Sosnowiec, 20 sierpnia 2021r.

dr hab. Beata Zielosko, prof. UŚ
Instytut Informatyki
Wydział Nauk Ścisłych i Technicznych
Uniwersytet Śląski w Katowicach

Recenzja rozprawy doktorskiej

Tytuł rozprawy: **SELECTED ASPECTS OF INTERACTIVE FEATURE EXTRACTION**

Autor rozprawy: **mgr Marek Grzegorowski**
Przewód doktorski: w dziedzinie nauk matematycznych, w dyscyplinie Informatyka
Promotor rozprawy: **prof. dr hab. Dominik Ślęzak**
Promotor pomocniczy: **dr Andrzej Janusz**

Recenzja wykonana na podstawie zlecenia Rady Naukowej Dyscyplin Matematyka i Informatyka Uniwersytetu Warszawskiego, pismo z 29 czerwca 2021r.

1. Obszar problemowy pracy

Rozwój technologiczny wiąże się z koniecznością radzenia sobie z coraz większą ilością danych, które należy nie tylko przechowywać ale przede wszystkim przetwarzać, w celu wydobycia użytecznej wiedzy wspomagającej procesy podejmowania decyzji na różnych szczeblach i w różnych dziedzinach życia. Sytuacja ta jest szczególnie trudna ze względu na ogromną ilość danych oraz występujące w nich błędy i ewentualne braki. W związku z tym, etap przetwarzania danych i ich przygotowania odgrywa istotną rolę w procesie eksploracji danych. Wyniki obróbki danych przekładają się na rezultat kolejnych faz procesu data mining, w tym jakość predykcji, czas obliczeń, wymagane zasoby oraz możliwość interpretacji uzyskanych wyników. Istotnymi etapami w procesie przygotowania danych są ekstrakcja cech oraz ich selekcja. W literaturze opisano wiele podejść i algorytmów stosowanych w tym zakresie, jednak znacząca większość z nich pomija kwestię użycia wiedzy dziedzinowej a przecież interakcja z ekspertami odgrywa istotną rolę zwłaszcza, kiedy wyniki analiz mogą mieć wpływ na bezpieczeństwo i życie ludzi. Metody te są istotne także ze względu na proces budowy i walidacji zespołów modeli uczących (ang. ensemble learning) tworzonych w oparciu o wybrane podzbiory cech.

Recenzowana praca dotyczy wymienionego wyżej obszaru badanego w kontekście interaktywnej ekstrakcji cech. Główne wyniki rozprawy dotyczą nowych metod konstrukcji i doboru cech umożliwiając wyrażenie wiedzy dziedzinowej związanej z najważniejszymi podzbiorami atrybutów. Wiedza dziedzinowa, której źródłem są eksperci dziedzinowi (np. operatorzy maszyn, lekarze) zazwyczaj stanowi cenne źródło informacji pozwalające dobrać właściwy podzbiór cech a tym sam podnieść jakość modeli predykcyjnych. Dotyczy to m.in. aproksymacji pojęć złożonych, opisanych za pomocą danych o charakterze tymczasowym. Tworzenie zespołów modeli uczących jest popularnym kierunkiem rozwoju metod budowy modeli predykcyjnych, zwłaszcza w odniesieniu do przetwarzania dużych zbiorów danych (ang. big data).

Autor proponuje metodę „odpornej" (ang. resilient) selekcji cech opartą na podejściu znanym z teorii zbiorów przybliżonych czyli reduktach oraz wykorzystaniu paradygmatu granulacji informacji m.in. w kontekście agregacji atrybutów. Metoda ta pozwala na uodpornienie algorytmu selekcji cech na ewentualne braki w danych a jednocześnie zachowanie satysfakcjonującego poziomu jakości modeli predykcyjnych. Metoda ta została wykorzystana jako element proponowanego przez Autora podejścia do procesu interaktywnej ekstrakcji cech dla złożonych, rzeczywistych problemów, wymagających przetwarzania dużych zbiorów danych pochodzących z różnego rodzaju czujników. Zaproponowana metodologia pozwala na tworzenie zespołu modeli predykcyjnych uzyskanych na podstawie analizy wielowymiarowych szeregów czasowych dotyczących danych z różnych dziedzin.

