dr hab. inż. Rafał Doroz
Instytut Informatyki, Wydział Nauk Ścisłych i Technicznych
Uniwersytet Śląski
ul. Będzińska 39
41-200 Sosnowiec
rafal.doroz@us.edu.pl

Sosnowiec, 15.02.2022

Recenzja rozprawy doktorskiej

mgr. inż. Jakuba Klikowskiego

Ensemble methods for imbalanced data stream classification

(*Metody zespołowe dla klasyfikacji niezrównoważonych strumieni danych*)

1. Tematyka rozprawy

Klasyfikacja strumieni danych jest obecnie jednym z dynamicznie rozwijających się obszarów uczenia maszynowego. Opracowując systemy dedykowane tego typu zagadnieniom musimy mieć świadomość istnienia pewnych problemów im towarzyszących m.in. niezbalansowania danych (liczność jednej klasy znacząco przewyższa liczność drugiej) oraz występowania zmian w charakterystyce analizowanych danych strumieniowych.

Podstawowym celem Autora rozprawy doktorskiej jest opracowanie nowych metod klasyfikacji niezbalansowanych strumieni danych. Klasyfikacja prowadzona będzie w oparciu o zespoły klasyfikatorów jednoklasowych.

2. Ocena treści rozprawy i wkładu oryginalnego

Recenzowana rozprawa jest obszerna: składa się z 5 rozdziałów, bibliografii, wykazu skrótów, symboli i liczy w sumie 200 stron. Praca została napisana w języku angielskim. Rozprawę, z nielicznymi wyjątkami, czyta się dobrze. Strukturę rozprawy oceniam pozytywnie, a wprowadzony podział na rozdziały wydaje się przemyślany. Pewne wątpliwości może budzić powielanie opisu pewnych elementów metod w kolejnych podrozdziałach. Jak rozumiem wynika to

z konwencji pracy jaką przyjął Doktorant, polegającej na opisywaniu kolejnych metod w sposób oddzielny a nie jako modyfikacji wcześniej opisywanych metod. Poniżej omówię zawartość poszczególnych rozdziałów, starając się przeprowadzić ich merytoryczną ocenę.

Rozdział pierwszy jest typowym wprowadzeniem w tematykę rozprawy z uwzględnieniem celu rozprawy i jego naukowego uzasadnienia. Autor rozpoczyna rozdział przywołując cytaty z prac Stanisława Lema oraz Alana Turinga, które jak sam pisze były dla niego inspiracją do zainteresowania się obszarem uczenia maszynowego. Wspomina o problemie niezrównoważenia danych oraz zjawisku dryftu modelu (*concept drift*), które występują w wielu obszarach sztucznej inteligencji. Doktorant słusznie zauważa, że powyższe problemy mogą w pewnych przypadkach prowadzić do całkowitego zaniku obiektów należących do jednej z klas. W celu pokonania tych niekorzystnych zjawisk Autor swoje zainteresowanie kieruje w stronę opracowanej przez siebie metody wykorzystującej strategię klasyfikacji jednoklasowej. Niekorzystne zjawiska niezrównoważenia klas nie będą łagodzone przez sztuczne zwiększanie liczności klasy mniejszościowej, co jest rozwiązaniem klasycznym, a poprzez metody wstępnego przetwarzania danych oparte na akumulacji danych. Doktorant podkreśla, że takie podejście jest także znane, jednak za cel stawia sobie opracowanie klasyfikatora, który będzie konkurencyjny w stosunku do obecnie stosowanych metod.

Na końcu rozdziału opisana została struktura kolejnych rozdziałów rozprawy doktorskiej oraz pojawia się jej cel:

One may design ensemble methods that use data sampling techniques and one-class classifiers, which can outperform state-of-the-art ensemble algorithms for imbalanced data streams classification.

(*Możliwe jest zaprojektowanie metod zespołowych, wykorzystujących techniki próbkowania danych i klasyfikatory jednoklasowe, które mogą przewyższać najlepsze algorytmy zespołowe do klasyfikacji niezrównoważonych strumieni danych.*)

Doktorant wskazuje także cele badawcze, które pozwolą zweryfikować postawiony cel rozprawy. Tak postawiony cel oceniam jako zasadny i ciekawy z naukowego punktu widzenia.

