POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki

Streszczenie rozprawy doktorskiej

ROZWÓJ METOD WYBORU ZMIENNYCH OPARTYCH O TEORIĘ INFORMACJI

DEVELOPMENT OF METHODS FOR FEATURE SELECTION BASED ON INFORMATION THEORY

mgr Radosław Piliszek

Promotor
dr hab. Witold R. Rudnicki

POZNAŃ 2023

Motywacja

Niniejsza rozprawa doktorska dotyczy obszernego tematu selekcji cech w uczeniu maszynowym. Mimo tej obszerności, co roku prezentowane są nowe metody selekcji cech, ulepszone w stosunku do poprzednich. Ponadto, coraz więcej kierunków ścisłych, humanistycznych i medycznych eksploruje możliwości związane z uczeniem maszynowym, którego jednym z etapów bardzo często jest selekcja cech.

W ostatnim czasie zwrócono szczególną uwagę, zwłaszcza w badaniach biomedycznych, na stabilność wyników. Powód jest dwojaki. Po pierwsze, etap wyboru cech jest tylko pierwszym krokiem w kierunku pełnego zrozumienia zjawiska, które to wymaga dalszej pracy ludzkiej, często kosztownej. Zatem ryzyko „niewłaściwych wskazówek" jest znaczące. Po drugie, ponieważ odkrycia mają charakter biomedyczny, mogą budzić obawy o odpowiedzialność przy ich wykorzystaniu. Jeśli występuje niewielka stabilność wyników, oznacza to, że istnieje duże ryzyko wyciągnięcia błędnych wniosków. Wykazano również, że wybór cech przyczynia się do odchylenia wyniku klasyfikacji bardziej niż samo dostrajanie parametrów modelu.

Rozważając dane zjawisko, często interesuje nas jego podłoże – w przypadku selekcji cech oznacza to wybranie cech najważniejszych, czyli takich, za pomocą których można jak najdokładniej opisać stany tegoż zjawiska. Wymaga to minimalno-optymalnego[1] podejścia do wyboru cech, który ma za zadanie odnaleźć najmniejszy podzbiór opisujący wynik „wystarczająco dobrze". Udowodniono, że jest to w ogólności zadanie NP-trudne, a zatem wszystkie popularne podejścia są z natury zachłanne – różnią się „sprytnością" swoich zachłannych heurystyk.

Motywacja pracy została omówiona bardziej szczegółowo w pierwszym rozdziale pracy i uzupełniona w rozdziale drugim.

Cele

Celem tej pracy było, przede wszystkim, zaproponowanie nowatorskiej, minimalno-optymalnej metody selekcji cech w oparciu o istniejące rozwiązanie na przecięciu selekcji cech i teorii informacji. W tym celu została zaproponowana odpowiednia miara odmienności cech, a także protokół walidacji do porównania z innymi ugruntowanymi podejściami do selekcji cech. Wszystkie cele zostały osiągnięte.

Przegląd stanu nauki

Temat pracy leży na przecięciu wielu dziedzin aktywnych badań, a mianowicie uczenia maszynowego, teorii informacji, analizy danych, statystyki i bioinformatyki. Drugi rozdział pracy zawiera kompleksowy przegląd zagadnień związanych z pracą. Rozpoczyna się krótkim przeglądem dziedzin uczenia maszynowego. Następnie następuje pogłębiony opis problemu redukcji wymiarowości w ogólności i selekcji cech, w szczególności w klasyfikacji

[1] W literaturze polskiej nie odnajduje się tej taksonomii metod selekcji cech, więc zastosowałem kalkę językową. Chodzi oczywiście o minimal-optimal.

z binarną decyzją. Omówione zostały różne podejścia przy użyciu odmiennych perspektyw, w tym porównanie podejść wyszukiwana wszystkich zmiennych istotnych[2] i minimalno-optymalnego zbioru cech – wraz z opisem filtrów, metod opakowujących[3] i metod osadzonych[4]. Bardziej szczegółowo opisany został również zestaw najważniejszych algorytmów selekcji cech. Następnie omówiłem algorytmy klastrowania, z pogłębionym opisem wybranych hierarchicznych algorytmów klastrowania. Potem zawarłem krótkie wprowadzenie do powiązanych zagadnień z zakresu teorii informacji. Rozdział obejmuje również kilka zagadnień technicznych wykorzystanych w pracy: problem wielokrotnego testowania, miary stabilności klastrów i miary wydajności klasyfikatorów.