Podsumowując tę część rozprawy stwierdzam, że Autor podjął ważny, aktualny a zarazem trudny problem naukowy. Przeprowadził analizę istniejących podejść, zaproponował autorskie rozwiązania, przedstawił ich zastosowania w wybranych projektach naukowych oraz zweryfikował ich skuteczność m.in. w ramach międzynarodowych konkursów analizy danych, co pozwoliło ocenić opracowane rozwiązania.

2. Kompozycja i zawartość pracy

Rozprawa została napisana w języku angielskim i składa się z sześciu rozdziałów, bibliografii oraz dwóch załączników.

Rozdział pierwszy stanowi wstęp do tematyki rozprawy. Autor przedstawił motywacje do podjęcia badań dotyczących interaktywnych tj. uwzgledniających wiedzę ekspertów dziedzinowych, metod konstrukcji i selekcji cech, strukturę rozprawy, główny wkład związany z realizacją rozprawy z podziałem na poszczególne rozdziały oraz podziękowania. W tej części rozprawy można zauważyć brak wyraźnego podkreślenia celu pracy i głównych zadań

badawczych. Ich wyeksponowanie pozwoliłoby lepiej zobrazować stopień złożoności pracy badawczej niezbędnej do zrealizowania doktoratu oraz zakres rozprawy.

Rozdział drugi stanowi przegląd literatury dotyczącej ekstrakcji cech. Autor przedstawił różnorodne techniki przygotowania różnego rodzaju danych (tekst, obraz, dane strukturalne, sensoryczne) do pracy z algorytmami uczenia maszynowego, w tym technikę przesuwnego okna czasowego stosowaną podczas analizy szeregów czasowych. Dalsza część rozdziału zawiera opis wykorzystania paradygmatu granulacji informacji w procesie tworzenia nowych atrybutów oraz metody selekcji cech stosowane m.in. w teorii zbiorów przybliżonych i oparte na reduktach. Autor przedstawił aktualny stan badań w obszarze którego dotyczy rozprawa, a sposób ich prezentacji można określić jako uogólnione resume literatury przedmiotu, co potwierdza szeroką wiedzę Autora.

Rozdział trzeci przedstawia zaproponowaną przez Autora metodę „odpornej" selekcji cech opartą na tzw. r-C-reduktach czyli nieredukowalnych podzbiorach atrybutów zapewniających zachowanie satysfakcjonującego poziomu jakości modelu predykcyjnego, według zadanego kryterium C, nawet po usunięciu r atrybutów. Autor podał definicję uogólnionego kryterium jakości atrybutów C, które zostało zastosowane w odniesieniu do reduktu dokładnego oraz przybliżonego. W ramach wyników teoretycznych Autor wykazał, że każde NP-trudne zadanie selekcji atrybutów polegające na znalezieniu minimalnego C-reduktu pozostaje NP-trudne również w wersji znalezienia r-C-reduktu. W kolejnej części rozdziału Autor przedstawił podejścia i algorytmy jakie można zastosować do wyszukiwania r-C-reduktów, m.in. na podstawie przeszukiwania kraty atrybutów wszerz (ang. breadth first search) oraz przeszukiwania kraty atrybutów w głąb (ang. depth first search). Został także zaproponowany nowy algorytm do wyznaczenia r-C-reduktu, wykorzystujący technikę permutacji.

W rozdziale czwartym Autor przedstawił podejście do tworzenia modeli predykcyjnych na podstawie dużych zbiorów danych pochodzących z różnego rodzaju czujników. Ważnym elementem tego podejścia jest wykorzystanie techniki okien przesuwnych w procesie ekstrakcji cech oraz tworzenie zespołów modeli na podstawie selekcji podzbiorów atrybutów. Pierwsza cześć rozdziału poświęcona jest dyskusji dotyczącej charakteru danych pozyskiwanych w rzeczywistym środowisku oraz technice przetwarzania wielowymiarowych szeregów czasowych z uwzględnieniem atrybutów pozyskanych od ekspertów dziedzinowych. Kolejna część rozdziału poświęcona jest granulacji atrybutów, stosowanej w procesie selekcji cech, jak i w kontekście przetwarzania dużych, złożonych zbiorów danych. Ostatnia część rozdziału przedstawia wykorzystanie

opisanych technik jak i zaproponowanych metod konstruowania i wyboru atrybutów do tworzenia zespołu zdywersyfikowanych modeli predykcyjnych.

