużyć do estymacji klas dla dokumentów niesklasyfikowanych. W podrozdziale 3.3 Doktorant przedstawia tę bazową metodę korzystając z opisów literaturowych, koncentrując się przede wszystkim na fazie estymacji parametrów, w której używa znanego algorytmu EM (Expectation Maximization). W tym zastosowaniu algorytm ten wymaga jednak w kroku E aproksymacji funkcji wiarogodności, do czego zastosowano metodę Variational Inference, w której zakłada się niezależność pewnych rozkładów, uzyskując jawne ograniczenie dolne (evidence lower bound – ELBO), które maksymalizuje się w kroku M. Ten fragment przydałoby się opisać dokładniej, z lepszą motywacją słowną wprowadzanych wzorów oraz lepszym opisem wprowadzanych oznaczeń. W podrozdziale 3.4 Doktorant wprowadza własne rozszerzenie opisanej wcześniej metody Multi-class sLDA przez zastosowanie zespołów klasyfikatorów, aby uniknąć konieczności wcześniejszego ustalania liczby tematów, co jest konieczne w opisanej wcześniej metodzie. Do tego celu użył metody ważonego głosowania większościowego, w którym wagi są ustalane za pomocą znanego algorytmu AdaBoost, który w kolejnych krokach zmienia wagi tak, aby zmniejszyć liczbę złych klasyfikacji. W rezultacie Doktorant opracował nową metodę, nazwaną przez niego Boost Multi-class sLDA, która rozwiązuje pomysł użycia zespołów klasyfikatorów do modeli Multi-class sLDA. W porozdz. 3.5 przedstawiono eksperymenty numeryczne wykonane na zbiorach SMS Spam Collection Data Set z UCI Machine

Learning Repository oraz Poliblog. Wykazały one znacznie większe dokładności klasyfikacji za pomocą zaproponowanej metody w stosunku do metod nieużywających zespołów klasyfikatorów.

W rozdziale 4, opartym na publikacji Doktoranta [30], Doktorant rozważa modelowanie zmiennych ukrytych za pomocą metody Variational Autoencoder (VAE) opartej na zastosowaniu sieci neuronowych, z powodzeniem stosowanej przede wszystkim w analizie i generowaniu obrazów. Składa się ona z dwóch podsieci: kodującej, która ustala wartości zmiennych ukrytych na podstawie podanych przykładów dokumentów, i dekodującej, która znając wartości zmiennych ukrytych generuje dokumenty. Algorytm VAE w zastosowaniu do rozważanego zagadnienia redukcji opisu dokumentów jest opisany w podrozdz. 4.4, a jego wersja przeznaczona do uczenia nadzorowanego w podrozdz. 4.5. W tej drugiej wersji (Supervised Variational Autoencoder – SVAE) z każdym dokumentem jest związana zmienna określająca klasę, do której on należy.

