**POLITECHNIKA WARSZAWSKA**

**WYDZIAŁ MATEMATYKI i NAUK INFORMACYJNYCH**

ul. Koszykowa 75
00-662 WARSZAWA
NIP: 525-000-58-34
Tel.: (22) 234-79-88, 621-93-12
Fax.: (22) 625-74-60

Warszawa, 9 kwietnia 2018 roku

Prof. dr hab. inż. Jacek Mańdziuk, prof. zw. PW
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

# Recenzja rozprawy doktorskiej mgr. Łukasza Zaniewicza zatytułowanej *Support Vector Machines for Uplift Modeling*

Niniejsza recenzja została przygotowana na prośbę Zastępcy Dyrektora ds. Naukowych Instytutu Podstaw Informatyki Polskiej Akademii Nauk prof. dr hab. inż. Wojciecha Penczka wyrażoną w piśmie z dnia 12 lutego 2018 roku.

## Tematyka rozprawy

Przedstawiona do recenzji rozprawa dotyczy zagadnienia modelowania różnicowego (ang. *uplift modeling*) stanowiącego istotny aspekt uczenia maszynowego (UM). Modelowanie różnicowe wykorzystywane jest w sytuacjach dotyczących wyboru ze zbioru podmiotów (np. pacjentów, abonentów jakiejś usługi czy klientów danej firmy) podzbioru tych osób, które powinny być poddane określonemu działaniu (np. terapii czy kampanii marketingowej). W takich sytuacjach zastosowanie typowych modeli UM ograniczone jest brakiem możliwości odniesienia uzyskanych wyników do sytuacji, w której działanie nie zostało wobec danego podmiotu zrealizowane. Przykładowo, jeżeli klient operatora telefonii komórkowej przedłuży umowę po zaproponowaniu mu korzystniejszej oferty wraz z końcem okresu obowiązywania umowy, nie mamy możliwości sprawdzenia czy przedłużenie umowy było w istocie wynikiem zaoferowania nowych warunków czy też abonent i tak przedłużyłby umowę na poprzednich warunkach, korzystniejszych dla operatora. W drugiej sytuacji koszt związany z przeprowadzeniem kampanii oraz poprawą oferty jest nieuzasadniony (nie uwzględniamy oczywiście w tym miejscu czynników długofalowych).

Jednym ze sposobów analizowania tego typu sytuacji jest wykorzystanie grupy kontrolnej czyli podmiotów, które nie są poddawane określonemu działaniu (w tym wypadku kampanii marketingowej) i oszacowanie różnicy pomiędzy skutecznością działania na grupie bazowej (poddanej działaniu) w stosunku do grupy kontrolnej. Stąd nazwa podejścia – *modelowanie różnicowe*.



wpłynęło
17 KWI. 2018

Praca doktorska mgr Łukasza Zieniewicza dotyczy adaptacji jednego z podstawowych modeli uczenia maszynowego – Maszyny Wektorów Wspierających (SVM) umożliwiającej jego zastosowanie w zagadnieniu modelowania różnicowego. Ponadto, Autor proponuje modyfikację funkcji straty pozwalającą na wpływanie przez operatora eksperymentu na wzajemne relacje pomiędzy licznościami zbiorów przykładów pozytywnych, negatywnych oraz neutralnych.

Rozważania dotyczące wykorzystania zaproponowanego przez Autora rozprawy modelu Uplift-SVM w modelowaniu różnicowym uzupełnia dyskusja na temat nielosowego przyporządkowania podmiotów do grup bazowej i kontrolnej. Wybór nielosowy grupy bazowej stanowi istotny problem, często występujący w praktyce, np. w obszarze medycznym, w którym wybór pacjentów podlegających terapii jest z natury rzeczy obciążony dotychczasowym przebiegiem choroby.

## Hipotezy badawcze

Podstawową hipotezą badawczą rozprawy jest potwierdzenie możliwości takiej modyfikacji modelu SVM, która pozwoliłaby na jego efektywne zastosowanie w problemie modelowania różnicowego.

