dr hab. prof. UŚ Michał Baczyński
Instytut Matematyki
Wydział Matematyki, Fizyki i Chemii
Uniwersytetu Śląskiego w Katowicach
ul. Bankowa 14
40-007 Katowice

Katowice, 7 czerwca 2017 roku

Recenzja

pracy doktorskiej pana mgra Andrzeja Wójtowicza

„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")

Niniejsza recenzja została napisana na podstawie pisma prof. dra hab. Jerzego Kaczorowskiego, Dziekana Wydziału Matematyki i Informatyki UAM, z dnia 27 marca 2017 r., w związku z prowadzonym przez Radę Wydziału Matematyki i Informatyki UAM przewodem doktorskim mgra Andrzeja Wójtowicza.

Recenzowana rozprawa doktorska jest napisana w całości w języku angielskim oraz liczy łącznie 112 stron, z czego część główna stanowi 62 strony, a pozostała zawartość to 4 dodatki (26 stron), 4 listy (symboli, algorytmów, rysunków i tabel – 14 stron) oraz bibliografia, licząca w sumie 76 pozycji (10 stron). Część główna pracy składa się z 5 rozdziałów. Do pracy nie dołączono żadnego nośnika z oprogramowaniem, jednakże w bibliografii (poz. [45] oraz [68]) oraz w samej pracy (m.in. na str. 37, 54) podano adresy publicznych strony Internetowych udostępniających repozytorium kodów wykorzystywanych w aplikacyjnej części pracy oraz adres strony Internetowej projektu OvaExpert (zob. `http://ovaexpert.pl/`).

Zagadnienie klasyfikacji binarnej jest dobrze znane w literaturze i ma zastosowanie tam, gdzie jedna z klas ma szczególne znaczenie. Jednym z takich zagadnień jest diagnozowanie chorób w medycynie. Autor w swojej dysertacji zajmuje się problemem klasyfikacji binarnej danych niekompletnych. Jak pisze mgr Wójtowicz w Streszczeniu „Motywacja do podjęcia badań ma swoje źródło w medycynie, gdzie bardzo często występuje zjawisko braku danych".

Możemy wyróżnić następujące cele pracy (por. „Introduction", str. 3):

- zaproponowanie algorytmów dla klasyfikatorów operujących na danych niekompletnych w taki sposób, że zwracają one przedział możliwych predykacji,
- opisanie oryginalnych metod agregacji oraz metod progowych dla danych podanych w postaci przedziałowej,
- zbadanie zaproponowanych metod w problemie diagnozowania guzów jajnika,

- zbadanie zaproponowanych metod dla typowych danych wykorzystywanych w nauczaniu maszynowym.

Aby zrealizować swoje cele, autor podzielił swoją pracę na kilka części, które teraz dokładnie opiszę.

Rozdział 1 zatytułowany "Introduction" (str. 1–4) stanowi wprowadzenie do zagadnień omawianych w pracy. Autor w pierwszej kolejności koncentruje się na opisaniu problemu diagnozowania guzów jajnika, wskazując różnego rodzaju statystyki medyczne, w szczególności fakt, że w ostatnich latach liczba wykrywanych przypadków tej choroby jest zdecydowanie wyższa w takich krajach jak Chorwacja, Polska, Węgry, niż przeciętnie w całej Unii Europejskiej. Zbudowanie ogólnego systemu wspierającego podejmowanie decyzji przez lekarzy w konkretnym zagadnieniu medycznym jest celem wielu różnych zespołów naukowych. Autor podaje odnośniki do różnych takich metod znanych w literaturze przedmiotu. W większości te systemy dzielą się na dwie grupy: klasyfikatory liberalne (ang. liberal models) oraz klasyfikatory konserwatywne (ang. conservative models). Oprócz problemu poprawności klasyfikatora, autor zwraca uwagę na jeszcze jeden problem występujący często w badaniach medycznych: braku kompletnych danych. Wskazując konkretne przykłady z literatury, wskazuje potrzebę zbudowania modelu, który nie jest oparty na idei imputacji, czyli albo usunięciu danych niekompletnych albo sztucznym wstawieniu pewnych wartości do danych (np. średniej ze wszystkich wartości danej zmiennej w próbce). Zaznacza przy tym, że lekarze nadal lepiej diagnozują choroby niż systemy komputerowe.

