Prof. dr hab. inż. Olgierd Hryniewicz
Instytut Badań Systemowych PAN

Warszawa, 2017-05-15

Recenzja

rozprawy doktorskiej mgr. Andrzeja Wójtowicza nt „Ensemble classification of incomplete data – a non-imputation approach with an application in ovarian tumour diagnosis support"

(„Grupowa klasyfikacja danych niekompletnych – podejście nieimputacyjne z zastosowaniem we wspomaganiu diagnostyki guzów jajnika")

1. Ogólna charakterystyka rozprawy

Niniejsza opinia została wykonana na prośbę prof. dr. hab. Jerzego Kaczorowskiego, Dziekana Wydziału Matematyki i Informatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu, z dnia 27-03-2017. Recenzowana rozprawa, napisana przez mgr. Andrzeja Wójtowicza, jest pracą doktorską na stopień doktora nauk matematycznych w zakresie informatyki i została wykonana pod kierunkiem prof. dr. hab. Macieja Wygralaka oraz dr. Krzysztofa Dyczkowskiego, jako promotora pomocniczego. Rozprawa ma charakter teoretyczno-praktyczny i jest poświęcona problemowi klasyfikacji obiektów opisanych zbiorem atrybutów, których wartości w pewnych przypadkach mogą nie być znane osobie dokonującej klasyfikacji. Równoległym i równorzędnym celem Doktoranta było wykorzystanie uzyskanych wyników do rozwiązania konkretnego zadania praktycznego, jakim jest diagnostyka guzów jajnika.

Zagadnieniu klasyfikacji, zwanemu także w środowisku informatyków jako zagadnienie uczenia z nadzorem, poświęcono setki, jeśli nie tysiące, prac naukowych. Opracowano dziesiątki występujących w różnych wariantach algorytmów, które są efektywnie wykorzystywane do klasyfikacji obiektów, których cechy (atrybuty) mogą mieć różną postać i być wzajemnie powiązane w skomplikowany sposób.

Największe problemy praktyczne, spotykane w przypadkach rzeczywistych implementacji algorytmów klasyfikacji, występują z powodu braków w danych. W przypadku obiektów opisywanych wieloma atrybutami powszechnie się zdarza, że wartości niektórych atrybutów nie są znane, przy czym zjawisko to może występować zarówno w zbiorze uczącym, wykorzystywanym do budowy klasyfikatora, jak też w przypadku obiektów poddawanych procesowi klasyfikacji. Budowa klasyfikatora jest stosunkowo prostsza, gdy w zbiorze uczącym mamy wystarczająco wiele przypadków, dla których znane są wartości wszystkich atrybutów. Możemy wówczas zbudować klasyfikator na podstawie części posiadanej informacji, a także – co jest rzadziej spotykane w praktyce – zestaw klasyfikatorów, wykorzystujących różne podzbiory atrybutów. To drugie podejście ma swój sens, gdy możemy z dużym prawdopodobieństwem wskazać atrybuty, których wartości mogą nie być znane. Powyższe podejście ma raczej ograniczone zastosowanie w przypadku zastosowań algorytmów klasyfikacji do diagnostyki medycznej, gdzie zbiory uczące są zazwyczaj niezbyt liczne, a także niepełne. Stosuje się, wobec tego, inne podejście, polegające na uzupełnianiu brakujących danych ich oczekiwanymi wartościami. Proces ten w środowisku osób zajmujących się analizą danych zwany jest procesem imputacji danych, a związanym z tym zagadnieniem problemom poświęcono wiele prac naukowych.

O ile problem imputacji danych może być względnie prosty na etapie budowy klasyfikatora, o tyle staje się bardzo trudny w rzeczywistym procesie klasyfikacji nowych przypadków. W takim przypadku użytkownikowi systemu klasyfikacji, np. lekarzowi w szpitalu, należy dostarczyć narzędzi informatycznych, które pozwoliłyby klasyfikować przypadki, dla których nie posiada on pełnych informacji o wartościach atrybutów modelu klasyfikacyjnego. W związku z tym, zaproponowane zostały tzw. metody nieimputacyjne, pozwalające w pewien sposób „ominąć" problem braku danych. Należy jednak podkreślić, że są to zazwyczaj metody mniej efektywne, ale za to znacznie prostsze w implementacji. Opiniowana rozprawa doktorska mgr. Andrzeja Wójtowicz dotyczy tego właśnie nurtu badań nad problemami klasyfikacji. Celem rozprawy jest opracowanie algorytmów, które atrybutom o nieznanych wartościach przypisują wartości przedziałowe, a także algorytmów, które wykorzystywane są w procesie podejmowania decyzji na podstawie takich niepełnych danych. Drugim celem rozprawy jest weryfikacja zaproponowanych rozwiązań na rzeczywistym przypadku diagnostyki medycznej.