Pozostałe hipotezy (cele) badawcze odnoszą się do specyficznych aspektów funkcjonalnych proponowanego modelu Uplift-SVM (USVM) i dotyczą: (1) możliwości zwiększania lub zmniejszania stosunku liczności grupy przypadków pozytywnych do negatywnych poprzez sterowanie ograniczeniami nałożonymi na grupę przypadków neutralnych oraz (2) możliwości wykorzystania modelu USVM w sytuacji, w której dobór przypadków do grup bazowej i kontrolnej nie jest w pełni losowy.

## Treść rozprawy

Rozprawa napisana jest w języku angielskim, liczy 88 stron, zawiera streszczenie w językach polskim i angielskim, składa się z 7 rozdziałów oraz spisu literatury zawierającego 50 pozycji. W rozdziale wprowadzającym Doktorant przedstawia zagadnienie modelowania różnicowego oraz wspomniane wyżej cele rozprawy w kontekście aktualnego stanu literatury. W dalszej części rozdziału zreasumowane są podstawowe wyniki badawcze rozprawy oraz przedstawiona jest struktura dysertacji.

Nie mam uwag do tej części pracy. Wprowadzenie do tematyki przedstawione jest w sposób nie budzący merytorycznych wątpliwości, podobnie opis celów badawczych i streszczenie najważniejszych wyników. Kontekst literaturowy mógłby być rozszerzony o ogólniejszą charakterystykę zagadnień uczenia maszynowego, niemniej z punktu widzenia istoty rozprawy, czyli tematyki modelowania różnicowego, zacytowane prace wybrane są prawidłowo a ich liczba jest wystarczająca.

Na podkreślenie zasługuje dojrzałe, uporządkowanie ujęcie omawianego wprowadzenia do tematyki modelowania różnicowego. Widać tutaj dobry wpływ Promotora rozprawy – prof. Szymona Jaroszewicza, uznanego specjalisty w dziedzinie uczenia maszynowego, w szczególności modelowania różnicowego.

Rozdział drugi zawiera formalne wprowadzenie do tematyki wykorzystania modelu SVM w zagadnieniu klasyfikacji. Zaczynając od modelu bazowego wykorzystywanego w przypadku zbiorów liniowo separowalnych, Autor kolejno przedstawia bardziej złożone wersje SVM, tzn. model ze zmiennymi osłabiającymi (*slack variables*) oraz model wykorzystujący tzw. trik kernelowy (*kernel trick*).

Omawiany rozdział napisany jest prawidłowo i z dużym zrozumieniem intuicji poszczególnych modeli oraz prawidłowym uzasadnieniem formalnym wprowadzanych modyfikacji względem modelu bazowego. Materiał zamieszczony w rozdziale nie zawiera wyników autorskich przypominając, generalnie dobrze znane, własności modelu SVM. Rozdział kończy się stwierdzeniem odnoszącym się do autorskiego modelu USVM w kontekście triku kernelowego: „*Moreover, experiments we have performed did not show significant improvements from using nonlinear kernels on benchmark datasets available to us*". Jestem ciekaw zdania Autora odnośnie przyczyn braku poprawy oraz możliwych kierunków rozwoju modelu USVM dających szansę na poprawę wyników w przypadku stosowania triku kernelowego.

Rozdział trzeci przedstawia autorski model USVM stanowiący główne osiągnięcie pracy doktorskiej, rozszerzający zakres stosowania modelu SVM do przypadku modelowania różnicowego, które obejmuje dwa zbiory obiektów: grupę bazową oraz grupę kontrolną. W dużym skrócie modyfikacja polega na uwzględnieniu w procesie klasyfikacji dwóch (w miejsce jednej) hiperpłaszczyzn rozdzielających zbudowanych w oparciu o dwa rozważane zbiory przypadków.

W dalszej części rozdziału Doktorant definiuje formalnie, w oparciu o warunki Karusha-Kuhna-Tuckera postać zadania optymalizacyjnego rozwiązywanego przez model USVM oraz dowodzi podstawowych własności teoretycznych zaproponowanego modelu odnoszących się do wpływu stosunku współczynników $C_2$ / $C_1$ w rozważanym zadaniu optymalizacyjnym na wynikowy podział przestrzeni na przypadki pozytywne, neutralne oraz negatywne.