Rozdział piąty dotyczy praktycznych zastosowań oraz ich ewaluacji a zarazem obrazuje efektywność zaproponowanych przez Autora metod. Badania eksperymentalne obejmują różne dziedziny i dotyczą rzeczywistych, złożonych problemów. Dla poszczególnych zastosowań Autor opisał złożoność i specyfikę problemu, użyte miary efektywności oraz uzyskane rezultaty (w formie opisowej, graficznej i liczbowej). Do najważniejszych zastosowań należą: (i) monitorowanie zagrożenia metanowego i sejsmicznego w kopalni węgla kamiennego; opracowany system wspomagania decyzji DISESOR wykorzystywał zaproponowane przez Autora rozprawy metody, (ii) monitorowanie aktywności i funkcji życiowych strażaków podczas akcji pożarniczych, (iii) predykcja cen zasobów w chmurach obliczeniowych. Nie bez znaczenia jest także udział Autora w zawodach data mining organizowanych podczas międzynarodowych konferencji (IJCRS'15, AAIA'16), gdzie na podstawie danych uzyskanych z czujników sensorycznych umieszczonych w kopalni węgla kamiennego organizatorzy zawodów upublicznili zbiór danych i zdefiniowali odpowiadający mu problem.

Rozdział szósty stanowi podsumowanie rozprawy. Autor wskazał główne osiągnięcia i możliwe kierunki kontynuacji prac badawczych.

Rozprawę kończy rozległa bibliografia, cytowane prace to przede wszystkim publikacje z czasopism naukowych oraz prace konferencyjne. Nie mam zastrzeżeń do doboru literatury, jednak w moim odczuciu jest on zbyt obszerny.

W pracy zamieszczono także dwa dodatki. W dodatku A przedstawiono charakterystykę danych uzyskanych z czujników zlokalizowanych w kopalni węgla kamiennego i wykorzystanych podczas międzynarodowych zawodów data mining IJCRS'15 oraz AAIA'16. Został tam również zawarty opis atrybutów dotyczących zachowań strażaków i dane dotyczące cen zasobów chmurowych. Dodatek B przedstawia krótki opis stosowanych przez ekspertów metod do klasyfikacji zagrożeń sejsmicznych.

Do pracy została także dołączona płyta DVD zawierająca zbiory danych, które podlegały analizie.

Podsumowując tę część recenzji uważam, iż praca jest obszerna i pokazuje dużą wiedzę Doktoranta. Widoczne jest naukowe podejście do realizowanego zadania. Rozprawa przedstawia dużą zawartość merytoryczną, w tym naukową oraz implementacyjną związaną z przedstawionymi zastosowaniami.

3. Opinia o rozprawie i oryginalność osiągnięć

Przede wszystkim należy podkreślić, iż problem badawczy podejmowany w rozprawie jest bardzo ważny i złożony. Zastosowanie interaktywnej konstrukcji i selekcji cech w procesie tworzenia modeli predykcyjnych i odkrywania wiedzy stanowi istotną kwestię poruszaną przez wielu badaczy. Dotyczy to zwłaszcza zjawisk obserwowanych w dłuższym okresie czasu, opisywanych przez obiekty o złożonej strukturze, które mogą ulegać zmianom.

