Wrocław, 16 kwietnia 2018 r.

dr hab. inż. Bogdan Trawiński, Prof. PWr.
Wydział Informatyki i Zarządzania
Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
e-mail: bogdan.trawinski@pwr.edu.pl
tel. kom. 697228303

Recenzja rozprawy doktorskiej
mgr. Barbary Probierz
p.t.
„ Automatyczna kategoryzacja wiadomości elektronicznych z zastosowaniem sieci społecznych oraz algorytmów mrowiskowych”

Przedmiotem recenzji jest rozprawa doktorska mgr. Barbary Probierz o podanym wyżej tytule, która została przygotowana na Wydziale Informatyki i Nauki o Materiałach Uniwersytetu Śląskiego w Katowicach pod kierunkiem naukowym Pani dr. hab. Urszuli Boryczki, prof. UŚ, a promotorem pomocniczym jest Pan dr Jan Kozak.

Niniejsza recenzja została przygotowana na podstawie pisma Pani prof. dr hab. Danuty Stróż, Dziekan Wydziału Informatyki i Nauki o Materiałach Uniwersytetu Śląskiego w Katowicach z dnia 13 grudnia 2017 r.

PRZEDMIOT ROZPRAWY
Przedmiotem rozprawy są zagadnienia zarządzania pocztą elektroniczną, a w szczególności problem automatycznego przypisywania wiadomości mejlowych do folderów (*ang. e-mail foldering problem*). Jest to ważny i aktualny problem w dyscyplinie informatyki, zarówno pod względem naukowym, jak i praktycznym. Dla wielu pracowników korporacji, organizacji i instytucji naukowych poczta elektroniczna jest podstawowym środkiem komunikacji. Nieprzerwane strumienie napływających wiadomości mejlowych wywołują potrzebę i zainteresowanie tworzeniem systemów, które w sposób automatyczny będą wspomagały użytkowników w zarządzaniu pocztą elektroniczną, a w tym również kategoryzowaniem i przypisywaniem mejli do folderów.

Zagadnienie to rozpatruje się jako przykład automatycznej metody klasyfikacji wieloklasowej, która ma za zadanie przyporządkowanie każdemu rozpatrywanemu obiektowi (*tu wiadomości mejlowej*) jedną z wielu dostępnych klas (*tu folderu o konkretnej nazwie*). Jest to zagadnienie trudne do rozwiązania metodami uczenia maszynowego, bowiem struktura wiadomości mejlowych może być złożona, informacje zawarte w temacie mejla mogą mieć inne znaczenie, aniżeli informacje znajdujące się w treści mejla lub załącznikach. Ponadto foldery często nie zawsze odpowiadają tematom mejli, mogą odnosić się do zadań do wykonania, konkretnych osób, zespołów projektowych, mogą mieć znaczenie tylko w połączeniu z wcześniejszymi wiadomościami itp. Częstokroć klasyczne modele klasyfikatorów, jak drzewa decyzyjne, naiwny klasyfikator Bayesowski, maszyny wektorów wspierających, czy sieci neuronowe, nie są wystarczająco skuteczne.

W 2017 Mujtaba i in. [1] opracowali przegląd najnowszych prac w zakresie klasyfikowania poczty elektronicznej, które ukazały się w latach 2016-2016 w czasopismach z listy filadelfijskiej i konferencjach indeksowanych w Web of Science. Kategoryzacja wiadomości mejlowych do folderów była jednym z najczęściej rozwijanym kierunkiem badań, obok

klasyfikacji mejli w celu wykrywania spamu i phishingu.
[1] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Nahdia Majeed, Mohammed Ali Al-garadi: Email Classification Research Trends: Review and Open Issues. IEEE Access 5, pp. 9044-9064 (2017)

Dlatego też pozytywnie należy ocenić podjętą przez doktorantkę próbę opracowania algorytmu kategoryzującego wiadomości mejlowe do folderów, wykorzystującego dodatkowe dane pozyskane z analizy sieci społecznej zbudowanej na podstawie kontaktów pomiędzy użytkownikami poczty elektronicznej, a następnie na zastosowaniu algorytmu mrowiskowego do budowy klasyfikatora. Kategoryzowanie wiadomości e-mail jest bowiem zależne od upodobań indywidualnych użytkowników, które można przedstawić w postaci sieci społecznych opartych na kontaktach pomiędzy nadawcami a odbiorcami mejli. Analiza sieci społecznych pozwala lepiej zrozumieć zachowania użytkowników poczty elektronicznej i pozyskać dodatkowe dane, które mogą wpłynąć na poprawienie dokładności przypisywania wiadomości mejlowych do folderów. Dodatkowo, wykorzystanie algorytmów mrowiskowych do budowy klasyfikatorów umożliwia przeszukiwanie większej przestrzeni rozwiązań, co również prowadzi do lepszej skuteczności proponowanych rozwiązań.