Po opisie tych algorytmów Doktorant przystępuje w podrozdz. 4.6 i 4.7 do prezentacji swoich oryginalnych rozwiązań. Zasadzają się one na modelowaniu zmiennych ukrytych jako mieszaniny rozkładów gaussowskich, co pozwala na lepsze rozdzielenie tematów (utożsamianych później z klasami) przez rozmieszczenie rozkładów poszczególnych tematów promieniście i na ich przejrzystą wizualizację na płaszczyźnie dwuwymiarowej. Parametry rozkładów zależą od dodatkowej zmiennej, zwanej etykietą (label), która jest generowana losowo, zgodnie z ustalonym rozkładem wag w mieszaninie rozkładów gaussowskich. Etykieta wpływa na rozróżnienie poszczególnych rozkładów. W pierwszej metodzie, nazwanej przez Doktoranta Deep Latent Gaussian Mixture Model (DLGMM) zmienna ta jest ustalana przez algorytm w sposób przypadkowy, co powoduje problemy, gdy klasy (tematy) okazują się do siebie podobne i wartości zmiennych ukrytych uzyskane dla dokumentów zaczynają się istotnie mieszać. W podrozdz. 4.7 zaproponowano rozwiązanie tych problemów używając etykiet w uczeniu nadzorowanym, co pomogło w wyeliminowaniu zauważonych problemów i doprowadziło do nowej metody nazwanej przez Doktoranta Deep Supervised Latent Gaussian Mixture Model (DSLGMM). Przeprowadzone w dalszych podrozdziałach eksperymenty z kilkoma zbiorami danych wykazały bardzo dobre działanie zaproponowanych przez Doktoranta metod zarówno pod względem dokładności klasyfikacji jak i szybkości obliczeń, w szczególności za pomocą modelu DSLGMM. Warto dodać, że Doktorant opracował matematycznie zaproponowane nowe metody oraz utworzył algorytmy komputerowe pozwalające na przeprowadzenie tych obliczeń, wykazując się znajomością stosowanych technik oraz biegłością zarówno w stosowaniu aparatu matematycznego, jak i nowoczesnych narzędzi informatycznych.

Rozdział 5 zawiera krótkie podsumowanie wykonanych badań.

Podkreślając duży wkład Doktoranta w algorytmizację wyznaczania zmiennych ukrytych charakteryzujących dokumenty tekstowe oraz wprowadzenie przez niego nowych metod, trzeba dodać, że rozprawa wygląda słabiej pod względem dokumentacji tych osiągnięć. Definicje używanych pojęć są często wprowadzane później niż w miejscu ich pierwszego użycia. Oznaczenia są zmieniane w poszczególnych rozdziałach i nie do końca bywają one wyjaśnione. Sformułowania są często nieprecyzyjne lub brakuje w nich dyskusji różnych przypadków. Przy wyprowadzaniu zależności matematycznych za mało jest słownych objaśnień, co znacznie utrudnia śledzenie rozumowania, szczególnie gdy we wzorach pojawiają się błędy, chociażby literowe. Te braki w opisie znacznie obniżają wartość rozprawy.

3. Uwagi szczegółowe

Do rozprawy nie dołączono streszczenia w języku polskim, ale zostało ono dosłane mejlowo w późniejszym terminie.

$9^{3\text{-}41}$ – Stwierdza się tu, że wiersz macierzy, którego podobieństwo kosinusowe do nowego przykładu jest największe, jednoznacznie wskazuje na klasę przykładu. Czy nie może się zdarzyć, że podobieństwo dwóch wierszy do nowego przykładu jest jednakowe? Jak wtedy wskazać jednoznacznie na klasę? Przy okazji, sama definicja podobieństwa kosinusowego została wprowadzona dopiero na str. 20.

10^{4} – W punkcie 1.2.1 ukryte zmienne, o znacznie niższej wymiarowości, których poszukuje się do skróconego opisu dokumentu, są nazywane „features". Tu nazywa się je „concepts", a zaraz dalej, w punkcie 1.3.2 „topics". Natomiast w rozdziale 4 wprowadza się jeszcze nazwy „hidden variables" (w p. 4.2) oraz „latent variables". Wydaje mi się, że wszystkie te zmienne pełnią tę samą rolę, a różne nazwy biorą się z tego, że są one wprowadzone w różnych publikacjach i ewentualnie przy rozważaniu różnych metod. Czytelność rozprawy znacznie by się poprawiła, gdyby ograniczyć różnorodność terminologii do niezbędnej, a do tego wprowadzić wyjaśnienia, skąd się biorą i jakie mają znaczenie te różnice w terminologii.

10_{12} – Jeżeli dobrze rozumiem, to *X'V* oznacza mnożenie dwóch wprowadzonych wcześniej macierzy. Skąd wiadomo, że mnożenie to jest dobrze określone (chodzi o zgodność odpowiednich wymiarów macierzy).

