Prof. dr hab. Włodzisław Duch
Katedra Informatyki Stosowanej i Laboratorium Neurokognitywne
Uniwersytet Mikołaja Kopernika, Toruń

Ocena osiągnięć naukowych doktora Macieja Zięby, w związku z ubieganiem się o stopień doktora habilitowanego, opracowana na prośbę Komisji ds. Stopni Naukowych w dyscyplinie informatyka techniczna i telekomunikacja.

Toruń, 7 marca 2021

Dr Maciej Zięba ma trzy dyplomy magisterskie: z informatyki uzyskany w 2009 roku w Instytucie Informatyki Politechniki Wrocławskiej, drugi za tą samą pracę w ramach umowy o podwójnym dyplomowaniu z Blekinge Institute of Technology, Szwecja, trzeci dyplom magisterski w Naukach o Zarządzaniu uzyskał w 2011 roku w Instytucie Organizacji i Zarządzania, Wydziału Informatyki i Zarządzania Politechniki Wrocławskiej. Doktorat z nauk technicznych w dyscyplinie informatyka uzyskał w 2014 roku w Instytucie Informatyki Politechniki Wrocławskiej, na podstawie rozprawy: Zespoły klasyfikatorów SVM dla danych niezbalansowanych. Widać więc, że uczeniem maszynowym zajmuje się od początku swojej kariery. Od 2016 roku zajmuje stanowisko adiunkta na Wydziale Informatyki i Zarządzania Politechniku Wrocławskiej, jest też wykładowcą na studia podyplomowe „Biznes.AI" Akademii Leona Koźmińskiego w Warszawie. Odbył półroczny staż doktorski na University of Wollongong w Australii.

Dokumentacja i autoreferat dr Zięby przygotowane zostały bardzo starannie. Jego działalność naukowa skupiona była wokół metod i algorytmów rozwiązujących podstawowe problemy z jakością i strukturą danych w zadaniach uczenia maszynowego. Rozprawa habilitacyjna oparta jest na cyklu 11 tematycznie powiązanych publikacji. Jedna z nich nie ma współautorów, 4 mają jednego współautora, pozostałe dwóch lub w więcej. Z oświadczeń wynika, że w 6 pracach współautorskich jego wkład był dominujący.

Celem badań było opracowanie szeregu metod rozwiązujących kilka problemów z wykorzystaniem danych w zadaniach uczenia maszynowego. Temat rozprawy jest bardzo aktualny, zastosowania metod uczenia maszynowego są coraz szersze, ale sukces zależy od jakości danych, używanych do treningu algorytmów. Dwie z publikacji doktora Zięby dotyczą pro-

WPŁYNĘŁO
08-03-2021
RDK ITiT/50/2021

blemu brakujących wartości niektórych elementów danych uczących, 4 dotyczą problemu stosunkowo niewielkiej liczebności danych dla niektórych klas, a pozostałe problemu ekstrakcji przydatnych cech z pierwotnych danych, by przygotować użyteczną strukturę danych uczących. Nie są to oczywiście zagadnienia nowe, jest wiele algorytmów, które w różny sposób pomagają je rozwiązać. Warto zaznaczyć, że propozycje dr Zięby nie dotyczą abstrakcyjnych algorytmów, ale pojawiają się w publikacjach na temat zastosowań uczenia maszynowego do konkretnych zagadnień, a więc ich motywacją było rozwiązywanie realnych problemów. Jak wypadają te prace na tle specjalistycznej literatury?

- Problem brakujących wartości atrybutów.

Pierwsza z publikacji na ten temat, w *IEEE Journal of biomedical and health informatics* z 2014 roku, dotyczy systemu wspomagania decyzji diagnostycznych w medycynie, gdzie często mamy do czynienia z brakującymi wartościami. Celem było stworzenie programu, który jest niewrażliwy na brakujące wartości atrybutów. Został on wdrożony jako komponent systemu działającego w oparciu o usługi webowe i przetestowany na rzeczywistych danych dotyczących przeżywalności pacjentów po poważnej operacji związanej z rakiem płuc. Zastosowano prostą metodę opartą na komitetach SVM uczonych na różnych podzbiorach danych, w których nie ma brakujących wartości. Takie podejściem jest uniwersalne, można zastosować dowolny algorytm tworząc modele i łącząc je w komitety. Jest też znacznym uproszczeniem, bo nie rozróżnia sytuacji, w której brakujące wartości wynikają z celowego działania: po zrobieniu wstępnych testów często nie ma potrzeby dalszych, niektóre testy zastępują podobne, inne się wykluczają. Rzeczywista sytuacja jest więc znacznie bardziej skomplikowana i wymaga głębszego zrozumienia diagnostyki medycznej dla określonego typu chorób. Drugi problem z takim podejściem widać w przypadku danych Breast Cancer z Ljubliany. Dokładność komitetu E-SVM wynosi 65.3%, a jego wynik trudno jest zinterpretować. Metody ekstrakcji reguł logicznych pokazują, że z tej bazy można wyciągnąć jedną prostą łatwo interpretowalną regułę o dokładności ponad 77%: jeśli są przerzuty a złośliwość jest duża będzie nawrót choroby (opublikowaliśmy to w 2001 roku w IEEE Transactions on Neural Networks). Zawsze warto zaczynać od najprostszych modeli zanim sięgnie się po bardziej złożone.

