Prof. dr hab. inż. Krzysztof Krawiec
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2
60-965 Poznań

Poznań, 20.08.2019

Recenzja rozprawy doktorskiej mgr Macieja Jankowskiego pt. "Dimension Reduction Methods in Text Classification with Probabilistic Graphical Models"

1 Tematyka rozprawy

Tematem rozprawy mgr Macieja Jankowskiego są metody redukcji wymiarowości w zadaniach klasyfikacji tekstów. Analiza tekstu to jedna z najbardziej interesujących i intensywnie zgłębianych obecnie gałęzi sztucznej inteligencji, uczenia maszynowego i analizy danych. Jest to obszar trudny, z racji wymagającej reprezentacji danych (ciągi słów o zmiennej długości), kontekstowego charakteru danych, oraz trudności w ich modelowaniu.

Mgr Jankowski w swojej rozprawie zdecydował się badać i rozwijać metody reprezentowane jako modele graficzne. Modele graficzne sprawdzały się wielokrotnie w wielu gałęziach sztucznej inteligencji, w tym także w analizie tekstów. W ramach studiowanych modeli sięgnął też po relatywnie nowe głębokie modele wariacyjne.

W związku z powyższym tematykę pracy mgr Jankowskiego mogę z pełnym przekonaniem uznać za aktualną, interesującą, nietrywialną i zdecydowanie umiejscowioną w dyscyplinie informatyka, a dokładniej w nośnym obecnie nurcie uczenia maszynowego, analizy danych i odkrywania wiedzy.

2 Ocena treści rozprawy i wkładu oryginalnego

2.1 Ocena treści rozprawy

Rozprawa jest bardzo zwięzła: składa się z 5 rozdziałów, bibliografii, i ma w sumie 93 strony.

Rozdział 1 wprowadza pojęcia podstawowe, w tym m.in. ukrytą analizę semantyczną/indeksowanie semantyczne (ang. latent semantic analysis, LSA, latent semantic

indexing), probabilistyczne modele zagadnień (ang. probabilistic topic models), alokację ukrytych zmiennych Dirichleta (ang. Latent Dirichlet Allocation, LDA), rozkłady prawdopodobieństwa często stosowane w analizie tekstu, oraz algorytm maksymalizacji wartości oczekiwanej (ang. expectation maximization).

Rozdziały 2-4 prezentują oryginalne badania i przyczynki Doktoranta. Rozdział 2 prezentuje wyniki badań nad modelami złożonymi (de facto *klasyfikatorami* złożonymi) w których modele bazowe mają postać LDA. Autor pokazuje empirycznie że modele złożone są w stanie osiągać wyższą trafność klasyfikowania niż modele proste (pojedyncze). Choć samo w sobie nie jest to może zbyt zaskakujące, to za oryginalne uważam prace nad metodami ważenia wyjść klasyfikatorów bazowych, a w szczególności tą wykorzystującą tzw. perplexity klasyfikatorów bazowych, która zresztą okazała się prowadzić do najlepszych wyników. Część eksperymentalna pracy dwa zbiory danych (benchmarki): Mobile Apps oraz Poliblog.

W rozdziale tym Doktorant zajął się także próbą automatycznego ustalania optymalnej liczby tematów (tematów w sensie topic modeling), w tym zaproponował metodę opartą na pomiarze entropii rozkładu tematów (sekcja 2.2.1) i skonfrontował ją empirycznie z metodą pełnego przeglądu opartą na technikach LSA i LDA, wykazując skuteczność tej pierwszej we wskazywaniu optimów.

W rozdziale 3 mgr Jankowski połączył nadzorowaną wersję metody LDA (sLDA) ze znaną od dłuższego czasu metodą boostingu, polegającą na konstrukcji wielu zdywersyfikowanych klasyfikatorów bazowych poprzez ich uczenie na danych o lekko zaburzonym rozkładzie. W szczególności, badania dotyczyły problemu wieloklasowego, stąd słuszny wybór jako punktu wyjścia wieloklasowej odmiany metody sLDA. Autor wykorzystał w tym celu popularny algorytm AdaBoost, który iteracyjnie modyfikuje wagi przykładów na podstawie trudności ich poprawnego zaklasyfikowania, i w ten sposób kieruje uwagę algorytmu uczącego na różne obszary w przestrzeni reprezentacji danych. Uruchomiony wielokrotnie, AdaBoost generuje w ten sposób zdywersyfikowany zestaw klasyfikatorów bazowych – którymi w tym przypadku były instancje wieloklasowego sLDA. W części empirycznej tego rozdziału Autor zademonstrował działanie proponowanej metody na problemach SmsSpam i Poliblog. Eksperyment obejmował także dobór liczby tematów K. Eksperyment potwierdził przewagę proponowanego klasyfikatora złożonego nad stosowaniem pojedynczych klasyfikatorów.