UKŁAD I ZAWARTOŚĆ ROZPRAWY
Recenzowana rozprawa doktorska liczy 121 stron i składa się z dziewięciu rozdziałów oraz spisu treści, wstępu, podsumowania, bibliografii (99 pozycji), spisu rysunków (33 pozycje) i wykazu tabel (28 pozycji). Rozprawa została napisana w języku polskim. Jest poprawna pod względem językowym.

Trzy pierwsze rozdziały rozprawy mają charakter wprowadzający i zawierają przegląd ogólnych zagadnień związanych z odkrywaniem wiedzy z danych, drzewa decyzyjnymi, algorytmami mrowiskowymi i sieciami społecznymi. W rozdziale pierwszym Autorka przedstawia główne pojęcia i modele analizy danych, a także opisuje podstawowe metody eksploracji danych. Rozdział drugi jest poświęcony budowie drzew decyzyjnych oraz lasów losowych z wykorzystaniem algorytmów mrowiskowych. Jako wprowadzenie to tych zagadnień, Autorka prezentuje algorytmy do konstruowania drzew decyzyjnych, a następnie omawia zespoły klasyfikatorów: bagging, boosting i lasy losowe. W rozdziale trzecim Autorka ujmuje wprowadzenie do sieci społecznych, przedstawia modele sieci społecznych oraz podstawowe wskaźniki charakteryzujące sieci społeczne, omawia znaczenie analizy sieci społecznych (SNA).

W rozdziale czwartym Autorka charakteryzuje zbiór danych Enron E-mail, który wykorzystała do przeprowadzenia badań eksperymentalnych. Ponadto Autorka dokonuje przeglądu badań nad metodami kategoryzacji i klasyfikacji wiadomości mejlowych. Pokazuje sposób przygotowania tego zbioru do własnych badań. Prezentuje opracowaną przez siebie metodę przekształcenia wiadomości ze skrzynek pocztowych do tabel decyzyjnych, które stanowią dane wejściowe do użytych algorytmów uczenia maszynowego.

W kolejnych dwóch rozdziałach Autorka zamieszcza wyniki badań eksperymentalnych nad kolejnymi wersjami algorytmów do kategoryzowania wiadomości email. Jako miary jakości klasyfikacji Autorka użyła dokładności klasyfikacji. Wyniki wskazują, że lepszą dokładność zapewniają algorytmy proponowane przez autorkę.

Rozdział piąty zawiera wyniki początkowych badań przeprowadzonych przez Autorkę z użyciem systemu RSES, opracowanego na Uniwersytecie Warszawskim. Autorka porównała rezultaty zawartych w tym systemie trzech algorytmów opartych na teorii zbiorów przybliżonych z wynikami dostarczonymi przez algorytm budowy drzew decyzyjnych CART, wykorzystujący dane, przygotowane w postaci tabel decyzyjnych.

Autorka porównała osiągnięte przez siebie dokładności predykcji z opublikowanymi wynikami uzyskanymi przez zespół Bekkermana w jego znanym raporcie z 2004 roku. Przeprowadziła nieparametryczny test statystycznej istotności Friedmana.