Rozdział drugi stanowi wprowadzenie do tematyki uczenia maszynowego. Doktorant przybliża definicje istotnych dla pracy zagadnień takich jak klasyfikatory jednoklasowe,

uczenie zespołowe, niezrównoważenie danych, strumienie danych oraz szumy odnoszące się zarówno do etykiet jak i atrybutów danych. Wiele uwagi w tym rozdziale poświęcono na kluczową dla rozprawy kwestię niezbalansowania danych strumieniowych, wprowadzając odpowiedni aparat formalny oraz miary jakości klasyfikacji. Definicje zostały przedstawione w sposób formalny i nie budzący wątpliwości.

Kolejne dwa rozdziały (3-4) stanowią oryginalny wkład Doktoranta opisujący zarówno badania jak i eksperymenty potwierdzające słuszność postawionych celów rozprawy. Rozdział trzeci jest opisem dwóch metod klasyfikacji niezbalansowanych strumieni danych z wykorzystaniem klasyfikatorów jednoklasowych. W celu podniesienia skuteczności klasyfikacji, zastosowano podejście bazujące na zespołach klasyfikatorów, które lepiej radzą sobie z trudnymi danymi. Model klasyfikatora zespołowego wymaga odpowiedniego podziału zbiorów uczących. W rezultacie każdy model jest trenowany na innym zestawie danych. Doktorant wykorzystał do podziału danych metodę klasteryzacji danych opartą na metodzie K-means. Zaproponował sposób wyznaczenia liczby klastrów metodą Silhouette Value (SV). Dane po podziale są wykorzystywane do trenowania komitetu klasyfikatorów. Ta jak i każda kolejna zawarta w rozprawie metoda jest opisywana i wyjaśniona w formie algorytmów. Pierwsza opracowana metoda nosi nazwę *Ensemble for Imbalanced data Stream* (*OCEIS*).

Rozwinięciem tej metody jest metoda nazwana *One Class support vector machine Weighted Ensemble* (*OCWE*), w której autor wprowadził do modelu cztery czynniki uwzględniające: kompetencje klasyfikatora, jakość klasyfikacji, wiek próbki oraz współczynnik niezrównoważenia klas. Każdy z tych czynników jest brany pod uwagę z wagą, której wartość wyznaczana jest eksperymentalnie.

Opracowane, autorskie metody zostały zweryfikowane eksperymentalnie, przy czym w przypadku metody drugiej liczba eksperymentów jest zdecydowanie większa.

W rozdziale czwartym Doktorant opisuje kolejne cztery metody klasyfikacji danych niezbalansowanych, w których tym razem zastosował wstępne przetwarzanie danych. W metodzie *Deterministic Sampling Classifier* (*DSC*) Doktorant założył, że przechowywanie do dalszej analizy całych bloków danych nie wpływa na redukcję poziomu ich niezbalansowania. Dlatego dane gromadzone z poprzednich bloków są poddane procesowi podpróbkowania, co pozwala rozwiązać problem niezrównoważonych danych, i dopiero wtedy mogą być użyte do wzmocnienia kolejnych bloków danych.

Druga metoda nazwana *The Deterministic Sampling Classifier with weighted Bagging* (*DSCB*) jest modyfikacją algorytmu *DSC* i polega na dodaniu zespołu klasyfikatorów w oparciu o metodę bagging. Ponieważ próbki pochodzące z poprzednich bloków strumieni danych są zapisywane w różnym czasie, Autor rozprawy opracował formułę, pozwalającą na wyznaczenie wieku każdej próbki. Ta informacja jest następnie wykorzystywana podczas podziału danych na podzbiory uczące.

Kolejnym rozszerzeniem klasyfikatora *DSC* wykorzystującego pojedynczy klasyfikator jest *Deterministic Sampling Ensemble* (*DSE*), gdzie zastosowano zespół klasyfikatorów z głosowaniem większościowym. Wykorzystano tutaj detektor dryftu, który za pomocą mechanizmu zapominania resetuje najstarsze modele i najstarsze zgromadzone dane. Zwieńczeniem prac jest metoda o nazwie *The Deterministic Sampling Ensemble of One Class Support Vector Machine classifiers* (*DSE-OC*). Klasyfikator ten łączy dwa opracowane wcześniej propozycje: klasyfikator *DSE* wykorzystujący dane z poprzednich bloków do dalszych etapów uczenia oraz klasyfikator *OCEIS* wykorzystujący zespoły klasyfikatorów do klasyfikacji niezbalansowanego strumienia danych. Podobnie jak w rozdziale trzecim przeprowadzono dogłębną analizę eksperymentalną popartą statystyczną walidacją uzyskanych wyników. Eksperymenty przeprowadzono na repozytorium 24 sztucznych strumieni oraz 30 strumieni rzeczywistych, pochodzących z dostępnych baz danych. W mojej opinii rozdziały trzeci i czwarty potwierdzają, że Autor posiada wiedzę i umiejętności uprawniające do konstruowania klasyfikatorów oraz modelowania ich zachowań w praktyce.