$10_{4\text{-}3}$ – Co to są parametry θ? Z rys. 1.3 można się domyślić, że chodzi o rozkłady tematów (topics) w dokumentach. Dlaczego tego nie objaśniono przy wprowadzeniu oznaczenia wektora θ?

$10_{3\text{-}1}$ – W tym przykładzie podaje się, że w elemencie β_1 wektora β zapamiętuje się parametry rozkładu wielomianowego, a więc zgodnie z definicją ze str. 15, liczbę *N* oraz wektor τ. Jak to się ma do wcześniejszej definicji wektora β oraz rys. 1.2, na którym elementami wektora β są częstości występowania słów (czyli termów) w tematach (topics).

$12_{9\text{-}6}$ – W definicji rozkładu Dirichleta z definicji 1.4.5 występuje parametr α (ale nie ma η), przy czym jest on zdefiniowany jako wektor, który, tak samo jak wektor θ, ma składowe liczbowe rzeczywiste. Natomiast tu napisano, że α i η są rozkładami. Brakuje wyjaśnienia, jak to interpretować.

13, rys. 1.5 – W algorytmie wybór $\beta_{z_n^{(d)}}$ zależy od $z_n^{(d)}$. Czy w związku z tym na schemacie nie powinno być strzałki od kółeczka *z* do kółeczka β?

$13^{1\text{-}6}$ – Rozkłady i użyte tu oznaczenia wprowadzono dopiero w następnym podpunkcie. A tu brakuje o tym nawet informacji.

16 – W obu wydzielonych wzorach górna granica w sumach powinna być *D* zamiast *M*.

17^{5} – Notacja użyta w zapisie $\mathbb{E}_{p\sim p(z|w)}\, p(w, z|\theta)$, często stosowana także dalej, powinna być objaśniona.

$19^{14\text{-}16}$ – Nie jest tu, co prawda, podana formalna definicja podejścia największej wiarogodności, jednak to sformułowanie jest zbyt dalekie od prawdy. Po pierwsze, w metodzie największej wiarogodności rozważa się funkcję wiarogodności utworzoną na podstawie prawdopodobieństwa warunkowego względem poszukiwanych parametrów, a nie po prostu prawdopodobieństwa. A do tego jest tak tylko dla zmiennych dyskretnych, bo dla zmiennych ciągłych rozważa się warunkową gęstość

---

[1] Zarówno tu, jak i w dalszym ciągu recenzji p^l oznacza l-ty wiersz od góry na stronie p, zaś p_l oznacza l-ty wiersz od dołu na stronie p.

prawdopodobieństwa, a nie warunkowe prawdopodobieństwo. Po drugie, za estymator największej wiarogodności przyjmuje się wartość maksymalizującą funkcję wiarogodności. Nie musi ona być na ogół równa wartości prawdziwej. Zachodzi to dopiero asymptotycznie ze wzrostem długości próby, a i to przy spełnieniu pewnych założeń.

19_4 – Chodzi tu o „topic distributions", a nie o „topic densities".

20_7 – Oznaczenie *KL* jest objaśnione dopiero na str. 50.

22, rys. 2.3 – Przy metodach wymienionych w legendzie, a przynajmniej tych, o których nie wspomina się w tekście, warto by było podać odpowiednią pozycję literatury.

23, rys. 2.4 – Podpis pod rysunkiem jest niezrozumiały. Co to znaczy „dokładność klasyfikacji ze średnią entropią"?

23-24 – Motywacja podejścia zespołowego (ensamble) jest niezrozumiała. Na początku twierdzi się, że chodzi o wybór *K* dla całego korpusu, ale dalej rozważa się tylko klasyfikację dokumentów i wybór dla nich najlepszego *K*. Prócz tego, w podrozdziale 2.2 położono nacisk na szybszy wybór *K* za pomocą LSA, gdy tutaj wraca się do przeglądu modeli LDA. Ten fragment przydałoby się rozwinąć i dokładniej opisać, o co chodzi w podejściu zespołowym.

