Kraków, 11.02.2022 r.

Marek Skomorowski (prof. dr hab. inż.)
Instytut Informatyki i Matematyki Komputerowej
Uniwersytetu Jagiellońskiego

Recenzja rozprawy doktorskiej
Pana mgra inż. Jakuba Klikowskiego
zatytułowanej
„Ensemble methods for imbalanced data stream classification"
(Zastosowanie zespołów klasyfikatorów w klasyfikacji niezbalansowanych strumieni danych)

Przewód doktorski jest prowadzony przez Radę Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Politechniki Wrocławskiej (RDNITiT PWr). Recenzja została napisana na zlecenie Przewodniczącego RDNITiT PWr, Pana prof. dra hab. inż. Michała Woźniaka (pismo z dnia 13.12.2021 roku).

1. Problem badawczy, jego znaczenie i aktualność

Klasyfikacja jest ważnym zadaniem w rozpoznawaniu wzorców. Standardowe algorytmy klasyfikacji, takie jak drzewo decyzyjne, sieć neuronowa, naiwny klasyfikator Bayesa, najbliższy sąsiad, czy maszyna wektorów nośnych są stosowane w praktyce. Jednak występujące w rzeczywistych zbiorach danych niezbalansowanie rozkładu obiektów w klasach (ang. class imbalance) powoduje trudności w fazie uczenia i obniża jakość predykcji w standardowych algorytmach klasyfikacji, które zakładają stosunkowo zbalansowany rozkład obiektów w klasach. Rozwój badań dotyczących klasyfikacji niezbalansowanych strumieni danych jest motywowany licznymi rzeczywistymi zastosowaniami, na przykład: diagnostyka medyczna i techniczna, wykrywanie oszustw dokonywanych za pomocą kart kredytowych i telefonów komórkowych, detekcja niechcianej poczty i włamań do sieci, czy też systemy nadzoru. W klasyfikacji niezbalansowanych strumieni danych oprócz niezbalansowania klas występują problemy typowe dla klasyfikacji strumieni danych, takie jak ograniczone zasoby obliczeniowe i pamięciowe, trudności w pozyskiwaniu etykiet, jak również zjawisko zwane dryfem koncepcji (ang. concept drift), powodujące zmiany w rozkładzie danych w trakcie ich napływu, co pogorsza jakość klasyfikacji.

Pomimo stałego rozwoju, problem klasyfikacji niezbalansowanych strumieni danych jest w dalszym ciągu traktowany jako trudny i jest przedmiotem intensywnych badań, zarówno od strony teoretycznej, jak również zastosowań praktycznych. Metody na poziomie danych i algorytmów, jak również podejście hybrydowe, łączące zalety dwóch poprzednich technik są stale ulepszane. Badania nad problemem niezbalansowania klas mają kluczowe znaczenie w uczeniu maszynowym i eksploracji danych.

Problematyka rozprawy doktorskiej Pana mgra inż. Jakuba Klikowskiego dotyczy klasyfikacji binarnej niezbalansowanych strumieni danych z dryfem koncepcji, to znaczy klasyfikacji do jednej z dwóch klas, z których liczebność jednej jest znacznie większa (klasa większościowa) niż drugiej (klasa mniejszościowa), a w trakcie napływu danych następują zmiany w ich rozkładzie. W przypadku niezbalansowanych strumieni danych standardowe algorytmy uczenia maszynowego preferują klasę większościową, co stanowi problem szczególnie dla klasy mniejszościowej dlatego, że koszty błędnej klasyfikacji klasy mniejszościowej są zazwyczaj wysokie. Najczęściej klasa mniejszościowa (pozytywna) jest ważniejsza niż klasa większościowa (negatywna) i dlatego problem ten jest bardziej wrażliwy na błędy klasyfikacji dla klasy mniejszościowej niż dla klasy większościowej. Na przykład, koszt nierozpoznania osoby chorej na rzadką ale niebezpieczną chorobę może być znacznie wyższy niż koszt błędnej klasyfikacji osoby zdrowej. Rozprawa powstała w wyniku realizacji projektu badawczego zatytułowanego „Algorytmy klasyfikacji niezbalansowanych strumieni danych", finansowanego ze środków Narodowego Centrum Nauki (nr 2017/27/B/ST6/01325), w którym Doktorant jest jednym z wykonawców.

Podsumowując, podjęta w rozprawie problematyka dotycząca klasyfikacji niezbalansowanych strumieni danych jest ważna i aktualna, zarówno z poznawczego jak również praktycznego punktu widzenia. Tematyka rozprawy lokuje się w dynamicznie rozwijającym się obszarze uczenia maszynowego w ramach sztucznej inteligencji.

