wpł. RDIT
15.09.2022
M. Skow

Warszawa, 12 września 2022 r.

dr hab. inż. Robert Nowak, prof. uczelni
Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
ul. Nowowiejska 15/19
00-665 Warszawa

Recenzja rozprawy doktorskiej mgr inż. Aleksandra Płaczka zatytułowanej „Wsparcie diagnostyki onkologicznej w opartym o dane omiczne procesie wielokryterianego wnioskowania parametryzowanego obrazem klinicznym pacjentów o podobnym profilu"

Recenzja powstała na prośbę Przewodniczącego Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej z dnia 27.07.2022, na podstawie: rozprawy doktorskiej liczącej 162 strony z czerwca 2022 r, dorobku naukowego Doktoranta uwzględnionego w bazach Scoupus i Google Scholar oraz kodów źródłowych prezentowanych rozwiązań udostępnionych przez Doktoranta.

1 Tematyka badań

Przedstawiona rozprawa doktorska przedstawia nową metodę oceny złośliwości guzków tarczycy stosując jednocześnie dane kliniczne oraz poziomy ekspresji wybranych genów. Przedstawiona metoda diagnostyczna pozwala zmniejszyć ilość niepotrzebnych zabiegów chirurgicznych.

Cel badawczy jest postawiony właściwie, jest on interesujący i istotny. Przedstawione rozwiązania są poprawne i potwierdzone eksperymentem.

2 Główne wyniki rozprawy

W rozprawie w sposób właściwy przeprowadzono analizę problemu. Opis wymagań na system diagnostyczny oraz opis zastosowań przedstawionej metody w praktyce klinicznej jest bardzo szczegółowy i dokładny. Uzyskanie tak szczegółowej analizy wymagań wymagało znacznej wiedzy medycznej, w tym przeglądu literatury światowej. Autor cytuje 160 źródeł. Wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonujący, pozwoliły one postawić opisane w pracy problemy badawcze.

Rozprawa ma 6 rozdziałów: wprowadzenie (rozdział 1), podsumowanie (rozdział 6), opis metod diagnostycznych guzków tarczycy (rozdział 2), opis sieci bayesowskiej (rozdział 3) oraz rozdziały 4 i 5 opisujące metody utworzone przez Doktoranta i ich badania.

Mgr inż. Aleksander Płaczek użył właściwej metody do rozwiązania postawionego problemu. Został on zdekomponowany na trzy pod-problemy: (1) analiza raportów cytologicznych, (2) obliczanie ryzyka złośliwości guzka na podstawie danych klinicznych

oraz (3) uwzględnienie danych klinicznych przy obróbce wyników ekspresji wybranych genów (nazywanych dalej danymi omicznymi). Wyniki algorytmów cząstkowych, wymienionych wyżej, zostały wykorzystane do prognozowania złośliwości guzka tarczycy na podstawie danych klinicznych oraz omicznych.

Każdy z problemów został rozwiązany poprzez opracowanie algorytmu, następnie budowę programu komputerowego w środowisku R, uruchomienie programu dla danych rzeczywistych pozyskanych z ośrodków medycznych i analizę jakości uzyskanych wyników.

Pierwszym pod-problemem, opisanym w rozdziale 4.1 oraz 5.1, jest ustalanie kategorii złośliwości guzków tarczycy, zwaną dalej kategorią Bethesda, na podstawie na podstawie tekstu raportu cytologicznego. Wykorzystano znane metody przetwarzania języka naturalnego i wyszukiwania informacji w tekście niesformatowanym. Potok przetwarzania jest typowy i wielokrotnie opisywany w literaturze, bazuje on wagach słów w oparciu o liczbę ich wystąpień (TF-IDF) i analizach N-gramów. Istniejące metody poprawnie dostosowano do przedstawionego problemu, tworząc program komputerowy o nazwie *RetroNGSC*. Poprawnie zauważono, że raporty cytologiczne nie są tworzone z myślą o automatycznej obróbce, dodatkowo różni specjaliści stosują różne słownictwo, w tym pojawiają się zwroty zmieniające znaczenie zdań (np. 'brak', 'nie znaleziono'). Algorytm testowano na zbiorze 3200 opisów uzyskując, w 10-krotnej walidacji krzyżowej, dokładność około 95%. Algorytm jest wykorzystywany do korekty decyzji specjalisty, w szczególności dla przypadków niejednoznacznych (skala Bethesda 3), oraz jako źródło atrybutów służące do grupowania podobnych przypadków, wykorzystywane do prognozowania złośliwości guzka tarczycy.