Ostatnia sekcja w tym rozdziale (3.4) omawia uogólnienie USVM do postaci $L_P$-USVM (jako analogonu uogólnienia SVM przez $L_P$-SVM), w którym zmienne osłabiające podniesione są do potęgi p (modelom standardowym w obu przypadkach odpowiada p=1). W dalszych trzech punktach (3.4.1 – 3.4.3) Autor przeprowadza w sposób formalny analizę własności modelu $L_P$-SVM. Dowody przedstawione przez Doktoranta są dokonaniami oryginalnymi (dotyczą zaproponowanego przez Niego modelu), niemniej dla jasności warto podkreślić, że są one przeprowadzane w sposób bardzo podobny do przypadku modelu SVM. Autor zwraca uwagę na to podobieństwo cytując wielokrotnie prace [1], [8] oraz [32].

Nie mam uwag do omówionego wyżej rozdziału. Propozycja modelu USVM (oraz $L_P$-USVM) stanowi bez wątpienia ważne osiągnięcie badawcze pozwalające na istotne rozszerzenie stosowalności odpowiednich wersji modelu SVM. Poziom opisu formalnego jest wysoki i nie znalazłem w nim nieścisłości.

Kolejne trzy rozdziały dotyczą praktycznych aspektów związanych z implementacją oraz eksperymentalna oceną skuteczności zaproponowanego modelu USVM. W szczególności rozdział 4 prezentuje metodę rozwiązywania zadania optymalizacyjnego realizowanego przez model USVM z wykorzystaniem solverów z biblioteki CVXOPT. W celu optymalizacji czasowej oraz zwiększenia stabilności numerycznej proponowanego podejścia Doktorant zaimplementował własne solvery dla warunków Karusha-Kuhna-Tuckera, dla zadania pierwotnego oraz dualnego. Techniki algebraiczne wykorzystywane w celu odpowiedniej reprezentacji zadania (uzupełnienie Schura i wzór Shermana–Morrisona zwany także wzorem Shermana–Morrisona-Woodbury'ego) omówione są w punkcie 4.1, a ich wykorzystanie w kontekście USVM w sekcjach 4.2 i 4.3 odpowiednio dla modeli $L_1$-USVM oraz $L_P$-USVM.

Przekształcenia matematyczne przedstawione w rozdziale 4 są formalnie poprawne i oryginalne w zakresie, w którym odnoszą się do autorskiego podejścia USVM. Podobnie jak w rozdziale 3 w dużej mierze bazują one na podobnych rozważaniach prezentowanych wcześniej w literaturze, w szczególności pracach [5], [11] oraz [16], które Doktorant prawidłowo cytuje we właściwych kontekstach.

W rozdziale piątym Doktorant odnosi się do sytuacji, w której podmioty przyporządkowywane są do grup bazowej i kontrolnej w sposób niecałkowicie losowy. Sytuacja taka, jak wspomnieliśmy powyżej występuje w praktyce bardzo często – stąd potrzeba mitygacji wpływu tego zjawiska na wynik klasyfikacji. W tym celu Autor proponuje modyfikację modelu USVM poprzez wprowadzenie czynnika regularyzacyjnego nazwanego przez Doktoranta regularyzają Székely'ego. Pomysł odnosi się do tzw. *odległości energetycznej* pomiędzy rozkładami prawdopodobieństwa próbek zaliczonych odpowiednio do grupy bazowej oraz grupy kontrolnej. Odległość energetyczna została zaproponowana w literaturze przez Székely'ego i Rizzo. Jej wartość zeruje się wtedy i tylko wtedy gdy rozkłady prawdopodobieństwa obu próbek są identyczne. Zakładając w pełni losowy rozkład w grupie kontrolnej, czynnik regularyzacyjny stanowi karę za odstępstwo od tego rozkładu w grupie bazowej.

W dalszej części rozdziału Kandydat definiuje w sposób formalny wersję modelu USVM z regularyzacją, dowodzi jego podstawowych własności oraz formułuje problem optymalizacyjny rozwiązywany przez wprowadzoną wersję modelu oraz przedstawia pseudokod stosownego algorytmu optymalizacyjnego.