W rozdziale szóstym przedstawione są wyniki eksperymentów przeprowadzonych na zbiorze Enron E-mail z użyciem algorytmu mrowiskowego aACDT, zmodyfikowanego o analizę sieci komunikacji, polegającej na badaniu listy odbiorców. Analogicznie do wcześniejszych badań porównano uzyskane wyniki z opublikowanymi wynikami zespołu Bekkermana i przeprowadzono analizę statystycznej istotności z wykorzystaniem testu Friedmana. Ponadto dla 17 zbiorów (skrzynek pocztowych), o dużej liczbie wiadomości mejlowych i folderów, wykonano badania porównawcze algorytmu mrowiskowego aACDT z sześcioma algorytmami klasyfikacji wybranymi z systemu WEKA. W tym wypadku wykazano również przewagę proponowanego algorytmu. Kolejna seria eksperymentów została wykonana z wykorzystaniem algorytmu mrowiskowego ACDF do konstruowania lasów decyzyjnych, autorstwa obojga promotorów pracy, oraz czterech metod tworzenia zespołów klasyfikatorów, wybranych z pakietu WEKA, a mianowicie AdaBoost, Dagging, Bagging i Random Forest. W tym wypadku proponowany w rozprawie algorytm także zapewnił lepsze wyniki klasyfikacji.

W rozdziale siódmym Autorka buduje sieć społeczną podstawie kontaktów mejlowych pomiędzy poszczególnymi pracownikami firmy Enron. Następnie dokonuje analizy struktur powiązań społecznych w tej sieci w skali makro, meso i mikro.

Rozdział ósmy zawiera główne wyniki badań zrealizowanych w ramach rozprawy doktorskiej. Autorka prezentuje w nim konstrukcję autorskiego algorytmu do automatycznego kategoryzowania wiadomości email, polegającego na tworzeniu sieci społecznej opartej na kontaktach pomiędzy użytkownikami, wyodrębnieniu grup użytkowników, a następnie na zastosowaniu algorytmu mrowiskowego do budowy klasyfikatora. Następnie Autorka przedstawia wyniki eksperymentów ewaluacyjnych na danych pozyskanych ze zbioru Enron Email i przekształconych do postaci tabel decyzyjnych utworzonych dla poszczególnych użytkowników. Wykazała, że jej autorski algorytm osiąga statystycznie istotnie lepszą dokładność przypisywania wiadomości mejlowych do folderów w porównaniu z algorytmami aADCT i ACDF, których rezultaty prezentowała w rozdziale szóstym pracy.

W rozdziale dziewiątym Autorka proponuje metodę, która sugeruje użytkownikom zakładanie nowych folderów, na podstawie struktury folderów innych użytkowników, wyznaczonych przez stworzoną sieć społeczną. Metoda predykcji nowych folderów bazuje na analizie macierzy śladu feromonowego dla wszystkich użytkowników w grupie. Macierz ta jest tworzona w czasie klasyfikowania wiadomości do folderów. Autorka zilustrowała działanie zaproponowanej metody na przykładowych danych ze zbioru Enron Email.

ORYGINALNE OSIĄGNIĘCIA

Badania w ramach pracy doktorskiej zostały wykonane na Uniwersytecie Śląskim w Katowicach, w zespole, który od kilku lat rozwija metody klasyfikowania i kategoryzacji wiadomości mejlowych z wykorzystaniem algorytmów mrowiskowych i przeprowadza eksperymenty z ich zastosowaniem. Do głównych osiągnięć Doktorantki należy:

- opracowanie autorskiego algorytmu do automatycznego przypisywania wiadomości mejlowych do folderów, polegającego na wykorzystaniu danych z grup użytkowników, wyodrębnionych w sieci społecznej, opartej na kontaktach pomiędzy użytkownikami, a następnie na zastosowaniu algorytmu mrowiskowego do budowy klasyfikatora,
- przygotowanie do badań zbioru danych Enron E-mail poprzez oczyszczenie go, dostosowanie do problemu oraz przekształcenie do struktury tabel decyzyjnych. Stworzone

przez Autorkę tabele decyzyjne stanowiły dane wejściowe do algorytmów klasyfikacyjnych,
- zbudowanie sieci społecznej na podstawie kontaktów pomiędzy nadawcami a odbiorcami wiadomości mejlowych w zbiorze danych Enron E-mail, przeprowadzenie analizy utworzonej sieci społecznej i wyodrębnienie w niej grup użytkowników posiadających podobną strukturę społeczną,
- zaplanowanie i przeprowadzenie eksperymentów ewaluacyjnych wykazujących, że zaproponowany algorytm charakteryzuje się lepszą dokładnością klasyfikacji w porównaniu z algorytmami opartymi wprawdzie na algorytmach mrowiskowych, lecz nie wykorzystującymi danych pozyskanych z analizy sieci społecznej użytkowników poczty elektronicznej,
- zaplanowanie i przeprowadzenie eksperymentów ewaluacyjnych wykazujących, że zaproponowany algorytm charakteryzuje się lepszą dokładnością klasyfikacji w porównaniu z klasycznymi klasyfikatorami oraz z zespołami klasyfikatorów,
- porównanie wyników przeprowadzonych własnych eksperymentów z rezultatami uzyskanymi przez innych badaczy i opublikowanymi w literaturze naukowej,
- przeprowadzenie analizy statystycznej istotności uzyskanych wyników z wykorzystaniem testów nieparametrycznych,
- zastosowanie opracowanego autorskiego algorytmu do sugerowania tworzenia nowych folderów, na podstawie struktury folderów innych użytkowników z tej samej grupy, wyodrębnionej w sieci społecznej.