Drugim pod-problemem, opisanym w rozdziale 4.2 oraz 5.2, jest algorytm oraz program komputerowy do obliczania ryzyka klinicznego na podstawie danych klinicznych. Uwzględniono 9 atrybutów: wiek, płeć, średnica guza, echogeniczność, średnica guza wg echogeniczności, kategoryzacja Bethesda, kształt zmiany, wieloogniskowość i zajęcie węzłów chłonnych. Doktorant wykorzystał sieci bayesowskie dla danych dyskretnych, dane ciągłe były przekształcane do postaci dyskretnej (dyskretyzowane) stosując progi. Algorytmy są typowe i podobnie jak w problemie pierwszym doktorant poprawnie dostosował je do przedstawionego problemu. Opracowana sieć pozwala m.in. wyznaczyć najbardziej prawdopodobną kategorię Bethesda (nazywaną w pracy 'zweryfikowaną kategorią Bethesda') oraz prawdopodobieństwa złośliwości zmiany nowotworowej dla poszczególnych kategorii Bethesda.

Trzecim pod-problemem, opisanym w rozdziale 4.6 oraz 5.3 jest wykorzystanie danych omicznych. Wybrano, na podstawie literatury, jeden z wcześniej używanych zbiorów genów, uznanych za właściwe w sygnalizowaniu złośliwości nowotworów. W cytowanych badaniach nie było ustalone, czy ekspresja poszczególnych genów samodzielnie pozwalają różnicować próbki łagodne od złośliwych. Dane o ekspresji genów (ciągłe) przekształcono do postaci dyskretnej wykorzystując ryzyko kliniczne w ten sposób, że próbki grupowano uwzględniając ryzyko kliniczne obliczone przez sieć bayesowską, a

następnie wykonywano niezależną dyskretyzacja dla każdego genu w każdej grupie. W ten sposób uzyskano możliwość separacji próbek z guzków łagodnych i złośliwych na podstawie dyskretnej informacji o ekspresji genów. Metody opisane wyżej została użyta poprawnie i poprawnie dostosowano ją do przedstawionych danych.

Do łączenia uzyskanych wyników wykorzystano sieć bayesowską, która pracuje na danych dyskretnych. Pozwala to analizować próbki, które nie mają wszystkich cech (sieć dostarcza wynik także dla rekordów, które nie mają określonych wszystkich atrybutów), a dodatkowo informacja jest czytelna dla człowieka. Doktorant opisał wyniki wielu badań przedstawionego rozwiązania, badał m.in.: jakość uzyskiwanych wyników, istotność poszczególnych cech, istotność ekspresji poszczególnych genów na kategorię guzka. Dodatkowo wyniki algorytmu zostały porównane z ocenami ekspertów.

Uzyskane wyniki są przedstawione poprawnie. Wywód jest zwięzły i jasny. Nie stwierdzam znaczących błędów redakcyjnych. W szczególności opracowana metoda jest szeroko dyskutowana w kontekście jej użycia w praktyce klinicznej.

Dostarczone oprogramowanie działa poprawnie, kod jest czytelny.

3 Elementy krytyczne

Autor nie odnosi się wprost do wyników prezentowanych przez innych badaczy pokazując wartości liczbowe uzyskiwane w innych badaniach.

Przykładowo, autor porównuje wyniki uzyskane dla przez opracowane narzędzie do analizy raportów cytologicznych (RetroNGSC) uruchamiając algorytmy dostępne w bibliotekach dla dostarczonych danych. Brak jest odwołań do wyników publikowanych w literaturze, przez inne zespoły, być może na innych danych. Nie wiadomo więc, czy osiągnięty wynik (dokładność 95%) jest porównywalny z wynikami innych badaczy. Sugeruję przejrzeć aktualną literaturę i odpowiednie wartości przedstawić na obronie rozprawy. Oczywiście często kluczowa jest jakość danych i wielkość zbioru zbioru, język wykorzystywany do tworzenia raportów itd., niemniej takie odniesienie zgrubnie pozwala upewnić się, że uzyskany wynik jest poprawny.

Moim zdaniem w pracy zabrakło szerszego omówienia algorytmów łączenia danych i pokazania własnego rozwiązania w tym kontekście.

Łączenie (fuzja) danych może odbywać się na wielu poziomach (poziom danych, logiczny, itd) opisanych w tzw. modelach fuzji danych. Jednym z bardziej znanych jest model JDL. Odniesienie do tego modelu (lub innego, który jest rozwinięciem JDL) wydaje mi się cenne.

W szczególności nie było jawnie stwierdzone, czy w pracy są elementy integracji danych, a nie tylko fuzji danych. Przy wykorzystaniu danych z różnych źródeł istnieją problemy związane z integracją, np. problem z identyfikacją, że dwa rekordy z różnych źródeł opisują ten sam obiekt rzeczywisty, problem wspólnego identyfikatora. Rozumiem, że dla rozważanych danych taki problem nie wystąpił, jednak warto byłoby napisać, skąd było wiadomo, że dane omiczne pochodzą z tego samego guzka, co raport cytologiczny

i dane kliniczne.