Rozdział 2 zatytułowany „Basic definitions" (str. 5–15) jest krótkim wprowadzeniem teoretycznym do zagadnień stanowiących podstawę recenzowanej pracy i jest on całkowicie oparty na ogólnie znanej literaturze. Na początku (podrozdział 2.2) autor podał definicje klasycznych (binarnych) klasyfikatorów, w tym zostały omówione funkcje oceniające (ang. scoring functions), drzewa decyzyjne (ang. decision trees) oraz klasyfikatory grupowe/zespołowe (ang. ensemble classifiers). Następnie (podrozdział 2.3) podał wzory na różnego rodzaju miary stosowane w analizie klasyfikacji binarnej: trafność (ang. accuracy) wrażliwość/czułość (ang. sensitivity), specyficzność (ang. specificity) oraz decyzyjność (ang. decisiveness). W tej części autor wprowadził również bardzo ważne pojęcie macierzy kosztów. W podrozdziale 2.4 podał, w pseudokodzie, trzy algorytmy związane ze sprawdzianem krzyżowym (ang. cross-validation). Ostatni podrozdział w tej części pracy dotyczy problemu klasyfikacji binarnej w przypadku brakujących (niekompletnych) danych. Autor podaje tutaj odnośniki do literatury, które opisują różne podejścia oparte na metodzie imputacji.

Rozdział 3 zatytułowany „Interval classification procedure" (str. 17–29) stanowi najważniejszą część teoretyczną pracy doktorskiej mgra Wójtowicza i opisuje nowe podejście do zagadnienia klasyfikacji binarnej w oparciu o rachunek przedziałowy. Ta część jest oparta na artykule „Solving the problem of incomplete data in medical diagnosis via interval modeling" opublikowanym w roku 2016 w czasopiśmie Applied Soft Computing [32], którego mgr Wójtowicz jest współautorem (i zarazem pierwszym autorem). W pierwszej kolejności (podrozdział 3.1) autor wprowadza notację dla modelu przedziałowego rozszerzoną o wartość „NA" i pokazuje jak można przekształcić dane (liczbowe) na dane przedziałowe. W podrozdziale 3.2 pokazał w jaki sposób można rozszerzyć

opisane w poprzednim rozdziale klasyfikatory binarne na sytuację, gdy dane wejściowe są przedstawione w postaci przedziałów, czyli przedstawił ideę jak rozszerzyć funkcję $f\colon X \to \mathbb{R}$, gdzie $X = X_1 \times \cdots \times X_n$ jest dziedziną instancji oraz X_i jest (numeryczną) dziedziną atrybutu x_i, do funkcji postaci $\hat{f}\colon \hat{X} \to \mathcal{I}_{[0,1]}$, gdzie $\hat{X} = \hat{X}_1 \times \cdots \times \hat{X}_n$ oraz $\hat{X}_i = \mathcal{I}_{X_i} = \{[a,b] : [a,b] \subseteq X_i\}$. Podane definicje i wzory zostały w każdym przypadku zilustrowane stosownym przykładem. Kolejny podrozdział 3.3 koncentruje się na problemie agregacji danych. Po podaniu podstawowych definicji dla przypadku klasycznego, czyli rozważeniu funkcji postaci $Agg\colon [0,1]^n \to [0,1]$, autor koncentruje swoją uwagę na operatorach agregujących postaci $\widehat{Agg}\colon \mathcal{I}^n_{[0,1]} \to \mathcal{I}_{[0,1]}$. Funkcje tego typu były badane m.in. przez Beliakowa i innych (poz. [51], [53] w Bibliografii) oraz Deschrijvera i Kerre (poz. [52]). Mając na uwadze wcześniejsze rozważania oraz klasyfikatory postaci $\hat{f}_i\colon \hat{X} \to \mathcal{I}_{[0,1]}$, autor pokazuje w ważnym, z praktycznego punktu widzenia, przykładzie 3.5, jak można stosować różnego rodzaju operatory agregujące dla danych przedziałowych (zarówno numeryczne operatory agregujące jak i przedziałowe operatory agregujące). Należy tutaj podkreślić, że lista operatorów agregujących wykorzystywanych w dalszej części pracy jest zawarta w dodatku A „Aggregation operators" (str. 63–68). Mając otrzymaną (w wyniku działania operatorów agregujących) wartość liczbową lub przedział, w kolejnym podrozdziale 3.4 autor opisuje podejście progowe, które może mieć jedną z następujących postaci funkcyjnych:

$$\begin{aligned} &\tau\colon [0,1] \to \{y_1, y_2, NA\}, && \text{numeryczna selekcja progowa,} \\ &\hat{\tau}\colon \mathcal{I}_{[0,1]} \to \{y_1, y_2, NA\}, && \text{przedziałowa selekcja progowa.} \end{aligned}$$

Krótka lista funkcji takiej postaci wykorzystywanych później w zagadnieniach praktycznych jest zawarta w dodatku B „Thresholding strategies" (str. 69–70). Zwieńczeniem tego rozdziału jest podrozdział 3.5, gdzie autor podsumował opisane wcześniej idee, zarówno w postaci graficznej (rysunek 3.1) jak i w postaci algorytmu w pseudokodzie (algorytm 3.1). Zdaniem recenzenta, z teoretycznego punktu widzenia, ten algorytm stanowi główny rezultat pracy doktorskiej mgra Wójtowicza.