2. Zawartość rozprawy

Recenzowana praca liczy w sumie 122 stron i składa się z wprowadzenia, rozdziału wprowadzającego podstawowe używane w rozprawie pojęcia, rozdziału zawierającego oryginalne wyniki pracy, dwu rozdziałów dotyczących praktycznej weryfikacji zaproponowanych rozwiązań, czterech załączników, w których opisano stosowane techniki obliczeniowe oraz wyniki eksperymentów. Rozprawę kończy krótkie (1 strona) podsumowanie. Ponadto rozprawa zawiera spis treści, listę używanych symboli, wykaz opracowanych algorytmów, wykaz rysunków, wykaz tabel oraz zawierający 76 pozycji wykaz cytowanej literatury.

We Wprowadzeniu (Introduction) Doktorant przedstawia medyczne motywacje podjęcia się prac na rozwiązanie przedstawionego w rozprawie problemu, w bardzo skrótowy sposób (w praktyce prawie wyłącznie odnośniki do literatury) przedstawia umiejscowienie swojej pracy w prowadzonych na świecie badaniach, definiuje jej cel oraz opisuje jej strukturę. W drugim rozdziale pracy (Basic definitions) mgr Andrzej Wójtowicz podaje wykorzystywane w rozprawie definicje, które przytacza na podstawie wskazanych pozycji literatury światowej z zakresu szeroko rozumianej eksploracji danych.

Trzeci rozdział rozprawy (Uncertaintification of classifiers) zawiera oryginalne teoretyczne wyniki rozprawy. Ich podstawą jest założenie, że w przypadku brakujących danych możemy dane te zastąpić przedziałami ich możliwych wartości. W tej sytuacji na wejściu stosowanych algorytmów (np. wykorzystujących pewne modele regresyjne lub drzewa klasyfikacyjne) mamy dane przedziałowe, zaś na ich wyjściu występują wartości mające także postać przedziałów. Przedziały te wskazują możliwe warianty klasyfikacji, uwzględniające niepewność związaną z występowaniem brakujących danych wejściowych. Jeżeli w procesie klasyfikacji wykorzystamy w podobny sposób szereg różnych klasyfikatorów, to uzyskujemy zbiór możliwych „niepewnych" klasyfikacji, który poddajemy procesowi agregacji. Polega ona na wykorzystaniu pewnej (zazwyczaj prostej) funkcji agregującej wraz z ustaloną wartością progową, przekroczenie której pozwala dokonać klasyfikacji w sposób jednoznaczny.

Kolejne dwa rozdziały poświęcone są eksperymentalnej weryfikacji zaproponowanych rozwiązań. W rozdziale 4 przedstawiono zastosowanie zaproponowanego podejścia w zagadnieniu diagnostyki guza jajnika, na podstawie rzeczywistych danych uzyskanych z

Uniwersytetu Medycznego w Poznaniu (prace prowadzone były wspólnie z lekarzami, naukowcami z tej uczelni). Z kolei, w rozdziale 5 przedstawiono wykorzystanie zaproponowanej metody do analizy 5 danych benchmarkowych, dostępnych w repozytorium danych z zakresu uczenia maszynowego prowadzonym na Uniwersytecie Kalifornijskim w Irvine. W obu przypadkach wykorzystano zestaw znanych klasyfikatorów, a celem badania był, miedzy innymi, wybór najlepszego operatora agregacji. Do oceny efektywności porównywanych rozwiązań wykorzystano zestaw powszechnie stosowanych wskaźników jakości klasyfikacji oraz odpowiednich testów statystycznych. Porównywane algorytmy były implementowane w powszechnie dostępnym języku programowania R.

3. Ocena rozprawy

Ocena merytoryczna

Rozprawa doktorska mgr. inż. Andrzeja Wójtowicza dotyczy ważnego obszaru komputerowej analizy danych, jakim jest zagadnienie klasyfikacji, a w szczególności zagadnienie budowy klasyfikatorów (uczenie pod nadzorem). Zaproponowane rozwiązanie można potraktować jako intuicyjnie proste połączenie metod stosowanych w różnych obszarach informatyki (klasyfikatory, operatory agregacji). Każda z zastosowanych metod ma swoje teoretyczne (matematyczne) uzasadnienie, ale próba formalnego uzasadnienia polegającej na ich połączeniu metody skazana jest chyba na niepowodzenie. Jest to sytuacja typowa w komputerowej analizie danych i wobec tego ograniczenie się przez Doktoranta wyłącznie do empirycznej weryfikacji zaproponowanych rozwiązań jest do zaakceptowania.