Kolejna uwaga dotyczy wybranego algorytmu: wykorzystanie sieci bayesowskiej dla wartości dyskretnych jest tylko jedna z możliwości łączenia danych. Warto byłoby wymienić inne metody łączenia, podobnie jak są wymieniane różne metody uczenia maszynowego.

Co mnie najbardziej zastanawia, to nie jest dla mnie do końca jasne, dlaczego został wybrany algorytm dyskretnej, a nie ciągłej sieci bayesowskiej. Algorytmy pracujące na danych ciągłych nie wymagają dyskretyzacji, która, bazując na wnioskach z tej pracy, w istotny sposób wpływa na jakość uzyskiwanych wyników. Proszę sprawdzić hasło 'Gaussian Bayes Classifier'. Moim zdaniem ta decyzja wymaga uzasadnienia na obronie rozprawy.

Mam także uwagi do sformułowań w tekście pracy. Myślę, że przez nieuwagę, są dam zamieszczone zdania, które są nieprawdziwe. Na stronie 33 jest zdanie: 'Należy jednak założyć, że nie wszystkie zależności rzeczywistych zmiennych są odwzorowywane w sieci, gdyż wymagałoby to olbrzymich reprezentatywnych zbiorów danych, z nieskończoną liczbą cech oraz niemal nieskończonych zasobów obliczeniowych.' Nieprawdą jest, że potrzebujemy nieskończonej liczby cech.

Na stronie 22 jest napisane: 'Uczenie maszynowe to metoda wnioskowania polegająca na budowaniu generycznych schematów klasyfikacji na podstawie dostępnych danych.' Klasyfikacja to tylko jeden z problemów, które potrafią rozwiązywać metody uczenia maszynowego. Innym problemem jest np. regresja.

Dalej na tej stronie: 'W procesie uczenia stosuje się metody zapobiegające przeuczeniu np: walidację krzyżową w celu utrzymania określonego poziomu generalizacji modelu'. Walidacja krzyżowa jest metodą pomiaru jakości, a nie metodą zapobiegającą przeuczeniu. W szczególności walidacja krzyżowa nie zmienia modelu. Metodą zapobiegającą przeuczeniu jest np. przycinanie drzew dla algorytmów opartych o drzewo decyzyjne.

W Tablicy 2.2, gdzie wymieniono popularne metody uczenia maszynowego zabrakło sztucznych sieci neuronowych. Moim zdaniem obecnie jest to najbardziej popularna metoda.

Kod źródłowy jest czytelny, jednak moim zdaniem warto byłoby go przejrzeć i uzupełnić, aby był łatwy w użyciu i w pielęgnacji. Sugeruję podzielenie na moduły (pliki) o jednoznacznej odpowiedzialności, usunięcie nieużywanych fragmentów, uzupełnienie komentarzy, dostarczenie testów jednostkowych oraz rozszerzonej dokumentacji projektowej i użytkownika (instrukcja użycia, instrukcja instalacji, zależności). Krótka instrukcja użytkownika, znaczenie poszczególnych parametrów mogłaby także być wbudowana w kod.

Powyżej przedstawione elementy krytycznie nie zmieniają pozytywnej konkluzji końcowej przedstawionej pracy. Stanowią podstawę do dyskusji na publicznej obronie rozprawy.

4 Ocena dorobku naukowego

W bazach Scopus oraz Google Scholar dorobek mgr inż. Aleksandra Płaczka zawiera jeden artykuł w IEEE Access, opublikowany w 2020, pt. „Bayesian Assessment of Diagnostic Strategy for a Thyroid Nodule Involving a Combination of Clinical Synthetic Features and Molecular Data", w którym Doktorant jest pierwszym autorem, oraz 5 prac w materiałach 3 konferencji: Advances in Intelligent Systems and Coputing, Communications in Computer and Information Science, CEUR (Central Europe) Workshop Proceedings, opublikowanych w latach 2016-2022, w których doktorant jest współautorem. Prace te są cytowane kilkanaście razy. W mojej ocenie dorobek jest wystarczający aby dopuścić mgr inż. Aleksandra Płaczka do dalszych etapów przewodu doktorskiego.

5 Podsumowanie

Temat badawczy uważam za bardzo istotny, teza jest poprawna i oryginalna, wykazana w stopniu wyczerpującym. Opracowane rozwiązanie jest nowatorskie, dostarcza wyniki na poziomie prezentowanym w literaturze światowej, a ponadto ma wysoki poziom gotowości technologicznej.

Stwierdzam, że **recenzowana rozprawa doktorska mgr inż. Aleksandra Płaczka spełnia warunki** określone w aktualnych przepisach i wnioskuję do Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej o dopuszczenie rozprawy doktorskiej mgr inż. Aleksandra Płaczka do publicznej obrony.

Robert Nowak