UWAGI KRYTYCZNE I DYSKUSYJNE
Poniższe uwagi mają charakter dyskusyjny i nie wpływają na pozytywną ocenę recenzowanej pracy doktorskiej:

Teza pracy (problemy terminologiczne)
Autorka sformułowała następującą tezę swojej rozprawy doktorskiej (recenzent zaznaczył terminy, do których formułuje swoje uwagi):
*„Zastosowanie algorytmów mrowiskowych i sieci społecznych w problemie automatycznego **kategoryzowania** wiadomości e-mail pozwala na poprawę **trafności** przypisywania wiadomości do folderów oraz umożliwia sugerowanie zakładania nowych folderów dla użytkowników."*

Autorka użyła w tezie pracy wyrażenia „*kategoryzowanie wiadomości e-mail*". Jednakże w tekście pracy używa wielokrotnie, oprócz tego wyrażenia, również zamiennie terminu „*klasyfikacja wiadomości e-mail*". Zresztą podobnie jest w literaturze światowej na ten temat: w tytułach prac pojawiają się albo terminy „*email classification*", albo „*email categorization*". Tym niemniej, Autorka powinna przedstawić w rozprawie definicje obu terminów i przedyskutować ich ewentualne różnice znaczeniowe.
Podobnie jest z wyrażeniem „*trafność przypisywania wiadomości do folderów*". Autorka używa go zamiennie z terminem „dokładność klasyfikacji wiadomości e-mail do folderów". Autorka stosuje wielokrotnie przy prezentowaniu wyników eksperymentów miarę „dokładność klasyfikacji". Jednakże, nie definiuje jej w sposób formalny w tekście pracy, a to jest konieczne, bowiem dla klasyfikacji wieloklasowej stosowanych jest kilka różnych miar jakości. Przegląd tych miar można znaleźć w pracy [2].
[2] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management 45, pp. 427-437 (2009).

Przegląd prac związanych
Przedmiotem rozprawy są metody kategoryzacji/klasyfikacji wiadomości pocztowych z wykorzystaniem podejść mrowiskowych oraz analizy sieci społecznych. Badania nad tymi metodami były intensywnie prowadzone w ciągu minionych dwóch dekad i nadal są rozwijane. Dlatego też, zamieszczenie w rozprawie analizy najnowszych prac w tym zakresie,

mogłoby wzbogacić przeglądy zamieszczone w rozdziałach 1,2,3 i 5. Przykładami takich prac są:

[3] Min-Feng Wang, Sie-Long Jheng, Meng-Feng Tsai, Cheng-Hsien Tang: Enterprise Email Classification Based on Social Network Features. IEEE International Conference onAdvances in Social Networks Analysis and Mining, Kaohsiung, Taiwan (2011), DOI: 10.1109/ASONAM.2011.89

[4] Anton Borg, Niklas Lavesson: E-mail Classification using Social Network Information. IEEE Seventh International Conference on Availability, Reliability and Security, Prague, Czech Republic (2012), DOI: 10.1109/ARES.2012.84

[5] Mihajlo Grbovic, Guy Halawi, Zohar Karnin, Yoelle Maarek: How Many Folders Do You Really Need? Classifying Email into a Handful of Categories. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014), Shanghai, China. DOI: 10.1145/2661829.2662018

[6] Izzat Alsmadi, Ikdam Alhami: Clustering and classification of email contents. Journal of King Saud University - Computer and Information Sciences 27:1, pp.46-57 (2015)