Pewne uwagi o charakterze polemicznym można mieć jednak do części weryfikacyjnej. W przypadku danych medycznych moje wątpliwości dotyczą analizy danych testowych. Dane te analizowano z wykorzystaniem algorytmów klasyfikacyjnych, które są stosowane w medycynie (odnośniki do czasopism medycznych) i służą do analizy danych pełnych. Z opisu eksperymentu nie wynika w sposób jednoznaczny sposób wykorzystania tych algorytmów do przypadków testowych z brakami danych. Czy w przypadku brakujących danych, które są używane przez dany klasyfikator, wynikiem klasyfikacji jest brak decyzji (NA), czy też klasyfikator jest modyfikowany, np. przez pominięcie w funkcji regresji odpowiedniej składowej, co może być przyczyną większej frakcji błędnych decyzji.

Z kolei, pewne wątpliwości budzi sposób porównania wyników działania porównywanego algorytmu z wynikami uzyskanymi przy zastosowaniu procesu imputacji. Jeśli dobrze zrozumiałem, to porównywano nowy algorytm agregacyjny z konkretnym algorytmem klasyfikacyjnym (*svmLinear*), w którym zaimplementowano algorytm imputacji (*mice*). Czy nie było możliwe porównanie algorytmu agregacyjnego z zespołem klasyfikatorów wykorzystujących mechanizm imputacji? Nie znalazłem też odpowiedzi na pytanie postawione w poprzednim akapicie. Co było robione, gdy analizowano dane testowe ze sztucznie wprowadzonymi brakami? Czy stosowano imputację, a jeśli tak, to w jaki sposób?

Ocena strony technicznej rozprawy.

Rozprawa dr. Andrzeja Wójtowicza napisana jest w sposób zwięzły, a czasami nawet zbyt „oszczędny". Na przykład, wykorzystane w rozdziale 5 klasyfikatory opisane są niewiele mówiącymi skrótami (nazwy funkcji w R?), które nie zostały nigdzie opisane. Podobnie, wykorzystane testy statystyczne (np. McNemara oraz t-Studenta z korekcją Benjaminiego-Hochberga) nie zostały opisane w pracy (np. nie podano jakie hipotezy statystyczne były

weryfikowane i jak należy interpretować wyniki tych testów), czyniąc jej odnośnie fragmenty niezrozumiałymi dla czytelników nie znających tych zastosowań statystyki.

Niezbyt czytelny jest opis wielu rysunków. Na przykład, z opisu Rys. 5.3 nie wynika, jak należy interpretować jedną krzywą opisaną jako „Original classifiers" lub „Uncertaintified classifiers". Podobny problem dotyczy innych rysunków. Czy krzywe te, lub słupki, są np. wynikiem uśrednienia wyników dla pojedynczych klasyfikatorów?

4. Ocena końcowa

Przedstawiona do recenzji rozprawa mgr. inżyniera Andrzeja Wójtowicza zawiera niewątpliwie propozycję oryginalnego rozwiązania z obszaru komputerowej analizy danych, co jest podstawowym wymogiem stawianym rozprawom doktorskim w zakresie informatyki. Praktyczna przydatność proponowanego rozwiązania jest – moim zdaniem – niewątpliwa. Wymienione w recenzji usterki redakcyjne utrudniają lekturę rozprawy, ale nie wpływają na jej pozytywną ocenę.

Z informacji jaką otrzymałem od Promotora rozprawy, prof. dr. hab. Macieja Wygralaka, wynika, że rozprawy doktorskie o charakterze teoretyczno-praktycznym, z mocnym akcentem na praktyczną przydatność otrzymanych rozwiązań, są przez Radę Wydziału Matematyki i Informatyki UAM akceptowane jako prace pozwalające nadać stopień doktora w dziedzinie „matematyka" i dyscyplinie „informatyka". W związku z tym uważam, że po **dokonaniu całościowej oceny rozprawy, po uwzględnieniu jej zalet i wad, można stwierdzić, że spełnia ona wymagania stawiane w odpowiednich przepisach rozprawom doktorskim w dziedzinie „matematyka" i dyscyplinie „informatyka". W związku z tym wnoszę o dopuszczenie mgr. Andrzeja Wójtowicza do dalszych, przewidzianych przepisami, etapów przewodu doktorskiego.**