Jednym z kluczowych parametrów proponowanego modelu jest współczynnik $\alpha$ w równaniu regularyzacji Székely'ego (5.2.3). Zgodnie z wymogami problemu jego wartość powinna należeć do przedziału [1,2). Autor dość arbitralnie przyjął $\alpha$=1.1 nie wskazując metody wyboru tej wartości parametru. W kontekście szerszym, interesująca byłaby dyskusja dotycząca zależności pomiędzy strukturą zbioru danych a doborem wartości $\alpha$. Nie oczekuję w tym miejscu konkretnych, formalnych wskazań na gruncie teoretycznym, chodzi raczej o głębsze odniesienie do kwestii skutecznego wyboru $\alpha$ dla zadanego problemu (czyli dystrybucji próbek w obu grupach).

Rozdział szósty poświęcony jest eksperymentalnej ocenie efektywności modelu USVM. Doktorant przedstawia sposób modelowania krzywej wzrostu w przypadku problemu klasyfikacji różnicowej polegający na odjęciu od siebie krzywych wzrostu dla zbiorów próbek kontrolnych i próbek badanych oraz wskazuje zbiory testowe wykorzystane do weryfikacji skuteczności modeli

różnicowych. Z uwagi małą liczbę reprezentatywnych zbiorów danych tworzonych z myślą o modelowaniu różnicowym, Autor definiuje także 16 dodatkowych zbiorów testowych powstałych w drodze sztucznego podziału na dane bazowe i kontrolne wybranych zbiorów UCI, zgodnie z metodą zaproponowaną w jednej z prac Promotora rozprawy [38].

Sekcja 6.3 ilustruje różnice pomiędzy modelami $L_1$-USVM a $L_P$-USVM (dla $p \neq 1$) w oparciu o zbiory *breast-cancer* oraz *australian* – oba pochodzące z UCI ML. Podstawowy wniosek płynący z wykonanych eksperymentów dotyczy możliwości sterowania względną licznością grupy próbek neutralnych poprzez odpowiedni dobór relacji $C_2/C_1$. W ten sposób eksperymentator może wpływać na liczności pozostałych grup (pozytywnej i negatywnej) w stosunku do rozmiaru grupy neutralnej. Niewątpliwie, taka możliwość stanowi silne narzędzie w rękach osoby prowadzącej eksperyment, pojawia się jednak w tym kontekście pytanie o *subiektywizm* uzyskanych w ten sposób wyników – wystarczy porównać wykresy na Rys. 6.3.1 uzyskane dla $p = 1.2$ oraz $p = 2.0$. Czy Doktorant rozważał możliwość i potrzebę jakiejś formy *zobiektywizowania* doboru $C_2/C_1$, np. poprzez nałożenie określonych warunków brzegowych na grupę wyników neutralnych?

W sekcji 6.4, w oparciu o 28 zbiorów testowych, Autor porównuje wyniki uzyskane przez model USVM z rezultatami pięciu innych modeli różnicowych wymienionych i krótko scharakteryzowanych w tej sekcji. Z punktu widzenia praktycznych rezultatów pracy doktorskiej rozdział ten ma istotne znaczenie, umożliwiając ocenę proponowanego modelu na tle stanu wiedzy w dziedzinie klasyfikacji różnicowej w oparciu o modele typu SVM. Porównanie wypada dla USVM pozytywnie, ale nie entuzjastycznie. Wyniki są porównywalne do uzyskanych przez trzy modele referencyjne (Double SVM, Diff-Pred SVM, Uplift Tree – ten ostatni jako jedyny nie należy do rodziny modeli SVM). Szczególnie istotny jest „remis" z modelem Double SVM stanowiącym prostą i bardzo naturalna implementację idei modelowania różnicowego. Pozostałe dwa modele referencyjne są wyraźnie słabsze, w tym jeden z nich (Treatment SVM) z oczywistych powodów, ponieważ nie wyróżnia grupy kontrolnej. Narzucającym się wnioskiem z analizy wyników jest *komplementarność podejścia USVM oraz Double SVM.* Czy Autor rozważał możliwość automatycznego doboru jednej z metod do zadanego problemu klasyfikacji różnicowej?