Rodział 4 opisuje propozycje Doktoranta dotyczące wykorzystania głębokich modeli generatywnych (and. deep generative models) do klasyfikacji dokumentów tekstowych. W tym celu sięgnął on po metody wariacyjne, a w szczególności autoenkodery (autokodery) wariacyjne (ang. Variational Autoencoders, VAE), stanowiące jedną z najbardziej interesujących osiągnięć uczenia głębokiego. Autor w pierwszej kolejności zilustrował zastosowanie VAE do klasyfikacji tekstu w trybie nienadzorowanym poprzez graficzną prezentację przestrzeni ukrytej (ang. latent space) (jako że pierwotnie właśnie do tej klasy metod należy VAE; sekcja 4.4). Następnie w sekcji 4.5 przedstawił (znany z literatury) sposób modyfikacji VAE do wariantu uczenia nadzorowanego (sVAE), w której od dekodera oczekuje się nie tylko rekonstrukcji przykładu wejściowego (reprezentowanego tu jako lekko zmodyfikowany wektor TF-IDF), ale także klasy do której należy dokument.

Doktorant pokazał też jak odpowiednio zmodyfikować funkcję straty i zaproponował ważenie jej komponentów celem (m.in.) usprawnienia procesu uczenia. Choć wizualizacja danych w przestrzeni zmiennych ukrytych wykazała lepszą separację klas w porównaniu z VAE, Doktorant zaproponował inny kierunek prac, polegający na zastąpieniu tradycyjnego unimodalnego rozkładu normalnego stosowanego w VAE jako rozkładu *a priori* wielomodalną mieszaniną rozkładów normalnych (ang. Gaussian Mixture Model, GMM), gdzie komponenty mieszaniną odpowiadają poszczególnym klasom decyzyjnym. Propozycja obejmuje m.in. sposób umieszczania mod (centroidów) komponentów mieszaniny w przestrzeni 2D i odpowiednie modyfikacje funkcji straty. W końcu, w sekcji 4.7, mgr Jankowski dokonał fuzji autoenkodera GMM z uczeniem nadzorowanym, odpowiednio aktualizując funkcję straty (metoda DSLGMM). Finalne wizualizacje przestrzeni latent istotnie charakteryzują się najlepszą separacją klas spośród wszystkich wariantów rozważanych w tym rozdziale.

Eksperymenty w tej części pracy przeprowadzone były na zbiorach Poliblog, SmsSpam, trzech podzbiorach znanej kolekcji dokumentów Reuters, oraz zbiorze obrazów MNIST, popularnym benchmarku w analizie obrazów (gdzie jak rozumiem w miejsce reprezentacji przykładów jako wektory TF-IDF wprowadzano na wejścia sieci bezpośrednio obrazy). W końcowej części tego rozdziału autor skonfrontował, w trybie klasyfikacji, proponowane techniki nadzorowane i wykazał przewagę podejścia DSLGMM. Wskazał także na zalety tego podejścia: możliwość generowania sztucznych przykładów oraz szybki czas uczenia, wielokrotnie krótszy niż dla metod studiowanych w rozdziałach 2 i 3.

2.2 Ocena kompletności i redakcji pracy

Redakcja pracy jest zasadniczo poprawna. Praca napisana jest klarownym i przystępnym językiem angielskim. Autor umiejętnie operuje terminologią i niemal zawsze prowadzi wywody w logiczny i przekonujący sposób (jest to szczególnie widoczne w rozdziale 5, gdzie kolejne rozszerzenia wyjściowej metody VAE logicznie wynikają jedne z drugich).

Niemniej Autor nie uniknął też jednak pewnych braków i uchybień:

- Autor nie formułuje jawnie celu i zakresu pracy, ani też hipotez, co jest nadzwyczaj rzadkie w rozprawach doktorskich i znacznie utrudnia lekturę – czytelnik jest zmuszony do samodzielnego 'wywiedzenia' ich z tekstu.

- Podobnie brakuje w pracy jawnie sformułowanej listy deklarowanych przyczynków oryginalnych (ang. claimed contributions). Znajdujemy ją (ale nadal w raczej zawoalowanej postaci) dopiero w podsumowującym rozdziale 5.