Rozdział 4 zatytułowany „Medical evaluation" (str. 31–48) przedstawia zastosowanie opisanej wcześniej metody klasyfikacji z danymi niepełnymi do problemu klasyfikacji guzów jajnika. Próbka zawierała dane 388 pacjentek Uniwersytetu Medycznego w Poznaniu, które poddane były badaniom w latach 2015-2016. Wśród posiadanych danych, 56% pacjentek zawierało dane kompletne, zaś 44% pacjentek posiadało dane niekompletne. Rozkład danych przedstawiony jest w pracy na rysunku 4.1. W eksperymencie komputerowym autor zastosował 6 modeli diagnostycznych znanych z literatury: dwa oparte na funkcji oceniającej, a cztery na modelach regresji. Przy kroku agregacji, cztery różnego rodzaju grupy operatorów agregujących były wykorzystane, przy czym za każdym razem były rozważane dwa scenariusze: agregacji poddawano całe przedziały (funkcje $\widehat{Agg}$) lub też ich numeryczne reprezentacje (funkcje Agg). Ostateczna klasyfikacja odbywała się na podstawie funkcji progowych opisanych w rozdziale 3.4 oraz dodatku B. Istotnym krokiem w całej procedurze było ustalenie współczynników w macierzy kosztów, które to została ustalone wraz z lekarzami tak jak podano w tabeli 4.2 na stronie 37. Cała procedura, pliki źródłowe oraz wyniki zostały zaimplementowane wykorzystując oprogramowanie R i są one dostępne na stronie Internetowej `https://github.com/ovaexpert/ovarian-tumor-aggregation`. W podrozdziale 4.6 autor przedstawił, w postaci tabel oraz wykresów, wyniki przeprowadzonych testów, w tym wartości miar opisanych w rozdziale 2.

Natomiast w podrozdziale 4.7 opisał działający w trybie rzeczywistym system OvaExpert, wspomagający diagnostykę guzów jajnika, gdzie zaproponowana metoda (przy ustalonych: funkcji agregującej, funkcji progowej, por. str. 46) została wykorzystana i wyniki zostały porównane z innymi metodami klasyfikacji. Głównym wnioskiem tego eksperymentu jest konkluzja, że przy zaproponowanym nowym podejściu można uzyskać lepszą wydajność w klasyfikacji guzów jajnika niż przy wykorzystaniu innych znanych metod opartych na imputacji.

W **rozdziale 5** zatytułowanym „Evaluation on UCI datasets" (str. 49–59) autor pokazał, jak przedstawiona w rozdziale 3 metoda sprawdza się w danych pochodzących z innych danych niż dane medyczne. Aby być obiektywnym w ocenie, mgr Wójtowicz zdecydował się wykorzystać 5 zbiorów danych dostępnych na repozytorium Uniwersytetu w Kalifornii, Irvine (UCI) (zob. [58]): „bank-marketing", „census-income", „credit-card", „magic" oraz „wine-quality". Autorska procedura z rozdzaiłu 3 została porównana z trzema metodami opartymi na imputacji. Opis przeprowadzonych badań jest zawarty w podrozdziale 5.3, gdzie podano również algorytmy wykorzystywane w eksperymencie. Analiza złożoności obliczeniowej jednego z tych algorytmów (a dokładniej algorytmu 5.2) jest zawarta w dodatku C „Algorithm complexity analysis". W omawianym rozdziale tak naprawdę autor opisał dokładniej tylko wyniki dla pierwszego zestawu danych, zaś zestawienia dla pozostałych czterech danych testowych zostały umieszczone w dodatku D „Results for UCI repository datasets". Przykładowo, na str. 55, zostały umieszczone uzyskane wartości dla poszczególnych miar zarówno w przypadku danych pełnych (tabela 5.2) jak i przy danych niepełnych i zastosowanych różnych strategii (tabela 5.3). W tym drugim przypadku ostatni wiersz zawiera dane związane ze strategią opartą na agregacji.