2. Teza i cel rozprawy

Teza rozprawy została sformułowana w rozdziale pierwszym w następujący sposób:

Można zaprojektować metody zespołów klasyfikatorów wykorzystujące próbkowanie danych oraz jednoklasowe klasyfikatory, które potrafią osiągnąć jakość predykcji lepszą niż znane z literatury klasyfikatory niezbalansowanych strumieni danych.

Sformułowano również 6 zadań badawczych realizowanych w ramach rozprawy w celu udowodnienia jej tezy. Zadania badawcze są dobrze uzasadnione i zostały zdefiniowane na podstawie wnikliwego przeglądu literatury, uwzględniającego aktualne wyniki badań w zakresie problematyki rozprawy, to znaczy do 2021 roku włącznie. Sformułowanie zadań badawczych dobrze świadczy o przygotowaniu Autora do działalności naukowej. W trakcie realizacji zadań badawczych Autor stawia pytania badawcze, na które odpowiada po analizie przeprowadzonych eksperymentów komputerowych.

Celem rozprawy było zaprojektowanie, implementacja i weryfikacja nowych klasyfikatorów niezbalansowanych strumieni danych.

3. Zawartość rozprawy

Rozprawa jest napisana w języku angielskim, liczy 188 stron i składa się z pięciu rozdziałów, spisu treści, wykazu bibliografii zawierającego 301 pozycji, wykazu skrótów i symboli.

W rozdziale pierwszym omówiono genezę problematyki badawczej, przedstawiono motywację podjętych badań, tezę rozprawy i realizowane zadania badawcze. Rozdział pierwszy zawiera również opis zawartości rozprawy.

Rozdział drugi dotyczy wybranych tematów klasyfikacji wzorców. Przedstawiono w nim zagadnienia uczenia maszynowego nadzorowanego, nienadzorowanego, częściowo nadzorowanego i ze wzmocnieniem. Opisano takie zagadnienia klasyfikacji jak ekstrakcja cech, maszyna wektorów nośnych, naiwny klasyfikator Bayesa, drzewo decyzyjne, perceptron wielowarstwowy i klasyfikatory minimalno-odległościowe. Omówiono algorytm klasyfikacji jednoklasowej One Class Support Vector Machine (OCSVM). Przedstawiono idę budowy zespołów klasyfikatorów. Opisano problemy dotyczące wyboru klasyfikatorów tworzących zespół i zagadnienia związane z ustalaniem wspólnej decyzji. Omówiono wybrane problemy klasyfikacji danych niezbalansowanych. Przedstawiono pojęcie stopnia niezbalansowania danych i podano przykłady instancji klas mniejszościowych. Opisano metryki (miary) do oceny jakości klasyfikacji. W ramach metod klasyfikacji niezbalansowanych danych na poziomie danych omówiono usuwanie obiektów z grup większościowych (ang. undersampling) i generowanie nowych obiektów do grup mniejszościowych (ang. oversampling), jak również metody łączące obie wymienione techniki. Przedstawiono również podejście do klasyfikacji niezbalansowanych danych na poziomie algorytmów. Omówiono typy dryfu koncepcji i detektory dryfu. Opisano metody online przetwarzania strumieni danych, jak również metody wykorzystujące bloki danych. Przedstawiono metody klasyfikacji niezbalansowanych strumieni danych za pomocą przetwarzania online i przetwarzania blokowego. Opisano metody eliminacji szumów związanych z etykietami i atrybutami. Rozdział drugi stanowi niezbędne wprowadzenie do problematyki rozprawy.