24, rys. 2.5 – Skrót RF nie został objaśniony.

26^{11} – Wcześniej przez model rozumiano algorytm, głównie LDA, z ustaloną liczbą tematów *K*. Natomiast tutaj jest wprowadzone pojęcie modelu $\mathcal{M}_l$, przy czym na dole strony 24 zdefiniowano $\mathcal{M}_l$ jako klasyfikator. Czym jest wobec tego tutaj model $\mathcal{M}_l$?

29_{9-7} – Przybliżony wybór *K* za pomocą modeli LSA został poparty tylko jednym przykładem, więc trzeba ostrożnie podejść do jego ogólności.

33, rys. 3.1 – Czy tu nie powinno być strzałki od kółeczka *z* do kółeczka β, z taką motywacją, jak dla rys. 1.5?

34, wzór (3.2) – W wyrażeniu $q(\theta|\gamma)$ brakuje indeksu górnego $^{(d)}$ przy θ.

34, wzór (3.3) – Nie jest zrozumiałe, skąd się wziął ten wzór. Można by było przypuszczać, że jest to po prostu definicja ELBO, ale wydaje się, że wcześniej jest próba wytłumaczenia lub motywacji tego wzoru, ale niezrozumiała. Jeżeli $\mathbb{E}_q$ jest operatorem wartości oczekiwanej względem zmiennej *z*, to po co obliczać wartość oczekiwaną pierwszego składnika po prawej stronie, który nie zależy od *z*.

35 – Wyprowadzenia wzorów na tej stronie są praktycznie niemożliwe do sprawdzenia. Jak są zdefiniowane funkcje Γ i Ψ? Czy $\gamma_{d,i}$ we wzorach (3.3) i (3.4) to to samo, co γ_{di} w poprzednim wzorze, a $\phi_{d,n,i}$ we wzorze (3.4) to to samo, co ϕ_{dni} we wzorze następnym? W ostatnim składniku w trzecim wierszu od góry sumowanie powinno być po *k*, a nie po *i*.

36, wzór (3.7) – Jak zależy prawa strona od *k*? Czy należy rozumieć to tak, że dla ustalonego *i* wartość $\beta_{k,i}$ jest taka sama dla dowolnego *k*?

36, wzór powyżej (3.8) – Górną granicą w głównej sumie jest w innych wzorach *M*, a tu *D*. Trzeba to ujednolicić.

36, wzór (3.8) – Zgodnie ze stosowanymi oznaczeniami w oznaczeniu pochodnych powinno być η_{c_i}, a nie η_{ci}. Prócz tego, zamiast I powinna być podwojona jedynka, tak jak w innych wzorach.

36, ostatni wzór na stronie – Wbrew temu, co napisano wyżej, tutaj nie wybiera się klasy o największym prawdopodobieństwie, ale klasę o największej wartości oczekiwanej rozpatrywanej we wzorze wielkości.

37, koniec punktu 3.3.4 – ten fragment rozumowania wymaga dokładniejszego objaśnienia.

44, tabl. 3.1 – Brak opisu, co oznaczają wartości w kolumnach 2 i 3. Można się tego oczywiście łatwo domyślić śledząc tekst, jednak powinno to być podane albo w nagłówku kolumn, albo w tytule tablicy.

47_4 – Skrót pdf jest powszechnie używany zamiast pełnej nazwy „probability distribution function", czyli dystrybuanty. Czy o to tu chodzi?

49^8 – Niezdefiniowany skrót VEM.

49_3 – Tak generowane wartości $z^{(d)}$ nie zależą bezpośrednio od parametru θ, co jest niezgodne ze schematem z rys. 4.2. Warto by było to wyjaśnić

50^3 – Co to jest „KL-divergence", jest zdefiniowane dopiero w punkcie 4.4.1, a i to brakuje tam słowa „divergence".

