Recenzja rozprawy doktorskiej mgr. Pawła Matykiewicza

dr hab. inż. Szymon Jaroszewicz
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
ul. Jana Kazimierza 5
Warszawa

25 maja 2015

Rozprawa doktorska mgr. Matykiewicza dotyczy problemu klasyfikacji tekstów medycznych z wykorzystaniem wiedzy dziedzinowej reprezentowanej przez sieci semantyczne. W pracy autor konsekwentnie wspomaga się analogiami kognitywistycznymi z procesami związanymi z kojarzeniem pojęć w mózgu człowieka. Podejściem, które stosuje autor, jest wzbogacanie macierzy słów-dokumentów poprzez dodawanie do niej nowych słów związanych relacjami semantycznymi ze słowami już w niej występującymi. Analogią kognitywistyczną jest tu propagacja aktywacji neuronów. Autor rozpoczyna od stwierdzenia sprzecznego z intuicją faktu, że wzbogacanie macierzy słów-dokumentów przy pomocy sieci semantycznych prowadzi często do pogorszenia jakości klasyfikacji tekstu. Jako istotną przyczynę tego faktu autor wskazuje problem wielokrotnego dziedziczenia, który powoduje, że część informacji dodawanej na podstawie sieci semantycznej nie pomaga w odróżnianiu klasy dokumentu. Aby poprawić skuteczność metody, autor ogranicza słowa dodawane do dokumentów, posiłkując się tzw. niesemantycznym hamowaniem. Proponowane są dwa rodzaje hamowania oparte o dwie różne klasyfikacje powiązań wprowadzanych słów. Metody te stanowią najważniejszy rezultat przedstawiony w pracy. Dodatkowo autor wprowadza szereg metod wizualizacji procesu wzbogacania danych przy użyciu sieci semantycznych oraz miar oceny jakości uzyskanych wyników i przedstawia hipotezę dotyczącą związku jednej z tych miar z miarą F_1.

Recenzję rozpocznę od ogólnej oceny pracy. Wywód autora jest logiczny, rozpoczyna się od identyfikacji przyczyn, dla których wzbogacanie seman-

tyczne danych nie daje zadowalających wyników. Następnie autor przedstawia opracowane przez siebie rozwiązania owego problemu, wprowadza metody oceny ich skuteczności, a w końcu przeprowadza analizę pokazującą ich użyteczność. Praca dowodzi, że autor jest w stanie samodzielnie rozwiązywać problemy badawcze. Dodatkowo zaznaczyć należy, że temat jest istotny, gdyż odpowiednie wzbogacenie danych uczących często prowadzi do znaczącej poprawy wyników, a prac na ten temat jest stosunkowo niewiele.

Metody proponowane w pracy stanowią istotny wkład do zagadnienia powiązania sieci semantycznych z problemem klasyfikacji tekstów, ale ocena tego, na ile ów wkład jest istotny dla samego problemu klasyfikacji tekstów, pozostaje utrudniona (patrz niżej).

Dorobek publikacyjny doktoranta jest znaczący, już same prace związane z tematem rozprawy obejmują dziewięć pozycji, w tym na uznanych konferencjach takich jak IJCNN. Autor ma na swoim koncie także kilka innych publikacji, w tym w wysoko punktowanych czasopismach, takich jak Journal of the American Medical Informatics Association. Z pewnością dotychczasowy dorobek doktoranta należy ocenić wysoko.

Po przedstawieniu ogólnej, pozytywnej, oceny przejdę do uwag krytycznych dotyczących pracy, których, niestety, nie brakuje.

Pierwszym z problemów jest nieumiejętność poprawnego stosowania przez autora formalizmu matematycznego. Definicje znajdujące się w pracy nie definiują w istocie nowych pojęć, ale służą wprowadzaniu notacji (definicje 4.1, 4.2 i dalej), podają schematyczne opisy algorytmów (definicja 2.1) czy ogólne opisy (definicje 2.2 i 2.3). Językowi stosowanemu w definicjach daleko do matematycznej ścisłości, jest to raczej nie do końca precyzyjny opis.

Oznaczenia stosowane przez autora utrudniają lekturę pracy. Na przykład, litera f stosowana jest (z różnymi indeksami) w bardzo wielu kontekstach: w definicji 4.2 jako funkcja macierzy słów-dokumentów zwracająca macierz słów dokumentów, dalej w tekście (z indeksem górnym P) jako funkcja trzech argumentów zwracająca macierz służącą do wzbogacania semantycznego, a w dalszej części tekstu jako klasyfikator oraz miara jego jakości. Znacznie lepiej byłoby użyć w każdym z tych przypadków innej litery.