Za jedno z najważniejszych osiągnięć przedstawionych w pracy uważam metodę selekcji cech opartą o tzw. r-C-redukty. Należy tutaj zwrócić uwagę na uniwersalność metody rozumianą w dwóch aspektach: (i) możliwość wykorzystania wielu różnych kryteriów m.in., stosowanych w teorii zbiorów przybliżonych podczas konstruowania reduktów, zapewniających, że proponowany podzbiór cech jest wystarczająco dobry z punktu widzenia informacji o zmiennej docelowej oraz (ii) możliwość wykorzystania metody nawet w sytuacji kiedy pewne atrybuty są chwilowo niedostępne, poprzez zaproponowane kryterium odporności. Autor dokonał analizy właściwości r-C-reduktu oraz udowodnił, że jeśli problem konstruowania reduktu o minimalnej liczności w oparciu o kryterium C jest NP-trudny to problem znalezienia r-C-reduktu o minimalnej liczności jest także NP-trudny. Ważnym elementem zaproponowanej przez Autora metody jest analiza dotycząca opracowania algorytmów dla problemu znalezienia r-C-reduktu oraz analiza złożoności obliczeniowej wybranych algorytmów. Algorytm 2, wykorzystując właściwości algorytmu Apriorii dotyczące przeszukiwania kraty atrybutów, dla danej wartości r, konstruuje wszystkie r-C-nadredukty lub minimalne r-C-redukty. Algorytmy 4 i 5, wykorzystując technikę przeszukiwania grafu w głąb oraz permutacje zbioru atrybutów A, konstruują odpowiednio, r-C-redukt i r-C-nadredukt, dla oczekiwanego poziomu odporności r.

Drugi w mojej opinii interesujący a zarazem złożony aspekt pracy to zaproponowane podejście do tworzenia zespołów modeli predykcyjnych na podstawie dużych zbiorów danych pochodzących z różnego rodzaju czujników. Podejście to wykorzystuje stosowaną dla wielowymiarowych szeregów czasowych metodę automatycznej ekstrakcji cech opartą na technice przesuwnego okna czasowego (ang. time sliding window), rozszerzoną o generowanie cech na podstawie ciągu okien czasowych opisujących obserwowany obiekt oraz wprowadzenie tzw. inter window features, wyrażających zmiany i trendy w danych zachodzące pomiędzy parami sąsiadujących okien. Duże znaczenie w tym procesie ekstrakcji cech odgrywa także interakcja z ekspertami dziedzinowymi i analitykami, którzy mogą ocenić istotność uzyskanych atrybutów jak i zaproponować własne. Istotnym elementem przedstawionego podejścia jest także wykorzystanie paradygmatu

granulacji informacji w kontekście atrybutów w procesie selekcji cech mającej na celu jak największe zróżnicowanie granuli atrybutowych wykorzystanych w późniejszych etapach do trenowania modeli.

Opis jak i przedstawienie w formie graficznej (Rysunek 4.5, 4.6 i 4.7) poszczególnych etapów proponowanego podejścia pokazuje złożoność i specyfikę problemu w przypadku pracy z dużymi zbiorami danych strumieniowych oraz podkreśla znaczenie procesu ekstrakcji i „odpornej" selekcji cech, w tworzeniu zespołu modeli predykcyjnych na podstawie różnych podzbiorów atrybutów i różnych próbek danych.

Kolejnym, ważnym elementem rozprawy jest aspekt praktyczny obrazujący różnorodne zastosowania zaproponowanych metod, w tym zastosowania realizowanie w ramach projektów naukowych jak i w ramach międzynarodowych konkursów analizy danych. Należy także zaznaczyć, iż Pan Grzegorowski zajął trzecie miejsce wraz ze współautorem w ramach konkursu organizowanego na konferencji IJCRS'15 oraz pierwsze miejsce w ramach konkursu organizowanego na konferencji AAIA'16.

Podsumowując uważam, że wykonana w ramach doktoratu praca stanowi istotny wkład w rozwój metod dotyczących interaktywnej ekstrakcji i selekcji cech. Rozprawa prezentuje wysoki poziom naukowy i w pełni zasługuje na ocenę pozytywną.

Praca napisana jest starannie, bez większych pomyłek językowych. Zawarte w rozprawie przykłady i ilustracje pozwalają lepiej zobrazować i wyjaśnić działanie proponowanych metod i algorytmów.