$50_{5\text{-}4}$ – Chodzi chyba o to, że zmienna z ma rozkład $\mathcal{N}(0, I)$? Czy nie lepiej tak napisać, zamiast wprowadzać pojęcie rozkładu izotropowego, który chyba nie jest zbyt powszechnie używany?

50, wzór (4.2) – Po prawej stronie są oznaczenia Σ, n, μ_1, które nie są objaśnione. Po zajrzeniu do pozycji [18] wygląda na to, że w rozważanym w rozprawie przypadku $\Sigma = \sigma^2 I$, a to oznacza, że wzór ten można łatwo jeszcze uprościć, bo $\log\det(\sigma^2 I) = n \log\sigma^2$, a $\mathrm{tr}(\sigma^2 I) = n\sigma^2$, gdzie n jest wymiarem macierzy jednostkowej I.

51^1 – Co oznacza skrót SGVB?

51^2 – Co oznacza skrót SGD?

51^{13} – Przy zamianie zmiennej pod wartością oczekiwaną (całką) z z na ϵ trzeba chyba uwzględnić jakobian?

52, rys. 4.3 – Które wielkości tworzą tutaj parametr θ występujący w ostatniej warstwie sieci neuronowej?

$52_{2\text{-}1}$ – Tutaj twierdzi się, że rozkład *a posteriori* zmiennej z faktycznie zależy od parametru θ, ale jest to zależność pośrednia, a schemat na rys. 4.2 wskazuje na zależność bezpośrednią. Jak to wytłumaczyć?

54_4 – Na rys. 4.5 nie ma x. Wygląda na to, że tu i dalej używa się dwóch oznaczeń w tym samym znaczeniu: w i x, bez wyjaśnienia, czy jest między nimi jakaś różnica.

54_4 – Wyrażenie $x \perp y \mid z$ jest niejednoznaczne, bo zarówno zapis $(x \perp y) \mid z$ jak i $x \perp (y \mid z)$ ma sens.

55^8 – Powinno być „$\mathbb{E}_{z\sim q_\phi}$".

56^9 – Czy wagi mogą być dowolne, czy też muszą być nieujemne?

56_{10} – O jaki obraz (image) tu chodzi?

56_9 – Czy przy równych wartościach największych przyjmujemy jeden z odpowiadających im indeksów losowo?

56_8 – Co oznacza indeks R na końcu wiersza?

57, rys. 4.7 – Tutaj w opisie ostatniej warstwy sieci neuronowej nie występuje jawnie parametr θ, ale jest on wymieniony w definicji funkcji b. Podobnie więc, jak w rys. 4.3, zachodzi pytanie, które wielkości tworzą ten parametr.

58, wzór (4.4) – Czy liczba rozkładów gaussowskich K oznacza, że jest ich dokładnie tyle, co liczba tematów (topics)? Jeżeli tak, to warto by o tym napisać.

60, rys. 4.10 – Jak należy rozumieć to, że zmienne ukryte (s_d, $z^{(d)}$) są bezpośrednio obserwowane? Przy okazji, dalej są używane także oznaczenia $s^{(d)}$.

62, algorytm 4 – Wartości wektora o są w drugim kroku ustalone na zero, a dalej nie są zmieniane. Można więc ich nie wprowadzać, a odpowiednie wielkości we wzorach pominąć.

62, algorytm 4 – Zmienna Σ występująca w kroku 3 (d) nie została wcześniej określona.

67_2 – O jakie „data-points" tu chodzi?

67_1 – Jeżeli jest to funkcja wiarogodności dla M obserwacji, to dlaczego nie zależy ona od M?

68^2 – Logarytm po prawej stronie wzoru nie jest logarytmem prawej strony poprzedniego wzoru (na funkcję wiarogodności).

68^3 – Chyba s_k jest wielkością występującą na k-tej pozycji K-wymiarowej zmiennej s, a nie k-tą wartością zmiennej s?