Dodatkowo, symbole nie zawsze są poprawnie zdefiniowane, np. na str. 23 funkcja f^H macierzy liczona jest element po elemencie, co nie zostało jasno napisane. W definicji 4.3 $\mathcal{C}$, $\mathcal{C}^A$ i $\mathcal{C}^B$ są wielozbiorami, nie zbiorami, gdyż zawierają wiele kopii tego samego elementu. Ich użyteczność jest zresztą wątpliwa, wystarczyłby sam wektor C. Poza tym w definicji 4.4 $\mathcal{C}^A$ i $\mathcal{C}^B$ używane są niepoprawnie, w charakterze zdarzeń losowych.

W pracy występują też błędy merytoryczne. Na przykład, stwierdzenie w linii 11 na stronie 43 jest nieprawdziwe: fakt, że brak korelacji nie oznacza niezależności, jest powszechnie znany. W definicji 5.5 cp_A i cp_B zdefiniowano

jako $P(X|C = A)$ i $P(X|C = B)$, dlaczego więc $cp_A > cp_B$ decyduje o przynależności do klasy? Należało w tym wypadku zastosować wzór Bayesa i uwzględnić prawdopodobieństwa apriori.

Drugim problemem pracy jest opis stanu wiedzy oraz dobór metod klasyfikacji tekstu, z którymi autor porównuje swoje podejście.

Przegląd literatury w rozdziale 2.1 urywa się na roku 2006 - dlaczego? Dodatkowo jest on zbyt skrótowy, autor wspomina o ponad stu artykułach związanych z tematem, należałoby przynajmniej podsumować najważniejsze kierunki badań.

Jako metodę klasyfikacji tekstów autor stosuje jedynie SVM. Nie używa innych metod, takich jak np. regresja z karą LASSO czy elastic net, które są obecnie popularne w analizie tekstów. Brak informacji o tym, czy zastosowano wstępne przetwarzanie danych, takie jak *stemming*, usuwanie słów stopu itp. Brak odniesień do metod redukcji wymiarowości i analizy tematycznej, takich jak ukryte indeksowanie semantyczne czy ukryta alokacja Dirichleta. Trudno stwierdzić, czy autor zna nowsze metody analizy danych tekstowych, które są obecnie w powszechnym użyciu.

W tabeli 6.17 metoda proponowana przez autora porównana jest z innymi podejściami. Brakuje w niej jednak metod z tabeli 2.1, które korzystają z sieci semantycznych, są zatem najbliższe propozycjom przedstawionym w pracy. Dodatkowo, najnowsza praca uwzględniona w tabeli pochodzi z 2002 roku, czy od tego czasu nie pojawiły się żadne inne (lepsze) podejścia?

Powyższe braki utrudniają ocenę znaczenia proponowanego podejścia dla problemu klasyfikacji tekstów.

W pracy autor przedstawił szereg niestandardowych miar jakości modelu. Część z nich jest interesująca i wydaje się być wartościowa, np. miary wykorzystujące aspekty semantyczne. Inne wydają się mniej przydatne. Na przykład miara zdefiniowana w rozdziale 5.3, wykorzystująca ortogonalizację niesymetrycznie, traktuje obie klasy, w związku z czym jej użyteczność jest wątpliwa. Dlaczego nie skorzystać z klasycznych metod wizualizacji, takich jak analiza komponentów głównych czy analiza dyskryminacyjna Fishera? Oprócz miary F_1 autor nie stosuje innych klasycznych wskaźników, takich jak krzywe ROC.

Inne uwagi i literówki:

strona 24, „edge that carry" powinno być „edge that carries"

strona 25, opis podejścia *unattended pruning* jest nieprecyzyjny, lepszy byłby opis algorytmu w formie pseudokodu.

strona 27 nad rysunkiem, powinno być „new nodes belonging to class B"

Rys. 4.5. Możliwy wydaje się też (przez symetrię) wariant z czterema rodzajami krawędzi, jak na rysunku, ale nie biorąc po uwagę wartości IG.

W konkluzji stwierdzam, że pomimo opisanych wyżej braków **praca spełnia wymagania stawiane rozprawom doktorskim przez obowiązującą ustawę i wnioskuję o dopuszczenie mgr. Matykiewicza do dalszych etapów przewodu doktorskiego.**

Szymon Grabowski