Rozdział trzeci stanowi realizację dwóch pierwszych z sześciu sformułowanych zadań badawczych. Na oryginalny dorobek naukowy Autora przedstawiony w tym rozdziale, składają się zaproponowane dwie metody klasyfikacji niezbalansowanych strumieni danych. Pierwsza z nich, opisana w podrozdziale 3.1 i nazwana One Class support vector machine classifier Ensemble for Imbalanced data Stream (OCEIS) wykorzystuje jednoklasowe maszyny wektorów nośnych (OCSVM). Zaprojektowano zespół złożony z wielu klasyfikatorów jednoklasowych wykorzystujący przetwarzanie blokowe i klastrowanie danych uczących, w którym każdy model jest uczony na innym zbiorze danych. Zaproponowano oryginalną regułę decyzyjną. Przeprowadzono eksperymenty komputerowe, których celem była weryfikacja metody OCEIS i porównanie jej z innymi znanymi metodami klasyfikacji niezbalansowanych strumieni danych. Eksperymenty przeprowadzono zarówno na danych pozyskanych z generatora strumieni danych, jak również na danych rzeczywistych. Z analizy eksperymentów wynika, że metoda OCEIS osiąga rezultaty porównywalne z innym metodami. Zaletą metody OCEIS jest to, że najlepiej sprawdza się na rzeczywistych strumieniach danych i nie ma tendencji do nadmiernej klasyfikacji obiektów do jednej z klas, tak jak ma to miejsce w niektórych porównywanych metodach (REA, OUSE). Metoda OCEIS została opublikowana w materiałach międzynarodowej konferencji „International Conference on Computational Science" w 2020 roku (pozycja [132] w spisie bibliografii rozprawy).

Druga z zaproponowanych przez Autora oryginalnych metod klasyfikacji niezbalansowanych strumieni danych, opisana w podrozdziale 3.2 i nazwana One Class support vector machine Weighted Ensemble (OCWE), jest udoskonaleniem metody OCEIS, przede wszystkim przez zastosowanie oryginalnej reguły decyzyjnej z głosowaniem ważonym. Przeprowadzono eksperymenty komputerowe, których celem była weryfikacja metody OCWE i porównanie jej z innymi znanymi metodami klasyfikacji niezbalansowanych strumieni danych. Eksperymenty przeprowadzono wykorzystując następujące klasyfikatory bazowe: k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB) i Decision Tree CART (DTC). Dla metody OCWE została przeprowadzona szersza analiza eksperymentalna niż dla metody OCEIS. Pierwszy eksperyment skupia się na testowaniu różnych algorytmów klastrowania w celu oszacowania potencjalnie najlepszej liczby klastrów. Drugi eksperyment dotyczy wyboru najlepszych wartości parametrów, które określają, w jakim stopniu czynniki wagowe wpływają na ostateczną decyzję. Trzeci eksperyment porównuje działanie metody OCWE z działaniem innych metod w przypadku strumieni danych o dynamicznym współczynniku niezbalansowania klas. Czwarty eksperyment koncentruje się na strumieniach danych z nagłym i przyrostowym dryfem koncepcji. Dane do wymienionych eksperymentów pozyskiwano z generatorów strumieni danych, natomiast piąty eksperyment przeprowadzono korzystając z danych rzeczywistych. Z analizy przeprowadzonych eksperymentów wynika, że metoda OCWE polepszyła jakość klasyfikacji w porównaniu do metody OCEIS. Metoda OCWE w porównaniu do innych metod osiąga podobną jakość klasyfikacji, a w pewnych warunkach lepszą. Zaletą metody OCWE jest to, że dobrze nadaje się do zadań klasyfikacji danych silnie niezbalansowanych i posiada zdolność adaptacji do aktualnie rozwiązywanego problemu za pomocą parametrów.

Rozdział czwarty stanowi realizację pozostałych czterech z sześciu sformułowanych zadań badawczych. Na oryginalny dorobek naukowy Autora przedstawiony w tym rozdziale, składa się zaproponowana rodzina czterech nowych metod wykorzystujących techniki wstępnego przetwarzania danych do klasyfikacji niezbalansowanych strumieni danych. Przeprowadzono eksperymenty komputerowe, których celem była weryfikacja i ocena zdolności predykcyjnych zaproponowanej rodziny metod i porównanie jej z innymi znanymi metodami klasyfikacji niezbalansowanych strumieni danych. Eksperymenty przeprowadzono wykorzystując następujące klasyfikatory bazowe: KNN, SVM, GNB i DTC.

W podrozdziale 4.1 Autor zaproponował oryginalną metodę nazwaną Deterministic Sampling Classifier (DSC). Metoda ta wykorzystuje bufor pamięci o stałym rozmiarze do propagacji wybranych instancji z poprzednich fragmentów strumienia danych. W zaproponowanej metodzie przyjęto, że po zapełnieniu bufora pamięci usuwane są najstarsze instancje. Metoda DSC umożliwia ustawianie parametrów w zależności od rozważanego problemu. Z analizy przeprowadzonych eksperymentów wynika, że metoda DSC dobrze sobie radzi z danymi niezbalansowanymi zarówno z nagłym, jak również z przyrostowym dryfem koncepcji. Wyniki uzyskane na danych rzeczywistych wskazują na dobrą zdolność klasyfikacyjną metody DSC do tego typu danych. W przypadku niektórych klasyfikatorów bazowych metoda DSC osiąga przewagę nad innymi metodami. Wykorzystując syntetyczne strumienie danych metodę DSC porównano z innymi znanymi metodami. W większości przeprowadzonych porównań metoda

DSC osiąga przewagę nad innymi metodami. Metoda DSC została opublikowana w materiałach międzynarodowej konferencji „Joint European Conference on Machine Learning and Knowledge Discovery in Databases" w 2019 roku ([28]).