[7] Shaukat Wasi, Syed Imran Jami, Zubair Ahmed Shaikh: Context-based email classification model. Expert Systems 33:2, pp. 129-144 (2016)

Dyplomantka nie cytuje w rozprawie najbardziej wartościowych pozycji swojej pani Promotor i Kopromotora:

[8] Boryczka U., Kozak J., Enhancing the effectiveness of ant colony decision tree algorithms by co-learning, Applied Soft Computing 30, pp. 166–178 (2015)

[9] Kozak J., Boryczka U.: Collective data mining in the ant colony decision tree approach. Information Sciences 372, pp. 126–147 (2016)

Dyplomantka nie cytuje czterech swoich prac napisanych wspólnie z panią Promotor i Kopromotorem, w których publikuje cześć wyników prezentowanych w niniejszej rozprawie. Prace te zostały opublikowane materiałach międzynarodowych konferencji ICCCI 2014, ACIIDS 2015, ICCCI 2015 oraz KES-IDT 2016, wydanych przez Springera, a następnie zaindeksowane w Web of Science. Są to następujące prace:

[10] Boryczka U., Probierz B., Kozak J. (2014) An Ant Colony Optimization Algorithm for an Automatic Categorization of Emails. In: Hwang D., Jung J.J., Nguyen NT. (eds) Computational Collective Intelligence. Technologies and Applications. ICCCI 2014. Lecture Notes in Computer Science, vol 8733. Springer, Cham

[11] Boryczka U., Probierz B., Kozak J. (2015) Adaptive Ant Colony Decision Forest in Automatic Categorization of Emails. In: Nguyen N., Trawiński B., Kosala R. (eds) Intelligent Information and Database Systems. ACIIDS 2015. Lecture Notes in Computer Science, vol 9011. Springer, Cham

[12] Boryczka U., Probierz B., Kozak J. (2015) A New Algorithm to Categorize E-mail Messages to Folders with Social Networks Analysis. In: Núñez M., Nguyen N., Camacho D., Trawiński B. (eds) Computational Collective Intelligence. Lecture Notes in Computer Science, vol 9330. Springer, Cham

[13] Boryczka U., Probierz B., Kozak J. (2016) Automatic Categorization of Email into Folders by Ant Colony Decision Tree and Social Networks. In: Czarnowski I., Caballero A., Howlett R., Jain L. (eds) Intelligent Decision Technologies 2016. Smart Innovation, Systems and Technologies, vol 57. Springer, Cham

W rozdziale 3 Autorka opisuje lasy losowe, jako przykład zespołu modeli klasyfikacyjnych. Przedstawienie samego wprowadzenia do baggingu, jako metody bazowej, jest niewystarczające, bowiem lasy losowe są złożeniem dwóch grup metod tworzenia zespołów modeli, a mianowicie baggingu oraz metody losowych podprzestrzeni (*ang. random subspace*), która polega na tworzeniu modeli składowych z wykorzystaniem podzbiorów losowo wybranych atrybutów. Metodę losowych podprzestrzeni zaproponował Ho [14].

[14] Ho T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), pp. 832-844 (1998)

Plan i wyniki eksperymentów

Przedstawiony w Rozdziałach 5, 6 i 8 opis przeprowadzonych eksperymentów pozostawia pewien niedosyt. Jest on bowiem lakoniczny i nie pozwala czytelnikowi na powtórzenie badań, ani też na stwierdzenie, czy wnioski wyciągane przez Doktorantkę oparte były na porównywalnych wynikach badań. Do najważniejszych niedociągnięć należą:

- brak formalnej definicji dokładności klasyfikacji, czy inni autorzy, z wynikami których porównywane były rezultaty własnych eksperymentów, stosowali taką samą miarę dokładności,
- brak opisu metody podziału zbiorów danych na podzbiory danych treningowych i testowych, w jakiej proporcji były one dzielone, czy też może była stosowana dziesięciokrotna walidacja krzyżowa, która jest domyślną metodą podziału w pakiecie WEKA,
- brak opisu, czy parametry klasycznych modeli i zespołów modeli w pakiecie WEKA były wstępnie strojone.