Punkt 6.5 zamykający omawiany rozdział weryfikuje skuteczność autorskiej metody wykorzystania modelu USVM w sytuacji, gdy rozkład w grupie bazowej nie jest w pełni losowy. Doktorant przedstawia proponowaną metodykę przeprowadzania eksperymentu, opisuje wykorzystane zbiory testowe, a następnie uzyskane wyniki i płynące z nich wnioski. Jedno ze sformułowań dotyczących opisu procedury testowania budzi moje wątpliwości, mianowicie zdanie „*Since different bias correction procedures are used for model construction and testing, we believe that it is less likely that the estimated model performance is a result of an uncorrected treatment assignment bias.*" Chętnie wysłuchałbym krótkiego wyjaśnienia dotyczącego istoty merytorycznej tego stwierdzenia. Drugą kwestią jest dobór współczynnika kary Székely'ego ($C_3$), który – jak pokazują testy – ma kluczowy wpływ na zachowanie modelu USVM. W jaki sposób (jaką metodą) powinna być dobierana wartość tego współczynnika?

Ostatni rozdział zawiera podsumowanie wyników przedstawionych w pracy oraz przypomina podstawowe wnioski płynące z przeprowadzonych badań. Wskazane byłoby jego uzupełnienie o dyskusję dotyczącą możliwych kierunków kontynuacji prac badawczych w obszarze poruszonych w rozprawie zagadnień.

Rozprawę dopełnia spis literatury obejmujący 50 pozycji, z których większość opublikowana została w okresie ostatnich 10 lat. Dobór pozycji bibliograficznych oraz sposób posługiwania się zawartymi w cytowanych pracach wynikami potwierdzają ugruntowaną wiedzę Autora w zakresie klasyfikacji z użyciem modeli z rodziny SVM, zarówno w warstwie teoretycznej jak i zastosowań praktycznych.

## Oryginalny wkład Autora rozprawy

Oryginalny wkład Autora w ramach rozważanego w rozprawie zagadnienia naukowego dotyczy generalnie trzech następujących obszarów:

1. Opracowania, implementacji, analizy teoretycznej oraz weryfikacji eksperymentalnej modelu klasyfikacji różnicowej USVM stanowiącego nietrywialna modyfikację modelu SVM;
2. Rozszerzenia oraz weryfikacji skuteczności powyższego modelu do rodziny klasyfikatorów $L_P$-USVM, w której model bazowy otrzymujemy dla p=1;
3. Opracowania modyfikacji modelu USVM dla przypadku, w którym grupy bazowa i kontrolna nie są wybierane w sposób w pełni losowy poprzez dodanie czynnika kary do postaci funkcji błędu.

Wymienione wyżej rezultaty badawcze zostały częściowo przedstawione w 3 publikacjach o zasięgu międzynarodowym: artykule w czasopiśmie z listy A MNiSW, rozdziale w monografii oraz materiałach z warsztatów naukowych.

## Konkluzja

Pracę doktorską mgr Łukasza Zaniewicza czyta się dobrze. Autor sprawnie operuje językiem angielskim, posiada bogate słownictwo, a poruszane zagadnienia przedstawione są w sposób spójny i logiczny, bazując na formalizmie matematycznym. Liczba błędów literowych czy stylistycznych, które zauważyłem w trakcie czytania jest znikoma i zdecydowanie mieści się w „zwyczajowych granicach".

Rozprawa zawiera szereg oryginalnych wyników Autora, istotnych z punktu widzenia rozwoju dziedziny modelowania różnicowego. Nie dostrzegłem w rozprawie istotnych braków czy nieprawidłowości, a wymienione w recenzji uwagi mają charakter polemiczny i nie zmniejszają mojej ogólnie wysokiej oceny dysertacji. Rozprawa dotyczy aktualnej i istotnej tematyki badawczej, a jej treść bez wątpienia dowodzi szerokiej wiedzy Autora, Jego dużej pomysłowości badawczej oraz głębokiego zrozumienia istoty rozważanych zagadnień.

Reasumując, **stwierdzam, że rozprawa spełnia wymagania stawiane przez odnośną Ustawę i wnoszę o jej przyjęcie oraz dopuszczenie jej Autora, mgr Łukasza Zaniewicza do dalszych etapów przewodu doktorskiego.** Ponadto, biorąc pod uwagę wysoki poziom merytoryczny rozprawy oraz jej wpływ na stan wiedzy w dziedzinie wykorzystania modeli uczenia maszynowego w zagadnieniu modelowania różnicowego **wnoszę o wyróżnienie rozprawy.**