Druga praca zajmująca się tym tematem ma 4 współautorów i została opublikowana w 2018 roku czasopiśmie *Applied Soft Computing*. Chociaż zaproponowany tu algorytm uzupełniania brakujących wartości dla danych sekwencyjnych przez przewidywanie kolejnych elementów

w sekwencji za pomocą procesu Gaussa jest uniwersalny, został opisany w publikacji dotyczącej 2-wymiarowych widm magnetycznego rezonansu jądrowego (NMR). Jest to aplikacja do ważnego, praktycznego problemu, co zasługuje na uznanie. Wyszukiwanie najbardziej prawdopodobnych położeń szczytów sygnałów i określenie ich niepewności jest zadaniem trudnym. Technika 2-D NMR nie zawsze produkuje sygnały tam, gdzie się ich należy w widmie spodziewać, z czym model procesów gaussowskich sobie radzi. Testy dały zaskakujący wynik, powszechnie używane programy do analizy takich widm robią wiele błędów. Ten algorytm nadaje się do innych problemów uzupełniania brakujących wartości dla atrybutów sekwencyjnych. Inspiracją dla jego opracowania była praca na temat przewidywania toru ruchu pieszych.

- Problem niezbalansowania danych uczących.

4 publikacje dotyczą problemu stosunkowo niewielkiej liczebności danych dla niektórych klas. Wiele metod klasyfikacji wykazuje w takiej sytuacji nadmierne preferencje dla klasy najliczniejszej. Jedno z podejść do radzenia sobie z tym problemem polega na generowaniu dodatkowych danych. Algorytm SMOTE (*Synthetic Minority Over-sampling Technique*) rozwijany od 15 lat i stosowany w bardzo wielu wariantach. Modyfikacja zaproponowana w 2015 roku przez dr Ziębę i współpracowników podobna jest w istocie do modnych obecnie metod generatywnych. Stosuje Ograniczone Maszyny Boltzmanna (RBM) do transformacji danych do przestrzeni ukrytej, w której generowane są syntetyczne dane z wykorzystaniem algorytmu SMOTE. Dzięki temu po transformacji odwrotnej do przestrzeni pierwotnych danych ich rozkład reprezentuje oryginalny rozkład na większej ilości danych. To intersujące i naturalne podejście, które spotkało się z zainteresowaniem innych badaczy.

W drugiej pracy na temat niezbalansowanych danych uczących, opublikowanej w Machine Learning (Tomczak, Zięba, 2015) zastosowano ważone kombinacje reguł logicznych, tworząc „miękkie reguły" decyzyjne. Nowością metody, nazwanej komitetami miękkich reguł (CSR, w dwóch miejscach chyba omyłkowo jest CSF) jest wprowadzenie dodatkowych zmiennych przez koniunkcję pierwotnych. To dobry kierunek badań, wiele metod uczenia maszynowego opiera się na tworzeniu użytecznych nowych cech, np. Functional Link Networks (Pao, 1995), czy też nasze algorytmy Universal Learning Machines (2009) i Support Feature Machines (2010). Problem XOR staje się liniowo separowalny jeśli wprowadzić iloczyn obu wejść jako nową zmienną. Drugą nowością omawianej pracy jest wprowadzenie reprezentacji grafowej

do obliczeń prawdopodobieństw warunkowych i zastosowanie wnioskowania Bayesowskiego. Pomimo tego, jak się należało spodziewać, mamy szybki wzrost czasu obliczeń przy większej liczbie atrybutów. Testy przeprowadzono na zestawie 45 zbiorów referencyjnych z repozytorium UCI, w tym 8 z obszaru diagnostyki medycznej (w tabelach diabetes i pimalmb to ten sam zbiór, powtarza się w części pozamedycznej). Te zbiory danych różnią się znacznie stopniem zbalansowania klas. Korektą na zbyt małą liczebność klas jest w tym przypadku znana *m*-estymata prawdopodobieństwa, która w liczniku i mianowniku ilorazu częstości danej klasy do liczby wszystkich przypadków ma dodatkowy parametr *m*, wyznaczany eksperymentalnie. Emipryczne testy jakości działania komitetów miękkich reguł wypadają podobnie jak inne zespoły klasyfikatorów i metoda Naiwnego Bayesa.