Przykładowo, w rozdziale 6 Autorka stwierdza, że klasyczne zespoły klasyfikatorów nie prowadzą dają dobrych wyników i często klasyfikują wiadomości ze znacznie mniejszą dokładnością nawet o 30-40%. Ponadto zastosowanie zespołu klasyfikatorów daje gorsze rezultaty niż pojedynczy klasyfikator z analizą odbiorców. Jednakże, Autorka nie opisuje, w jaki sposób wykorzystała zawarte w pakiecie WEKA zespoły klasyfikatorów. Maja one bowiem kilka istotnych parametrów, które wymagają wykonania wstępnych badań, celem ich dostrojenia. Wówczas można uzyskać znacznie lepsze wyniki, aniżeli stosując domyślne wartości parametrów ustawione w pakiecie WEKA.

Testy statystyczne

Doktorantka zastosowała testy nieparametryczne do wykazania statystycznie istotnych różnic pomiędzy dokładnością klasyfikacji badanych algorytmów. Jest to prawidłowe podejście, ze względu na bardzo małą liczbę obserwacji: od 7 do 17. W podsumowaniu Doktorantka stwierdza, że dla celów porównania użyła nieparametrycznego testu Manna-Whitneya-Wilcoxona (testu sumy rang Wilcoxona dla dwu próbek). Jednakże w rozdziałach 5, 6 i 8 zamieszcza wyniki tylko nieparametrycznego testu Friedmana i nie wspomina nic o teście Manna-Whitneya-Wilcoxona, co sprawia, że czytelnik mógłby mieć wątpliwości, czy ten test w ogóle był stosowany. Ponadto, test Manna-Whitneya-Wilcoxona przeznaczony jest dla przypadku próbek różnolicznych. W wypadku posiadania obserwacji równolicznych, dających się połączyć w pary, a to ma miejsce w prezentowanych w rozprawie eksperymentach, właściwsze byłoby użycie testu Wilcoxona dla par obserwacji, który jest nieparametryczną alternatywą dla testu t-Studenta. Recenzent chciałby zwrócić również uwagę, że w wypadku analiz Doktorantki mamy do czynienia z porównaniami wielokrotnymi, a do tego testy dla par obserwacji są niewystarczające. Niezbędne jest zastosowanie procedur post-hoc, które pozwalają skompensować tzw. *family-wise error*, np. opisanych w kilku następujących artykułach przez zespół prof. Herrery z Uniwersytetu w Grenadzie w Hiszpanii:

[15] García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for All Pairwise Comparisons. Journal of Machine Learning Research 9, pp. 2677–2694 (2008)

[16] García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced Nonparametric Tests for Multiple Comparisons in the Design of Experiments in Computational Intelligence and Data Mining: Experimental Analysis of Power. Information Sciences 180, pp. 2044–2064 (2010)

[17] Derrac, J., García, S., Molina, D., Herrera, F.: A Practical Tutorial on the Use of Nonparametric Statistical Tests as a Methodology for Comparing Evolutionary and Swarm Intelligence Algorithms. Swarm and Evolutionary Computation 1, pp. 3–18 (2011)

Złożoność obliczeniowa
Dokładność i wydajność są podstawowymi miarami służącymi do oceny proponowanych metod klasyfikacji. W rozprawie Doktorantka nie prowadzi systematycznego badania złożoności obliczeniowej proponowanych algorytmów, ani poprzez analizę formalną, ani też poprzez pomiar czasów przetwarzania.

Generalizacja zaproponowanych rozwiązań
Doktorantka opracowała zaproponowane przez siebie rozwiązania na podstawie przetwarzania i analizy zbioru danych Enron E-mail. Wszystkie badania ewaluacyjne przeprowadziła również z wykorzystaniem tego zbioru danych. Powstaje zatem pytanie, jaka jest ogólność prezentowanych w rozprawie metod. Czy nadają się one tylko i wyłącznie do przetwarzania zbioru Enron E-mail, czy również mogą być stosowane do innych zbiorów wiadomości mejlowych?