- Niektóre z wyników eksperymentalnych nie są zbyt przekonujące, ponieważ uzyskano je na tylko jednym zbiorze danych (np. sekcja 2.2, zbiór Film Reviews).

- Założenie że liczba klas jest równa liczbie tematów ($K = C$, sekcja 4.6.3) jest bardzo silne, dość dyskusyjne, i jako takie zasługuje na obszerniejszy komentarz. Spodziewam się że znaczna część analiz prowadzonych w dalszej części tego rozdziału nie dałaby tak przekonujących wyników gdyby założyć że K jest nieznane.

- Autor nie wspomina o walidacji krzyżowej/skrośnej czy innych metodykach oceny klasyfikatorów, z czego domyślam się że eksperymenty przeprowadzono dla jednokrotnego podziału danych na zbiór uczący i testujący, co nieco osłabia wiarygodność wyników. Niedociągnięcie to jest szczególnie zaskakujące w kontekście metod neuronowych (rozdział 4), dla których autor sam podkreśla ich znakomitą wydajność obliczeniową.

- Wybór prezentowanych pojęć podstawowych jest dyskusyjny. Autor prezentuje np. tak elementarne pojęcia jak rozkład Bernoulliego, ale nie definiuje bardziej (moim zdaniem) wyrafinowanego pojęcia dywergencji Kullbacka-Leiblera.

- Mimo zwięzłości pracy, jest w niej zauważalna liczba powtórzeń, np. niektóre skrótowce wprowadzane są wielokrotnie (np. LDA na s. 30).

- Czasami stosowanie pojęć wyprzedza ich definicje; np. zbiory danych omówione w sekcji 4.10.1 wykorzystywane były już we wcześniejszych sekcjach.

- Zdarzają się niekonsekwencje notacyjne i rzeczowe; np. dokumenty/przykłady, oznaczane zazwyczaj przez $w^{(d)}$, w części rozdziału 4 (rys. 4.7 i sekcja 4.5.1) oznaczane są przez x. Skrót SGVB w tytule sekcji 4.4.2 wydaje się nie być wprowadzony wcześniej. Niektóre etykiety serii danych w rys. 2.3 nie wydają się mieć odpowiedników w tekście. Wykres 2.4 prezentuje błąd klasyfikowania, a jest podpisany 'classification accuracy'; człon normalizujący B w formule na dole s. 15 nie jest zdefiniowany (ani wspomniany w tekście); etykiety osi w niektórych wykresach (np. fig. 2.1) to enigmatyczne 'Value' w miejsce właściwej nazwy wielkości;

- Autor nadużywa przecinków; w szczególności, choć w języku polskim „że" poprzedzamy przecinkiem, to analogiczna reguła nie obowiązuje w języku angielskim.

2.3 Ocena wkładu oryginalnego

Mimo powyższych uchybień redakcyjnych, praca jest merytorycznie spójna, przekonująca, i zaawansowana w sensie metodycznym. Za jej główne przyczynki uznaję:

- Propozycję ustalania liczby tematów K opartą na pomiarze entropii rozkładu (sekcja 2.2.1),

- Wykorzystanie miary perplexity do ważenia klasyfikatorów składowych w klasyfikatorze złożonym opartym na metodzie LDA (sekcja 2.3.3),

- Metodę boostingu dla wieloklasowej odmiany metody sLDA (sekcja 3.4),

- Rozszerzenia autoenkoderów wariacyjnych na przypadek mieszaniny rozkładów normalnych (metoda DLGMM, sekcja 4.6),

- Dostosowanie metody DLGMM do uczenia nadzorowanego, tj. metodę DSLGMM (sekcja 4.7).

3 Konkluzja końcowa

Rozprawa doktorska mgr Jankowskiego zawiera w mojej ocenie oryginalne i wartościowe osiągnięcia, zilustrowane wynikami empirycznymi uzyskanymi na wymagających danych rzeczywistych. Wymienione powyżej uwagi polemiczne odnośnie treści i prezentacji pracy nie podważają głównych konkluzji rozprawy i mojej pozytywnej jej oceny. Szczególnie przekonują mnie: (i) stopień zaawansowania koncepcyjnego i metodologicznego pracy, (ii) dogłębna znajomość tematyki w obszarze analizy i klasyfikacji tekstów, (iii) oryginalność niektórych przyczynków, oraz (iv) aktualność obranego warsztatu i wybranych podejść bazowych.

Wobec powyższego stwierdzam, że **rozprawa doktorska mgr Macieja Jankowskiego spełnia z nawiązką warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony, o co wnoszę do Rady Wydziału Cybernetyki Wojskowej Akademii Technicznej w Warszawie.**