

Recenzja rozprawy doktorskiej

Autor: mgr Tomasz Klonecki

Tytuł: Cost-constrained feature selection using information theory

Promotor: dr hab. Paweł Teisseyre

Ogólna charakterystyka rozprawy

Przedłożona do recenzji rozprawa doktorska jest napisana w języku angielskim, liczy 116 stron, składa się z polskiego i angielskiego streszczenia, 8 rozdziałów, dodatku z kilkoma tabelami oraz bardzo obszernej bibliografii obejmującej 169 pozycji.

Praca jest zogniskowana na problematyce konstrukcji klasyfikatorów w sytuacji, gdy pomiar (pozyskanie) cech do klasyfikacji wiąże się z konkretnymi, możliwymi do ustalenia i koniecznymi do poniesienia kosztami. Jako jeden z przekonujących przykładów Doktorant podaje koszty diagnostyki medycznej, gdzie decyzje podejmuje się na podstawie pakietu badań o konkretnych kosztach oraz o pewnej zdefiniowanej strukturze, która może mieć charakter blokowy.

Problematyka jest bardzo dobrze zdefiniowana i całość pracy jest zorientowana na jej wszechstronnej analizie. Badane problemy są bardzo ważne zarówno od strony teoretycznej jak też w aspekcie praktycznych zastosowań. Przedstawione są trzy podejścia do powiązania narzędzi teorii informacji z aspektami kosztowymi klasyfikatorów oraz dokonuje się analizy zarówno symulowanych jak i rzeczywistych zbiorów danych.

Rozdział 1 pracy charakteryzuje motywację do rozwijania zaproponowanych badań, przedstawia przeprowadzone badania na tle szerokiej i bardzo dobrze dobranej literatury, wymienia najważniejsze oryginalne osiągnięcia pracy oraz listę publikacji, których współautorem jest Doktorant, tematycznie ściśle związanych z przeprowadzonymi

badaniami naukowymi. Omawiana jest także struktura i zawartość pracy oraz definiuje się stosowane w pracy oznaczenia.

Rozdział 2 pracy poświęcony jest przedstawieniu metod i mechanizmów selekcji cech w nadzorowanych algorytmach klasyfikacji. W tym rozdziale dyskusję dzieli się na trzy podstawowe zagadnienia. Pierwszym jest ogólne omówienie problemu selekcji cech w nadzorowanej klasyfikacji wraz z przywołaniem szerokich odniesień literaturowych. Drugim zagadnieniem jest problem ukierunkowany na temat doktoratu – dołączenia kosztów uzyskania / pomiaru cech do systemu selekcji cech i do konstrukcji klasyfikatora. Omawia się ważne zagadnienia szczegółowe. Pierwsze z nich to selekcja i koszty cech bazujące na modelu statystycznym (zwykle modelu regresji) lub w systemie „wolnym” od modelu. Drugie zagadnienie to grupowanie (blokowanie) cech w aspekcie ich kosztów. Bardzo często eksperyment, wymagający poniesienia określonego kosztu, generuje nie jedną cechę ale całą ich grupę. Musi to być uwzględnione w konstrukcji algorytmów. Ostatnim zagadnieniem w rozdziale 2 jest klasyfikacja wielo – etykietowa (multi label classification). Definiuje się model klasyfikacji wieloetykietowej. Dokonuje się porównania do modeli jedno i wieloklasowych. Omawia się specyfikę tego modelu, w szczególności problemy obliczeniowe generowane w takim modelu, między innymi wymagające estymacji wielowymiarowych rozkładów prawdopodobieństwa.

Rozdział 3 poświęcony jest zdefiniowaniu pojęć teorii informacji wykorzystywanych w pracy. Definiuje się pojęcia entropii, entropii warunkowej, odległości (dywergencji) Kullbacka Leilblera informacji wzajemnej, warunkowej informacji wzajemnej. Omawia się własności tych miar w aspekcie zastosowania w algorytmach klasyfikacji. Następną ważną miarą, omawianą w rozdziale 3, jest wskaźnik informacji o interakcji pomiędzy zmiennymi losowymi (interaction information, II). Przedstawia się definicję tego wskaźnika oraz jego związki z poprzednio wprowadzonymi miarami (może on być rozumiany jako uogólnienie poprzednich miar informacyjnych). Dyskutuje się jego użyteczność do konstrukcji klasyfikatorów. Wreszcie jako ostatnie zagadnienie w rozdziale 3 omawia się problematykę estymacji miar / wskaźników w teorii informacji.

Rozdział 4 poświęcony jest naświetleniu literaturowego tła dla zagadnienia selekcji cech bazującej na wskaźnikach teorii informacji. Na wstępie formułuje się dwa podstawowe problemy, znalezienia wszystkich relewantnych cech (All Relevant Feature Selection, ARFS) oraz znalezienia optymalnego zbioru cech o minimalnej wielkości (Minimal Optimal Subset Selection, MOSS). Następnie, w podpunkcie 4.2 formułuje się problem MOSS z wykorzystaniem wskaźnika informacji wzajemnej. W kolejnym podpunkcie 4.3 przedstawia się sekwencyjne algorytmy wyboru cech dla problemu MOSS, w przypadku klasyfikacji jedno – etykietowej, a podpunkcie 4.4 dla przypadku klasyfikacji wielo – etykietowej.