Trzecia praca z tej serii, opublikowana na konferencji w 2016 roku, oparta jest na wariancie metody uczenia się we własnym tempie (ang. self-paced learning), czyli wyborze w pierwszej kolejności prostych przypadków, dodając trudniejsze na dalszym etapie uczenia. Trudność klasyfikacji oceniana jest na podstawie analizy funkcji błędu dla rozpatrywanego przykładu. Wymaga to oceny wartości funkcji błędu, która została przez autorów zmodyfikowana przez dodanie parametrów kosztów pomyłek dla różnych klas. W omawianej pracy zrobiono to przy pomocy niezależnej sieci neuronowej, co nie jest najlepszym rozwiązaniem. W przypadku sieci neuronowej łatwe przypadki dają wartość funkcji bliską oczekiwanej. Wystarczy więc ustalić próg by ocenić, które dane są łatwe (daleko od granicy decyzji), a które trudne. W pracy *Support Vector Neural Training* (LNCS 2005) użyłem podobnej metody w bardziej efektywny sposób, pozostawiając po każdej iteracji tylko te wektory, dla których wartości funkcji realizowanej przez sieć nie były dostatecznie blisko oczekiwanym. W efekcie na końcu uczenia zostają tylko nieliczne wektory brzegowe, uczenie jest szybkie i nie trzeba dodatkowych parametrów. Np. na danych dotyczących chorób tarczycy z podziałem klas ok. 2.5, 5,0 i 92.5% udało się w ten sposób osiągnąć rezultaty na najwyższym poziomie.

Ostatnia praca z tej serii, opublikowana w 2018 roku, dotyczy metody konstrukcji komitetu modeli klasyfikacyjnych przy użyciu metody boostingu, czyli specyficznej metody próbkowania danych. Przez analogię z popularnym algorytmem AdaBoost, wykorzystanie w próbkowaniu rozkładu beta-dwumianowego nazwano BetaBoost. Dla każdej z klas tworzony jest osobny rozkład beta-dwumianowy, co pozwala na wyrównanie liczby danych w klasie. Zgodnie z filozofią boostingu robi się ranking bazowych klasyfikatorów, osobno dla każdej klasy.

Zastosowano dwie strategie, w zależności od zaufania do jakości danych uczących. Te założenia wpływają na parametry rozkładu a przez to sposób próbkowania klas. Jeśli zaufanie do jakości danych jest niskie przywiązuje się mniejszą wagę do błędów, czyli przypadków „trudnych", jeśli zaufanie jest wysokie próbuje się je uwzględnić. To przypomina optymalizację marginesu w metodzie SVM, gdzie parametr C wpływa na lepsze dopasowanie lub większą generalizację modelu i jest oceniany w procesie kroswalidacji. Można było problem wyboru strategii tak właśnie rozwiązać. Obydwa podejścia przetestowano na dwóch silnie dużych zbiorach z obszaru oceny ryzyka kredytowego, w których liczba przypadków w klasie pozytywnej stanowi poniżej 10% negatywnej. Porównania jakości za pomocą krzywych AUC pokazały, że BetaBoost działa znacznie lepiej niż zwykle stosowane zbalansowane próbkowanie (balanced bagging). Niestety takie metody nie dają nam wglądu w najważniejsze atrybuty dla podjętej decyzji.

Pozostałe prace dr Zięby przedstawione jako osiągnięcie naukowe dotyczą właśnie

- interpretowalności modeli uczenia maszynowego.

Explainable AI to obecnie bardzo aktualny temat. W pracy z 2015 roku w czasopiśmie *Expert Systems with Applications* Tomczak i Zięba opisali zastosowanie Ograniczonych Maszynach Boltzmanna (RBM) do konstrukcji interpretowalnego klasyfikatora do scoringu kredytowego. Zwykle stosuje się tablice scoringowe do zliczania liczby punktów przypisanych do danego przypadku, wykorzystując model regresji logistycznej. Pozwala to na uzasadnienie decyzji, do którego zobligowane są instytucje finansowe w wielu krajach. Jest to popularny temat badań i zaproponowano tu wiele rozwiązań. Motywacją do zastosowania RBM jest możliwość stworzenia dobrego, interpretowalnego modelu klasyfikacyjnego na niezbalansowanych binarnych danych. Sieci RBM uczą się rozkładu prawdopodobieństw łącznych aktywacji wszystkich jednostek. Dla każdej z klas można określić prawdopodobieństwo pojawienia się danej cechy wejściowej zakładając zerowe wejścia pozostałych cech. Nie zawsze jest to najlepsze rozwiązanie, nie podano jednak uzasadnienia takiego założenia. Tak obliczone prawdopodobieństwo określa istotność cechy w tablicy scoringowej, a punkt odcięcia, powyżej którego przyznawany jest kredyt, wyznaczany jest za pomocą zbioru walidacyjnego.