Rozdział piąty stanowi podsumowanie rozprawy. Autor przedstawił tutaj najważniejsze osiągnięcia i krótko nakreślił kierunek przyszłych badań.

3. Ocena oryginalności rozprawy

Do najważniejszych osiągnięć naukowych Doktoranta Pana mgra inż. Jakuba Klikowskiego, należy zaliczyć zaproponowanie szeregu nowych lub zmodyfikowanych metod klasyfikacji niezbalansowanych strumieni danych. W metodzie *OCEIS* Doktorantowi udało się osiągnąć skuteczność klasyfikacji porównywalną z konkurencyjnymi metodami. Przeprowadzone testy wykazały, że skuteczność jest szczególnie wysoka podczas klasyfikacji danych rzeczywistych, przy jednoczesnym braku skłonności do promowania obiektów z jednej z klas. Dalsze prace nad metodą polegały na wprowadzeniu ważonej reguły decyzyjnej. Zaproponowana modyfikacja poprawiła skuteczności klasyfikacji niezrównoważonych strumieni danych, a eksperymenty wykazały, że metoda bardzo dobrze nadaje się do klasyfikacji silnie niezbalansowanych danych.

W kolejnych czterech metodach Autor zaproponował ciekawe hybrydowe podejścia do klasyfikacji niezrównoważonych strumieni danych, które bazują na wykorzystaniu technik preprocesingu danych. W pierwszej metodzie (*DSC*) zapamiętuje się próbki zarówno z klasy negatywnej jak i pozytywnej pochodzące z poprzednich bloków i wykorzystuje je do nadpróbkowania danych. Dobrym pomysłem okazało się wykorzystanie w metodzie *DSCB* informacji o wieku próbek podczas tworzenia podzbiorów uczących. W rezultacie metoda wykazała się bardzo dobrą skutecznością w porównaniu z innymi reprezentatywnymi metodami. Ostatnia zaproponowana metoda (*DSE-OC*) łącząca metody *DSE* oraz OCEIS również cechuje się wysoką skutecznością w porównaniu z metodami znanymi z literatury.

Lektura doktoratu nasuwa spostrzeżenie o dobrym przygotowaniu Doktoranta do podejmowania wyzwań naukowych związanych z klasyfikacją danych i uczeniem maszynowym. Uznanie budzi sposób projektowania i przeprowadzenia eksperymentów naukowych w celu potwierdzenia słuszności założonych celów i tez badawczych.

Uznaję, że Doktorantowi udało się osiągnąć cel rozprawy, jakim była konstrukcja nowych algorytmów klasyfikacji danych strumieniowych.

4. Uwagi dyskusyjne i polemiczne

Nie kwestionując wartości wyników zawartych w rozprawie, chciałbym zgłosić kilka uwag krytycznych i dyskusyjnych.

1. Jednym z czterech czynników, które zaproponował doktorant w metodzie *OCWE* jest wiek próbki. W rezultacie decyzjom klasyfikatorów pracujących na nowszych danych są przypisywane wyższe wagi. Rodzi się pytanie jak zaproponowane podejście ma się do zjawiska sezonowości danych (data seasonality)? Czy doktorant przeprowadził badania aby to sprawdzić? Proszę też o wyjaśnienie dlaczego wiek próbki w metodzie *OCWE* i *DSCB* obliczany jest według różnych formuł?
2. Duża liczba parametrów potrzebna w proponowanych metodach może być traktowana zarówno jako zaleta jak i wada. Problem doboru wartości wag jest wyraźnie zauważalny podczas testów metod *OCWE* oraz *DSCB*. Czy skuteczność wykorzystanego sposobu wyznaczania wag parametrów poprzez łącznie ich w pary (α; β) oraz (γ; δ) została w jakiś sposób zweryfikowana, tzn. czy wyznaczone w ten sposób wartości odbiegają od wartości wyznaczonych na podstawie wszystkich

kombinacji wag? Czy dobór par paramentów jest przypadkowy? Czy wszystkie parametry są potrzebne?