W podrozdziale 4.2 Autor zaproponował oryginalną metodę nazwaną Deterministic Sampling Classifier with weighted Bagging (DSCB). Metoda ta, będącą udoskonaleniem metody DSC stanowi zespół klasyfikatorów, którego ideą jest tworzenie nowych modeli za pomocą danych, które zostały zgromadzone z poprzednich części strumienia danych. Z analizy przeprowadzonych eksperymentów wynika, że metoda DSCB dobrze sobie radzi z danymi o różnym współczynniku niezbalansowania klas, nawet w wtedy kiedy klasa mniejszościowa zawiera tylko 5 % obiektów. Metoda DSCB jest dość odporna na szumy związane z etykietami. W przypadku strumieni danych o dynamicznym współczynniku niezbalansowania klas metoda DSCB osiąga przewagę nad innymi znanymi metodami. W ogólnej ocenie metoda DSCB uzyskuje lepszą wydajność predykcyjną w porównaniu z innymi metodami. Dotyczy to zarówno danych syntetycznych, jak również rzeczywistych.

W podrozdziale 4.3 Autor zaproponował oryginalną metodę nazwaną Deterministic Sampling Ensemble (DSE), będącą rozszerzeniem metody DSC. Metoda DSE bazuje na akumulacji wybranych próbek z poprzednich fragmentów w celu późniejszego wzmocnienia i zrównoważenia danych. Takie podejście nie wymaga generowania sztucznych danych. Metoda DSE rozszerza metodę DSC z pojedynczego klasyfikatora na zespół z głosowaniem większościowym. Kolejnym rozszerzeniem metody DSE w porównaniu do metody DSC jest wprowadzenie mechanizmu przeciwdziałania dryfowi koncepcji. Mechanizm ten składa się z dwóch części. Pierwsza z nich to detektor dryfu, a druga to procedura zapominania, która usuwa najstarsze modele i najstarsze zgromadzone fragmenty danych. Z analizy przeprowadzonych eksperymentów wynika, że metoda DSE dobrze sprawdza się zarówno w przypadku danych zaszumionych, jak również niezaszumionych. Metoda DSE jest statystycznie lepsza od innych znanych metod dla większości strumieni danych, na których była testowana za pomocą różnych klasyfikatorów bazowych, w szczególności DTC i SVM. Metoda DSE osiąga dobre wyniki zarówno przy nagłych, jak również przyrostowych dryfach koncepcji.

W podrozdziale 4.4 Autor zaproponował metodę nazwaną Deterministic Sampling Ensemble of One Class Support Vector Machine classifiers (DSE-OC), będącą połączeniem wybranych cech metod DSE i OCEIS. To co jest oryginalne w tej metodzie dotyczy głównie gromadzenia danych z klasy większościowej i modyfikacji metody OCEIS stosowanej jako klasyfikator bazowy. Z analizy przeprowadzonych eksperymentów wynika, że metoda DSE-OC osiąga dobrą jakość i jest statystycznie lepsza od innych znanych metod.

Wszystkie zaproponowane przez Doktoranta w rozprawie metody zostały porównane z następującymi znanymi klasyfikatorami niezbalansowanych strumieni danych: REA, KMC, Learn++CDS, Learn++NIE, OUSE, MLPC i ocenione za pomocą pięciu następujących metryk: $Gmean_s$, F_1score, Recall, Specifity, Precision. Oszacowano złożoność obliczeniową wszystkich zaproponowanych algorytmów. Wszystkie zaproponowane metody, jak również eksperymenty

zaimplementowano w języku Python, a ich implementacje, jak również środowiska eksperymentalne i wyniki przeprowadzonych eksperymentów są dostępne publicznie.

Rozdział piąty stanowi podsumowanie rozprawy i zawiera uwagi na temat dalszych badań.