W rozdziale 5 rozpoczyna się przedstawianie oryginalnych wyników pracy. W podpunkcie 5.1 formułuje się problem selekcji cech jako zadanie wskaźnika maksymalizacji wzajemnej informacji przy ograniczeniu na koszt wybieranych cech. W podpunkcie 5.2

przedstawiona jest zachłanna procedury wyboru cech z ograniczeniem na koszty oraz jej schematy blokowe (rysunek 5.1). W podpunkcie 5.3 zamieszczone są uwagi dotyczące sposobu doboru współczynnika λ , który skaluje wagę kosztów cech w stosunku do wnoszonego przez nie wkładu do wskaźnika informacji wzajemnej. W podpunkcie 5.4 opisane są eksperymenty obliczeniowe dotyczące zaproponowanych algorytmów oraz ich porównania z wynikami opublikowanymi. Eksperymenty zostały zrealizowane zarówno dla danych symulowanych jak i rzeczywistych. Doktorant stwierdza, że w literaturze / w Internecie raczej brakuje rzeczywistych zbiorów danych odpowiadających sytuacji, gdy oceny / pomiary cech generują konkretne koszty. Proponuje kilka strategii tworzenia takich zbiorów lub uzupełniania danych o cechach o ich dodatkowe koszty. Dokonuje porównań wyników zastosowania różnych algorytmów. Porównania są bardzo szerokie. Obejmują algorytmy selekcji cech jedno - etykietowych MIM, MRMR, JMI oraz wiele – etykietowych AMI, MDMR, MVML, SCLS, STFS, DCRMFS. Doktorant opisuje także oraz dokonuje porównań z literaturową metodą selekcji cech z ograniczeniami na koszty CFSM (Cost-constrained Feature Selection on Multi-label data). Także dobór kryteriów / wskaźników porównawczych jest szeroki i wyczerpujący. Wyniki są prezentowane w postaci tabel i wykresów. Zwraca uwagę dużą przewagą proponowanej metody nad metodą CFSM. W osobnym podpunkcie przedstawiona jest analiza rzeczywistego zbioru danych MIMIC II, dla którego dostępne są wartości kosztów uzyskiwania cech. Przedstawia się analizy wartości wskaźnika AUC (area under the curve) w zależności od budżetu, dla proponowanej metody oraz dla „tradycyjnej” metody selekcji cech.

W rozdziale 6 rozważany jest szczególny przypadek problemu selekcji cech z ograniczeniami kosztowymi, obejmujący częstą sytuację, gdy cechy pozyskuje się grupowo. Podobnie jak w poprzednim rozdziale, na wstępie formalnie definiuje się problem optymalizacji (formuła 6.1). Opisuje się odpowiednio zaprojektowany dla tej sytuacji algorytm selekcji cech CC-GFS (Cost Constrained Group Feature Selection). Dokonuje się porównania tego algorytmu z algorytmem selekcji, w którym grupowa struktura cech nie jest uwzględniana CC-SFS (Cost Constrained Single Feature Selection). Porównuje się algorytmy CC-GFS, CC-SFS oraz „tradycyjny” algorytm selekcji cech dla zbioru danych MIMIC II, dla którego grupowa struktura cech jest opisana w tabeli 6.1. Pokazuje się następnie, na rysunku 6.1 przewagę algorytmów CC-GFS i CC-SFS nad metodą tradycyjną. W podpunkcie przedstawia się dwukrokową modyfikację algorytmu GCC-FS zaprojektowaną tak aby poprawiać jego efektywność.

W rozdziale 7 przedstawione są algorytmy selekcji cech z ich strukturą kosztową bazujące na modelach statystycznych. Formalnie problem sformułowany jest w podpunkcie 7.1 przez wyrażenie (7.1). W tym przypadku problem polega na odpowiednim doborze parametrów modelu statystycznego. W podpunkcie 7.2 przedstawia się wersję algorytmu Lasso uwzględniającą koszty pozyskiwania cech. Adaptacyjną wersję tego algorytmu opisuje się w podpunkcie 7.3. W podpunkcie 7.4 przedstawia się algorytmy bazujące na różnych

funkcjach kary uwzględniające dodatkowo koszty pomiaru cech. W podpunkcie 7.5 omawia się wyniki eksperymentów obliczeniowych.

Rozdział 8 stanowią wnioski. Wymienia się w nim najważniejsze osiągnięcia pracy, wspomina się o znaczeniu praktycznym badania problematyki syntezy klasyfikatorów w sytuacji konieczności uwzględniania kosztów pozyskiwania cech. Wreszcie, omawia się występujące przy konstrukcji algorytmów ograniczenia oraz charakteryzuje się możliwe kierunki dalszych badań.