Ponadto należy zaznaczyć, iż Pan Marek Grzegorowski jest autorem lub współautorem szeregu wartościowych publikacji (14 prac w bazie Web of Science Core Collection oraz 15 prac w bazie DBLP, w tym dwie prace w czasopiśmie Information Sciences (stan na koniec sierpnia 2021)).

4. Uwagi i problemy do dyskusji

W pracy brakuje wyraźnie sformułowanego celu pracy, nie są dostatecznie wyeksponowane zadania badawcze niezbędne do realizacji postawionego celu. Co prawda Autor w rozdziale pierwszym opisał tematykę i zakres swojej pracy oraz przedstawił motywacje do podjęcia realizowanego tematu, jednak w moim odczuciu wyróżnienie pewnych elementów w treści tekstu jasno wskazuje niezbędne zadania do wykonania, zarówno naukowe jak i praktyczne, a tym samym podkreśla nakład pracy ze strony Doktoranta.

Rozdział Main Contributions opisuje wkład Autora w poszczególne rozdziały rozprawy. Wydaje mi się, że określenie Main Achievements lepiej

odwzorowałoby główne osiągnięcia pracy, które poza opisem w ramach poszczególnych rozdziałów powinny zostać wypunktowane.

W zakresie ekstrakcji cech można wyróżnić cztery aspekty: konstrukcja cech; generowanie podzbiorów cech; definicja kryterium oceny; oszacowanie kryterium oceny (lub metoda oceny). Nawiązując do tematu pracy, oczekiwałabym większego podkreślenia w rozprawie, np. w rozdziale drugim „Feature Extraction", iż przedstawione wyniki badań odzwierciedlają różne aspekty procesu konstruowania cech.

W rozprawie brakuje spisu tabel, rysunków i algorytmów, co ułatwiłoby nawigację po treści pracy.

Dla zaproponowanej metody selekcji cech dotyczącej r-C-reduktów Definicje 5 i 7 odnośnie funkcji kryterium C opierają się na właściwości monotoniczności (w odniesieniu do zawierania w zbiorze), czy możliwe jest zastosowanie tych definicji dla niemonotonicznych miar?

Czy Autor widzi możliwość automatyzacji całego procesu eksploracji danych tj. stworzenia systemu, który wchodząc w interakcje z ekspertem (analitykiem) zapewni odpowiednie przygotowanie danych (w tym ekstrakcja i selekcja cech) i utworzenie modeli predykcyjnych pozwalających na uzyskanie wyników na satysfakcjonującym poziomie? Czy system mógłby być na tyle uniwersalny aby pozwalał na prace z danymi z różnych dziedzin, czyli także zapewniał jakąś ogólną metodykę dotyczącą konstrukcji cech definiowanych przez eksperta dziedzinowego?

Zanim przejdę do wniosków chciałabym zaznaczyć, że opisane powyżej uwagi nie mają istotnego wpływu na jakość i wagę przedstawionych rozwiązań i w żadnym stopniu nie obniżają wartości pracy a zadane pytania wynikają z chęci podjęcia dyskusji.

5. Wnioski

Pan Marek Grzegorowski przedstawił rozprawę doktorską stanowiącą oryginalne rozwiązanie problemu naukowego z zakresu wykorzystania wiedzy dziedzinowej w procesach ekstrakcji i selekcji cech oraz opracowania podejścia do tworzenia zespołów modeli dostosowanych do przetwarzania dużych zbiorów danych. Opracowane metody zostały zastosowane i zweryfikowane dla różnych problemów rzeczywistych.

Autor wykazał się dużą wiedzą w zakresie tematyki rozprawy, umiejętnością pracy naukowej oraz znajomością metod badawczych. Osiągnięte wyniki świadczą o bardzo dobrym przygotowaniu Autora do pracy naukowej.

Recenzowana praca spełnia wszystkie wymagania stawiane rozprawom doktorskim przez ustawę o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki. Zatem wnoszę, by mgr Marek Grzegorowski został dopuszczony do publicznej obrony.

Beata Zielosko