Metoda oceny ryzyka kredytowego testowana była przy użyciu 4 zbiorów, od niecałych 500 przypadków do 150.000. Na największym z tych zbiorów tablice scoringowe oparte na RBM osiągnęły bardzo dobry wynik, wyraźnie lepszy niż drzewa decyzji i metoda regułowa RIPPER.

Jedynie w przypadku najmniejszego zbioru danych, *Australian Credit*, szereg metod okazało się lepszych. Znam kilka przykładów, w których decyzje kredytowe można było wyjaśnić za pomocą pojedynczego atrybutu lub podobieństwa do jednego prototypu. Jeden z nich dotyczy właśnie zbioru *Australian credit*. Widać więc, że trzeba najpierw sprawdzać najprostsze modele bo złożone metody nie działają w takich przypadkach najlepiej.

W najczęściej cytowanej publikacji dr Zięby i współpracowników *"Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction"* (2016) wykorzystano ewolucyjną procedurę selekcji najistotniejszych cech. Ciekawym pomysłem jest tu wykorzystanie syntetycznych cech, czyli arytmetycznych kombinacji danych wejściowych. Ponieważ uczenie drzew decyzji jest szybkie i automatycznie odrzuca nieprzydatne cechy można to zrobić na dużą skalę. Jako algorytm adaptacji zastosowano metodę Extreme Gradient Boosting. Końcowy model to komitet losowych lasów z ważonym głosowaniem.

Ta metoda została zastosowana na realnych danych dotyczących przewidywania bankructwa polskich firm w skali czasowej od 1 do 5 lat w oparciu o 64 wskaźniki kondycji firmy. Do porównania użyto 13 metod klasyfikacji. Rezultaty opracowanej metody są znacząco lepsze niż metod konkurencyjnych, a ranking wygenerowanych ewolucyjnie kombinacji cech daje pewien wgląd w syntetyczne wskaźniki, chociaż interpretacja większości tych cech nie jest prosta. Warto rozważyć inne metody generacji przydatnych cech, o czym wspomniałem powyżej.

Na konferencji ICRL (International Conference on Learning Representations) w 2017 roku Maciej Zięba i Lei Wang przedstawili pracę *Training triplet networks with GAN*. Temat częściowo nadzorowanego uczenia się rozwijany jest od dawna. W tej pracy zastosowanie model generatywny typu GAN (*Generative Adversarial Nets*), do uczenia się cech na warstwach pośrednich sieci dyskryminującej. Trójkowe funkcje strat w przypadku binarnej klasyfikacji oparte są na prawdopodobieństwach wyliczonych z odległości pomiędzy wybranym wektorem i pojedynczymi wektorami z obu klas. Zadaniem generatora jest tworzenie rozkładów aktywacji warstwy ukrytej dyskryminatora nieodróżnialnych od prawdziwych, a dyskryminatora uczenie dopasowywania cech w trybie nadzorowanym. Celem jest stworzenie jak najlepszej reprezentacji rozkładu analizowanych danych.

Testy przeprowadzono na zbiorze małych obrazków CIAR-10 oraz zbiorze ręcznie pisanych cyfr MNIST, zawierających odpowiednio 50000 i 60000 elementów podzielonych na 10 klas.

Wybrano niewielką liczbę przypadków z przypisanymi klasami. Do klasyfikacji zastosowano algorytm najbliższych sąsiadów w oparciu o metrykę trójkową. Porównanie ze zwykłą siecią trójkową i modelem typu GAN bez regularyzacji sieci trójkowej wypada bardzo korzystnie dla kombinacji obu metod. Praca oparta jest na najnowszych doniesieniach naukowych i napisana jest w bardzo zwarty sposób, zapewne ze względu na ograniczenia konferencyjne. Nie jest jasne jak dobrano parametry (liczbę cech i sąsiadów).