Dynamiczna natura komunikacji pocztowej
Komunikacja za pośrednictwem poczty elektronicznej ma charakter dynamiczny. Zaproponowany w rozprawie algorytm oraz przeprowadzone eksperymenty są adekwatne dla zbiorów danych stacjonarnych, co nie odpowiada dynamicznemu charakterowi komunikacji za pośrednictwem poczty elektronicznej. Recenzent sugeruje Doktorantce podjęcie próby opracowania algorytmów uwzględniających zmienność strumienia nadchodzącej poczty elektronicznej w czasie. Np. eksperymenty można byłoby przeprowadzić, ucząc modele klasyfikacyjne na pewnym stacjonarnym zbiorze danych z początkowego okresu czasu, a następnie uporządkować pozostałą część danych wg czasu i dostosowywać modele do zmieniającego się strumienia danych. Przykłady najnowszych badań nad klasyfikacją/klasteryzacją strumieni danych z zastosowaniem metod mrowiskowych można znaleźć w następujących artykułach:
[18] Nesrine Masmoudi, Hanane Azzag, Mustapha Lebbah, Cyrille Bertelle, Maher Ben Jemaa: CL-AntInc Algorithm for Clustering Binary Data Streams Using the Ants Behavior. Procedia Computer Science 96, pp. 187-196 (2016)
[19] Conor Fahy and Shengxiang Yang: Dynamic Stream Clustering Using Ants. 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK, pp. 495-508 (2016)
[20] Conor Fahy, Shengxiang Yang, Mario Gongora: Finding Multi-Density Clusters in non-stationary data streams using an Ant Colony with adaptive parameters. 2017 IEEE Congress on Evolutionary Computation (CEC), pp. 673-680 (2017)

Drobne uwagi o charakterze formalnym
Wzór (1.15) na odchylenie standardowe na stronie 5 jest błędny. Wyrażenie po prawej stronie znaku równości powinno być spierwiastkowane.

Podpisy pod rysunkami 6.1 na stronie 69 oraz 6.2 na stronie 72 są niezrozumiałe. Mają one następujące brzmienie:
Rysunek 6.1: Poprawność dokładności klasyfikacji proponowanej metody w stosunku do artykułu [10]
Rysunek 6.2: Poprawność dokładności klasyfikacji proponowanej metody

Na stronie 71 Autorka stwierdza: „Dla wszystkich zbiorów danych stworzonych na podstawie dziesięciu użytkowników proponowany algorytm za każdym razem uzyskuje lepsze rezultaty.”, podczas gdy badania wykonywała na 17 zbiorach.

Pozycje literaturowe [86] i [87] są te same. W spisie literatury różnią się tylko rokiem wydania, przy czym pozycja [87] zawiera błędny rok.

W spisie literatury opisy bibliograficzne przedstawione są w sposób niejednolity. Z kolei, opisy bibliograficzne [23], [26], [43], [79] [79], [81], [86] i [87], [89], [90] są niekompletne.

KONKLUZJA

Porównanie wymienionych powyżej oryginalnych wyników Autorki o charakterze naukowym z pewnymi niedociągnięciami, czy problemami dyskusyjnymi wypada na korzyść rozprawy, która wnosi wartościowy wkład do dziedziny zagadnień związanych z kategoryzacją wiadomości mejlowych. Doktorantka ma już dorobek naukowy liczący osiem pozycji, a w tym cztery referaty opublikowane na międzynarodowych konferencjach ICCCI 2014, ACIIDS 2015, ICCCI 2015 oraz KES-IDT 2016 a następnie zaindeksowane w Web of Science. Z powyżej wymienionych powodów moja ocena przedłożonej do recenzji rozprawy jest pozytywna. Uważam też, że należy również wysoko ocenić nakład pracy Autorki oraz opiekę pani Promotor oraz Kopromotora.

Uważam, że w recenzowanej rozprawie doktorskiej mgr. Barbary Probierz rozwiązany został oryginalny problem badawczy. Doktorantka wykazała się dobrą znajomością literatury przedmiotu rozprawy, umiejętnością konstruowania algorytmów, a także zaplanowania i przeprowadzenia eksperymentów obliczeniowych. W konkluzji stwierdzam, że rozprawa doktorska ***Automatyczna kategoryzacja wiadomości elektronicznych z zastosowaniem sieci społecznych oraz algorytmów mrowiskowych*** spełnia wymogi ustawy o stopniach i tytule naukowym (Dz. U. Nr 65 z dnia 14 marca 2003r., ze zmianami w Dz. U. z 2005r. Nr 164 oraz w Dz. U. z 2011r. Nr 84). Wnoszę o jej przyjęcie i dopuszczenie do publicznej obrony.

dr hab. inż. Bogdan Trawiński, Prof. PWr.