Moim zdaniem recenzowana praca nie poddaje się jednoznacznej ocenie, a decyzja którą muszę podjąć jest dość trudna. Z czysto teoretycznego punktu widzenia nie zawiera ona matematycznych twierdzeń popartych dowodami. Najważniejszym autorskim wynikiem teoretycznym jest ogólna metoda oraz algorytm podany pod koniec rozdziału 3 na stronach 28 oraz 29. Jednakże z punktu widzenia zastosowań, w szczególności zastosowań metod matematycznych w informatyce, praca jest interesująca, gdyż nie tylko dotyka bardzo ważnego problemu medycznego, ale również dlatego, że wyniki uzyskane w trakcie różnego rodzaju eksperymentów pokazują, że zaproponowane podejście może być stosowane w praktyce. Autor zaimplementował zarówno autorską metodę jak i inne strategie w języku R oraz dokonał analizy otrzymanych rezultatów. Z praktycznego punktu widzenia liczba przeprowadzonych eksperymentów jest bardzo duża i widać, że w takich też badaniach eksperymentalnych autor czuje się najlepiej. W swoich badaniach mgr Wójtowicz wykorzystał różnorodne biblioteki programistyczne – pokazuje to biegłość doktoranta w programowaniu.

Zasadniczo nie mam zastrzeżeń do redakcji rozprawy. Praca jest napisana poprawnie, układ rozdziałów jest czytelny. Dysertacja jest napisana w języku angielskim i nie znalazłem tutaj rażących błędów językowych lub składniowych. Jednakże mam pewne uwagi szczegółowe.

Uwagi szczegółowe:

1. str. 19, zdanie „These two definitions are equivalent whenever the scoring function is continuous". Sądzę, że autor mógł trochę szerzej skomentować ten fakt. Czy warunek ciągłości jest tutaj konieczny?

2. Podrozdział 3.3: uważam, że można było więcej miejsca poświęcić teorii operatorów agregujących działających na przedziałach. To było idealne miejsce na większy wkład czysto matematyczny, który uzupełniłby część praktyczną. Na przykład autor mógł pokazać jak przekładają się wybrane własności klasycznych operatorów agregujących na własności operatorów rozszerzonych. Taka część wzmocniłaby moją opinię, że praca jest napisana w dziedzinie nauk matematycznych – teraz jest ona na jej pograniczu.

3. str. 37: wybór stałych w tabeli 4.2 jest dość arbitralny. Czy autor przeprowadził badania jakie wyniki zostaną uzyskane, jeżeli zmienimy chociaż minimalnie tę macierz kosztów?

4. str. 66 oraz 67: Czy autor może wyjaśnić jak należy rozumieć podany operator agregujący dla różnych α?

5. str, 103, References: moim zdaniem przy takiej dużej liczbie pozycji w bibliografii, porządek alfabetyczny byłby o wiele wygodniejszy i czytelniejszy. Ponadto nie wiem dlaczego autor nie podaje wszystkich autorów danej pozycji w bibliografii.

6. str. 107, poz. [31]: edytorzy są błędnie podani.

Należy podkreślić, że praca doktorska oparta jest na dość imponującej liczbie artykułów naukowych (9 publikacji), których mgr Wójtowicz jest współautorem, z czego 3 artykuły [29], [32] oraz [35] to prace opublikowane w czasopismach naukowych, 5 artykułów [30], [31], [33], [34], [36] to prace zawarte w materiałach konferencyjnych, zaś praca [28] jest artykułem w monografii.

Podsumowując uważam, że rozprawa doktorska pana mgra Andrzeja Wójtowicza wpisuje się w nurt prac na pograniczu matematyki oraz informatyki. Cele postawione na początku dysertacji zostały zrealizowane. Najważniejszym wynikiem teoretycznym jest zaproponowanie algorytmu grupowej klasyfikacji danych niekompletnych w wykorzystaniem agregacji danych (przedziałowych). **W mojej opinii recenzowana praca doktorska spełnia wymagane ustawą warunki, to jest art. 13 ust. 1 *Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki***. Choć dysertacja nie jest na najwyższym poziomie teoretycznym (brak wyników czysto matematycznych jest jednak dla mnie istotnym mankamentem), stanowi ona oryginalne rozwiązanie problemu naukowego. Wykazuje, że autor potrafi prowadzić badania naukowe na pograniczu matematyki i informatyki oraz wykazuje ogólną wiedzę autora w zakresie eksploracji danych, w szczególności w klasyfikacji danych. Z punktu widzenia zastosowań, autor zaimplementował omawiane metody klasyfikacji danych w języku R. Moim zdaniem dysertacja stanowi podstawę do nadania autorowi stopnia naukowego doktora nauk matematycznych w zakresie informatyki. Dlatego wnoszę o jej przyjęcie i dopuszczenie pana mgra Andrzeja Wójtowicza do dalszych etapów przewodu doktorskiego, w szczególności do publicznej obrony rozprawy doktorskiej.

Michał Baczyński