68^5 – Co to jest Σ? Może chodzi o Σ_k?

68, wzór (4.8) – Dlaczego wśród argumentów funkcji celu nie umieszczono z? Jakie zmienne wchodzą w skład wektorów parametrów θ i ϕ?

68, wzór (4.10) – Jak ta funkcja po prawej stronie zależy od ϕ? We wzorze tej zależności nie widać i dopiero w algorytmie 5 można prześledzić, jak następuje minimalizacja tej funkcji po ϕ.

72^8 – Nazwa „confidence of classification" jest niezbyt fortunna. W statystyce są dobrze ugruntowane takie pojęcia jak "confidence interval" czy "confidence level" i przyjęta nazwa od razu kojarzy się z tymi pojęciami statystycznymi. Przydało by się też wyjaśnić, jak interpretować dużą lub małą wartość $||z||$.

73, tabl. 4.1 – Wyjaśnienie (jedna iteracja) powinno być sformułowane w języku angielskim.

74, tabl. 4.3 – Zbiór SmsSpam został pominięty w opisie w p. 4.10.1. Czy rzeczywiści zawiera on tylko 16 elementów?

75, koniec p. 4.10.2 – Jakie wnioski wynikają z tych obliczeń?

78, tabl. 4.4 – Na rys. 4.19 wartość progu (threshold) mieści się w przedziale [0,1], a w tablicy są wartości większe od 1. Jak to rozumieć?

80, rys. 4.21 – Co oznaczają kolorowe skale po prawej stronie rysunków? Jak je interpretować znaczeniowo?

86, tabl. 4.10 – W siódmym wierszu i drugiej kolumnie powinno być $\sum_{k=1}^{K} s_k^{(d)} = 1, p\left(s_k^{(d)}\right) = \pi_k$.

92 – W pozycji [28] zabrakło informacji, że jest to rozdział w książce *Artifitial Intelligence and Soft Computing* pod redakcją L. Rudkowskiego i R. Scherera wydanej w serii *Lecture Notes in Computer Science*, a przy podawaniu stron brakuje kreski rozdzielającej stronę początkową od końcowej.

4. Podsumowanie

Rozprawa jest poświęcona ważnej i aktualnej tematyce w przetwarzaniu tekstów, a mianowicie redukcji wymiarowości opisu dokumentów tekstowych przez wprowadzenie ukrytej reprezentacji tych dokumentów, co może być użyte do przetwarzania tych dokumentów w rozsądnych czasach obliczeniowych. Doktorant wykazał się dobrą znajomością literatury przedmiotu oraz dużą innowacyjnością, wprowadzając nowe rozwiązania do istniejących metod oraz proponując nowe metody własnego pomysłu, szczególnie zaawansowane rozdziale 3 – metoda Boost Multi-class SLD – oraz rozdziale 4 – metody DLGMM i DSLGMM. Pod tym względem rozprawa wyróżnia się wyraźnie na

tle wielu innych rozpraw doktorskich. Niestety, znacznie słabiej należy ocenić rozprawę pod kątem opisu wprowadzonych pomysłów i metod.

Wydaje mi się, że uzyskane w rozprawie wyniki mogłyby być z powodzeniem opublikowane w wysoko punktowanych czasopismach, po dopracowaniu i ewentualnym uzupełnieniu badań. Do tego celu byłaby jednak potrzebna bardziej dojrzała prezentacja materiału, także pod względem językowym. Zachęcam Doktoranta do podjęcia takich prób, bo z pewnością przyczyniłyby się one do jego rozwoju naukowego.

Rekapitulując całość recenzji stwierdzam, że przedłożona rozprawa spełnia warunki stawiane rozprawom doktorskim przez odpowiednią ustawę oraz warunki zwyczajowe i wnoszę o dopuszczenie Doktoranta do dalszych etapów przewodu doktorskiego.