4. Redakcja rozprawy

Tytuł rozprawy właściwie odzwierciedla jej zawartość. Wykazy skrótów i symboli w kolejności alfabetycznej, umieszczone na początku rozprawy ułatwiają jej lekturę. Rozprawa jest napisana czytelnie, a jej zawartość dobrze przemyślana zarówno od strony merytorycznej, jak również metodologicznej. Kolejne rozdziały i podrozdziały logicznie następują po sobie. Rozprawa prezentuje wysoki poziom edytorski. Wzory matematyczne są czytelnie wyeksponowane. Każdy algorytm zaproponowany przez Autora jest przedstawiony, zarówno w postaci schematu opisującego ideę metody, jak również w postaci pseudokodu. Bogata i starannie dopracowana strona graficzna zawiera liczne ilustracje przedstawiające wyniki przeprowadzonych eksperymentów komputerowych w postaci czytelnych wykresów i tablic.

5. Oryginalny wkład Autora

Przedstawiona w rozprawie analiza wyników eksperymentów przeprowadzonych za pomocą zaprojektowanych i zaimplementowanych przez Autora klasyfikatorów potwierdza, że wszystkie zadania badawcze zostały zrealizowane, a tym samym została udowodniona teza rozprawy. Na oryginalny dorobek Autora przedstawiony w rozprawie w zakresie problematyki klasyfikacji niezbalansowanych strumieni danych składają się:

1. Zaprojektowanie zespołu klasyfikatorów jednoklasowych (OCEIS) do klasyfikacji niezbalansowanych strumieni danych (podrozdział 3.1, [132]).
2. Modyfikacja zespołu klasyfikatorów jednoklasowych (OCEIS) przez wprowadzenie udoskonalonej ważonej reguły decyzyjnej do klasyfikacji niezbalansowanych strumieni danych (OCWE, podrozdział 3.2).
3. Zaproponowanie klasyfikatora z próbkowaniem i akumulacją danych do klasyfikacji niezbalansowanych strumieni danych (DSC, podrozdział 4.1, [28]).
4. Modyfikacja klasyfikatora z próbkowaniem i akumulacją danych (DSC) do ważonego zespołu z wykorzystaniem baggingu do klasyfikacji niezbalansowanych strumieni danych (DSCB, podrozdział 4.2).
5. Modyfikacja klasyfikatora z próbkowaniem i akumulacją danych (DSC) do metody zespołowej z detektorem dryfu koncepcji (DSE, podrozdział 4.3).
6. Zaprojektowanie metody zespołowej do klasyfikacji niezbalansowanych strumieni danych, łączącej technikę próbkowania i akumulacji danych z klasyfikatorami jednoklasowymi (DSE-OC, podrozdział 4.4).

6. Podsumowanie

Podsumowując, stwierdzam że:

1. Teza rozprawy została udowodniona.

2. Rozprawa stanowi oryginalne rozwiązanie problemu naukowego w zakresie tematyki klasyfikacji niezbalansowanych strumieni danych, na co składają się osiągnięcia Autora wymienione w rozdziale piątym niniejszej recenzji, w punktach od 1 do 6. Osiągnięcia wymienione w punktach 1 i 3 zostały opublikowane w materiałach prestiżowych międzynarodowych konferencjach (CORE A, 140 punktów według Ministerstwa Nauki i Edukacji), co oznacza pozytywną weryfikację osiągnięć naukowych Doktoranta przez specjalistów zajmujących się problematyką klasyfikacji niezbalansowanych strumieni danych.

3. Analiza aktualnego stanu wiedzy w zakresie problematyki rozprawy została przeprowadzona na podstawie literatury w sposób wyczerpujący, świadczący o dobrej znajomości tematyki. Doktorant jest współautorem ośmiu publikacji z zakresu problematyki rozprawy, przy czym w sześciu z nich jest pierwszym Autorem, co dowodzi, że ma już osiągnięcia naukowe w tym zakresie. Sześć z wymienionych artykułów opublikowano w materiałach międzynarodowych konferencji, a jeden w czasopiśmie Journal of Computational Sciences.

4. Rozprawa dowodzi, że Doktorant posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja.

5. Rozprawa dowodzi również, że Doktorant wykazał się umiejętnością samodzielnego prowadzenia pracy naukowej.

Na podstawie punktów 1, 2, 3, 4 i 5 podsumowania niniejszej recenzji stwierdzam, że recenzowana praca doktorska spełnia ustawowe wymagania stawiane rozprawom doktorskim. W związku z tym wnioskuję o przyjęcie rozprawy doktorskiej i dopuszczenie Pana mgra inż. Jakuba Klikowskiego do publicznej jej obrony w dziedzinie nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

Marek Skomorowski