Kolejna praca, opublikowana z udziałem 3 współautorów na najwyżej notowanej w tej dziedzinie konferencji NIPS w 2018 roku, dotyczy również uczenia się reprezentacji danych w nienadzorowany sposób przy wykorzystaniu techniki GAN. Model BinGAN wykorzystuje sieć dyskryminującą, której warstwa pośrednia podlega regularyzacji. W warstwie o dużej liczbie wymiarów policzono odległości Hamminga pomiędzy reprezentacjami obiektów wejściowych, które zostały użyte do regularyzacji głębszej warstwy, mającej dostarczyć binarnej reprezentacji obiektów. To podejście określono jako Distance Matching Regularizer (DMR). Drugą formą regularyzacji jest maksymalizacja entropii reprezentacji binarnych warstwy docelowej sieci dyskryminującej, zwiększająca różnorodność binarnych kodowań. Dzięki temu otrzymano zwartą reprezentację binarną z warstwy pośredniej dyskryminatora. Pozwala to zwiększyć niezależność pomiędzy reprezentacjami różnych klas. Trening modelu nie wymaga etykiet, na zmianę uruchamiany jest generator i dyskryminator. Reprezentacje binarne utworzone w ten sposób można wykorzystać jako nowe deskryptory obrazów.

Testy przeprowadzono na bazie CIFAR-10. W zadaniu dopasowywania obrazów użyto sieci o całkowitej liczbie warstw równej 10, a docelowe reprezentacje binarne miały wymiar 256. Podobna sieć została zastosowana w zadaniu wyszukiwania obrazów. Wyniki porównano z 10 innymi metodami opartymi na technikach haszowania i dla 16, 32, 64 bitów BinGAN wykazał się znaczącą przewagą. Metoda pozwala na generację nowych fragmentów obrazów, które można używać do uczenia nadzorowanego. Autorzy udostępnili też oprogramowanie pozwalające odtworzyć ich wyniki. Jest to bez wątpienia ważna publikacja. Nie podano szczegółów dotyczących złożoności obliczeń, ale nie powinna to być przeszkoda by rozwinąć praktyczne zastosowania.

Ostatnia omawiana publikacja ma 6 współautorów a swój wkład dr Zięba szacują na około 1/3. Praca ta dotyczy zastosowania uczenia reprezentacji danych w trybie nienadzorowanym przy użyciu generatywnego modelu dla chmur punktów w 3 wymiarach. W tym przypadku

użyto modelu opartego na głębokim autoenkoderze (*adversarial* nie ma dobrego tłumaczenia, może „zwodniczy" oddaje tu sens) AAE dostosowanego do danych w 3 wymiarach. Jest to nietrywialne rozszerzenie modelu używanego do analizy obrazów. Proponowany model 3dAEE pozwala na wyuczenie się kompaktowej reprezentacji binarnej rozkładów punktów w warstwie ukrytej przy użyciu techniki GAN. Na podstawie wyuczonej reprezentacji możliwa jest generacja rozkładu punktów w 3 wymiarach.

Testy wykonano na zbiorze referencyjnym ShapeNet. Możliwości generatywne zostały sprawdzone z wykorzystaniem standardowych miar. Operacje na reprezentacjach warstwy ukrytej pozwalają na generację nowych obiektów, taka sieć może wykazać się kreatywnością. Jakość binarnych reprezentacji oceniono wykorzystując liniowego klasyfikatora SVM i jest ona porównywalna z najlepszymi znanymi metodami. W tej publikacji jest też wiele innych nowości.

Google Scholar na początku marca 2021 roku uwzględniał 49 publikacji dr Zięby, a liczba jego cytowań sięgała 650, indeks Hirscha równy wynosił 12. W Semantic Scholar (Allen Institute for AI) rezultaty są prawie identyczne, podana jest też liczba cytowań o dużym wpływie (*infuential citation*), która wynosi 34. W ciągu ostatnich dwóch lat liczba cytowań prac dr Zięby wzrosłą dwukrotnie i tylko w ubiegłym roku była wynosiła prawie 200. W bazie Scopus było 29 dokumentów i 339 cytowań, oraz indeks h=9. Widać też rosnące wyrafinowanie opracowywanych przez niego algorytmów.

W sumie jest to znaczący dorobek naukowy w ważnej i bardzo aktualnej specjalności informatyki technicznej. Nie mam wątpliwości, że dr Zięba posiada kwalifikacje do samodzielnej pracy naukowo-badawczej a uznanie jego wyników naukowych będzie dalej szybko rosło. Przedstawiony mi do oceny materiał uważam za w pełni wystarczający, by spełnić ustawowe i zwyczajowe wymagania dotyczące nadania mu stopnia doktora habilitowanego.

Andrzej Dudek

Toruń, 11.12.2017