Ocena rozprawy

Recenzowana rozprawa jest oryginalnym dziełem naukowym. Jest powiązana bardzo ściśle z pięcioma oryginalnymi, współautorskimi publikacjami naukowymi Doktoranta (w czterech Doktorant jest pierwszym autorem).

Praca jest bardzo dobrze skomponowana i zaprojektowana tematycznie. Jest napisana jasnym i ładnym językiem. Zawiera jasno sformułowane problemy oraz ich rozwiązania, a także dobrze i precyzyjnie wprowadzone formalizmy matematyczne, które są podstawą konstrukcji algorytmów proponowanych w pracy. Jest starannie opracowana edycyjnie, zawiera przejrzyste i estetycznie opracowane rysunki i tabele.

Warsztat naukowy Doktoranta jest bardzo wartościowy. Doktorant potrafi dopasować metody i modele matematyczne do stawianych problemów oraz typu danych. Ma także kompetencje w zakresie budowania nowatorskim modeli statystycznych oraz ich weryfikowania. Praca dokumentuje także umiejętność Doktoranta posługiwania się warsztatem programistycznym i obliczeniowym obejmującym opracowanie i implementację algorytmów optymalizacji numerycznej. Opracowanie stabilnych implementacji modeli, ich przetestowanie i udostępnienie dowodzi dysponowania przez Doktoranta dobrym warsztatem badawczym w zakresie technik informatycznych.

Najważniejsze oryginalne osiągnięcia pracy obejmują

- Bardzo kompetentne i precyzyjne podsumowanie stanu wiedzy dotyczącej zastosowania wskaźników teorii informacji do selekcji cech, wyprowadzenie szeregu oszacowań przydatnych w konstrukcji algorytmów selekcji cech.

- Opracowanie nowego algorytmu selekcji cech z uwzględnieniem ich kosztów, bazującego na zachłannej procedurze optymalizacji numerycznej. Wykorzystanie w tym algorytmie oszacowanej dolnej granicy wartości wskaźnika informacji wzajemnej.

- Opracowanie teorii selekcji cech dla problemu, w którym struktura kosztów pozyskiwania cech ma charakter blokowy / grupowy. Opracowanie kilku (trzech) wersji odpowiedniego algorytmu, przetestowanie ich działania.

- Dokonanie weryfikacji i wszechstronnych porównań opracowanych algorytmów z algorytmami literaturowymi.

Wszystkie powyższe uwagi skłaniają do bardzo pozytywnej oceny recenzowanej pracy.

Uwagi dyskusyjne i krytyczne

W pracy brak jest bezpośrednio sformułowanych tez. Praca ma bardzo konsekwentną i logiczną strukturę, która bardzo dobrze pasowałaby do wspierania odpowiednio sformułowanych tez doktoratu. Mimo, że dużo uwagi poświęca się na opisanie oryginalnych elementów pracy i wkładu wniesionego przez pracę w rozwój badań nad klasyfikacją nadzorowaną, to jednak brak tez jest w mojej opinii wadą pracy.

Silną motywacją do prowadzenia badań nad klasyfikacją z uwzględnieniem kosztów pomiaru cech są aspekty praktyczne. Jest to jak najbardziej słuszne, jednak przy czytaniu pracy powoduje pewien niedosyt. Doktorant stwierdza, że raczej brakuje dostępnych danych rzeczywistych związanych z kosztami cech. Praktyczne znaczenie kosztów cech powinno powodować dostępność takich danych. Być może powinno się przeprowadzić jakieś badania, które pozwoliłyby odkryć (oszacować) koszty, które są ponoszone przy pomiarach cech, ale nie są opisywane w wynikach eksperymentów.

Pytania

Przy czytaniu opisów numerycznych, zachłannych algorytmów selekcji cech z uwzględnianiem kosztów nasuwają się pytania odnośnie szczegółów ich działania. W jaki sposób inicjalizowano algorytmy? Jak wyglądał proces zbieżności? Czy obserwowano wyraźną wielomodalność rozwiązań? Jakie były warunki stopu?

W tabeli 5.3 widać, że proponowany algorytm traci (może tracić) swoją przewagę przy zwiększaniu budżetu. Może to sugerować niewystarczająco intensywne przeszukiwanie przestrzeni rozwiązań. Czy próbowano, lub czy zdaniem Doktoranta ma sens próbować zastosować metody takie jak iteracje MCMC lub np. algorytmy genetyczne / memetyczne?

Konkluzja

Praca stanowi podsumowanie oryginalnych badań naukowych, w których sformułowano i weryfikowano oryginalne hipotezy badawcze. Dokumentuje wiedzę i kompetencje Doktoranta.

Osiągnięcia i oryginalne elementy pracy są na pewno wystarczające do jej ogólnej bardzo dobrej oceny. Uwagi dyskusyjne i krytyczne nie pomniejszają jej wartości. Stwierdzam, że rozprawa spełnia warunki stawiane pracom doktorskim i wnioskuję o jej dopuszczenie do publicznej obrony.