Odmienność cech ograniczona do zmiennej decyzyjnej

Po zbudowaniu fundamentów w rozdziale drugim, rozdział trzeci zawiera opis pierwszej kontrybucji do stanu nauki. Standardowe miary podobieństwa i odmienności cech na ogół nie uwzględniają relacji między cechami a zmienną decyzyjną. Zaproponowana została nowa miara odmienności nazwana symetrycznym wzrostem informacji o celu[5] (STIG). Mierzy ona przyrost informacji o zmiennej decyzyjnej w dwóch przypadkach: gdy znamy wartość tylko jednej cechy oraz gdy znamy wartości obu cech. Na prostym przykładzie pokazano, że użycie tej miary może prowadzić do innego (i bardziej pożądanego) wyniku w porównaniu ze standardową miarą odmienności, w postaci zmienności informacji, gdy obydwa zastosowane zostaną do systemu, w którym cechy opisowe są skorelowane niezależnie od zmiennej decyzyjnej. W rozdziale opisanych zostało również kilka bardziej technicznych aspektów implementacji miary STIG.

Ulepszenia MDFS

W następnym rozdziale opisane zostały różne ulepszenia w bibliotece MDFS do wyboru cech. MDFS jest używany jako silnik obliczeniowy dla algorytmów proponowanych w pracy. W szczególności, do biblioteki wprowadzona została możliwość liczenia różnych miar teorii informacji, wraz z ulepszonym schematem dyskretyzacji i poprawioną wydajnością.

Odporna na przeuczenie i stabilna metoda wyboru cech

Rozdział piąty pracy poświęcony jest wprowadzeniu nowego algorytmu minimalno-optymalnej selekcji cech o nazwie Robust Agglomerative Feature Selection (RAFS). RAFS opiera się na trzech kluczowych ideach. Po pierwsze, wykorzystuje grupowanie cech i wyłanianie przedstawicieli klastrów (w przeciwieństwie do optymalizacji zestawów cech) w celu zmniejszenia liczby cech wykorzystanych w klasyfikacji binarnej. Po drugie, ekstensywnie wykorzystuje zespół wyników uzyskanych w schemacie z walidacją krzyżową w celu uzyskania stabilnego zestawu cech. Po trzecie, wykorzystuje omówioną wcześniej miarę STIG.

[2] all-relevant
[3] wrappers
[4] embedded
[5] symmetric target information gain

Zasadniczo, RAFS może używać dowolnej miary odmienności między cechami, ale domyślna konfiguracja wykorzystuje STIG wraz ze zmiennością informacji i miarą opartą na korelacji. Algorytm RAFS został szczegółowo opisany, a ponadto przedstawiłem przykłady zastosowań dla syntetycznych zbiorów danych – w formie samouczka.

Zastosowania w świecie rzeczywistym

Ostatni główny rozdział pracy opisuje zastosowanie metody RAFS do trzech rzeczywistych zestawów danych, obejmujących pełny zakres trudności klasyfikacji binarnej: średni sygnał (BLCA), silny sygnał (PTLD) i słaby sygnał (KIRC). Wszystkie zbiory danych pochodzą z badań omicznych na pacjentach z różnymi typami i podtypami raka. Pierwsze dwa to zbiory danych dotyczących ekspresji genów.

Zbiór BLCA opisuje pacjentów z rakiem pęcherza moczowego. U niektórych z nich rozwinął się cięższy przypadek raka in situ[6] (CIS). Ten zestaw danych jest oparty na sekwencji RNA. Zastosowanie RAFS do zbioru danych BLCA skutkuje zestawami cech, które zarówno dają lepsze wyniki klasyfikacji, jak i są bardziej stabilne w rygorystycznej zewnętrznej kontroli krzyżowej.

Zbiór PTLD opisuje pacjentów z potransplantacyjnym zaburzeniem limfoproliferacyjnym z rozlanym chłoniakiem z dużych komórek B (DLBCL) – najczęstszym podtypem PTLD. W danych występują dwie grupy pacjentów: ci z wirusem Epsteina-Barra i ci bez niego. Ten zestaw danych jest oparty na analizie mikromacierzy. RAFS ze STIG ujawniają więcej ciekawych genów niż poprzednie badania.

Zbiór KIRC opisuje pacjentów z rakiem nerki, których podzielono na grupy wysokiego i niskiego ryzyka. W przeciwieństwie do dwóch poprzednich, ten zestaw danych jest genomowy i obejmuje zmienność liczby kopii[7] (CNV). Pomimo słabego sygnału, proponowane podejście jest w stanie osiągnąć najwyższą obserwowalną wydajność klasyfikacji i stabilność wyboru cech w rygorystycznej zewnętrznej kontroli krzyżowej.

[6] carcinoma in situ
[7] copy number variation