3. Proszę o wyjaśnienie jak w oparciu o tabele zawierające wartości wag (np. Rysunek 3.10) był wyznaczany zestaw ich optymalnych wartości? Czy decydującym kryterium wyboru była maksymalna wartość w obu tabelach?
4. Każda z zaproponowanych metod została dokładnie przebadana. Brakuje jednak zbiorczej tabeli, która kategoryzowała by ewolucję proponowanych metod. Autor mógł także rozważyć wykonanie testów z wykorzystaniem bazowego klasyfikatora gdzie nie uwzględniono niezrównoważenia danych oraz nie przeprowadzono detekcji dryftu koncepcji.
5. W metodzie *OCWE* liczba klastrów N dla danych z klasy większościowej wyznaczana jest według wzoru (3.3). Sama idea ograniczenia liczby klastrów nie budzi wątpliwości, jednak w praktyce tak wyznaczone liczby klastrów dla kasy większościowej i mniejszościowej mogą się od siebie bardzo różnić. Czy Doktorant przeprowadził jakieś eksperymenty na potrzeby opracowania tej formuły? Proszę także o informację, czy wpływ klasteryzacji był oceniany w kontekście liczby cech?
6. Doktorant modyfikuje dane testowe tak, aby uzyskać zadany dryft koncepcji. Jednocześnie sam przyznaje, że istnieje niewiele danych rzeczywistych, które charakteryzują się danymi zakłóceniami. Tutaj rodzi się pytanie o praktyczne zastosowanie opracowanej metody.
7. W rozprawie znajdziemy dwukrotny opis algorytmu *OCEIS*, pierwszy na stronie 61, a kolejny na stronie 146. Opisywanie tego samego algorytmu wydaje się być niepotrzebne, zwłaszcza, że brakuje nawet krótkiego uzasadnienia tego zabiegu.

5. Charakterystyka działalności naukowej, w tym publikacyjnej, Doktoranta

Do wyraźnie zauważalnej aktywności naukowej Doktoranta należy zaliczyć zestaw ośmiu artykułów opublikowanych zarówno w renomowanych czasopismach naukowych jak i międzynarodowych konferencyjnych. Na szczególną uwagę zasługują publikacje z listy MEiN z wagą 100 oraz 140 punktów. W sześciu z nich Doktorant jest pierwszym autorem. Dwa kolejne artykuły są w recenzji. Chyba jako oznakę skromności można uznać przypisanie przez Doktoranta jednej z nich 100 zamiast 140 punktów. Wspomniane współczynniki bibliometryczne należy uznać za ponadprzeciętne. Wszystkie osiągnięcia publikacyjne świadczą o wysokim poziomie

naukowym Pana mgra inż. Jakuba Klikowskiego, a także całego zespołu naukowego, z którym Doktorant współpracuje.

6. Konkluzja końcowa

Tematyka rozprawy mgra inż. Jakuba Klikowskiego jest istotna w kontekście możliwości automatycznej klasyfikacji danych strumieniowych. Rozprawa zawiera oryginalny i wartościowy wkład w rozwiązanie ważnego problemu naukowego, a uzyskane wyniki empiryczne i eksperymentalne wskazują również na możliwość wykorzystania zaproponowanych metod w praktyce.

Doktorant swobodnie posługuje się aparatem matematycznym oraz dobrze zna metody i algorytmy uczenia maszynowego oraz klasyfikacji danych. Poprawnie opisuje przebieg eksperymentów oraz wykorzystuje metody statystyczne do ich analizy. Wszystko to świadczy o dojrzałości naukowej Pana mgra inż. Jakuba Klikowskiego.

Chciałbym zauważyć, że krytyczne uwagi oraz komentarze mają zachęcić do dyskusji w celu lepszego zrozumienia rozprawy oraz zainspirować do dalszych badań. Uzyskane rezultaty stanowią oryginalny własny wkład Autora rozprawy Pana mgra inż. Jakuba Klikowskiego w rozwój dyscypliny naukowej Informatyka, w szczególności obszarze uczenia maszynowego.

Wobec powyższego stwierdzam, że rozprawa doktorska mgra inż. Jakuba Klikowskigo **spełnia wymogi stawiane przez Prawo o szkolnictwie wyższym i nauce w odniesieniu do rozpraw doktorskich. Wnoszę o dopuszczenie opiniowanej dysertacji do dalszych etapów przewodu doktorskiego, w tym do jej publicznej obrony.**

Biorąc pod uwagę dojrzały charakter recenzowanej pracy oraz fakt, że Autor opublikował 8 prac w czasopismach i na konferencjach o dużej randze naukowej wnioskuję także o jej wyróżnienie.