Częstochowa, dn. 06 lipca 2021 r.

prof. dr hab. inż. Rafał Scherer
Katedra Inteligentnych Systemów Informatycznych
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska
al. Armii Krajowej 36
42-200 Częstochowa

Recenzja

rozprawy doktorskiej mgra inż. Pawła Zyblewskiego, pt.: Classifier selection for imbalanced data stream classification.

Promotor: prof. dr hab. inż. Michał Woźniak

Niniejszą recenzję opracowano na wniosek Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Politechniki Wrocławskiej, która mocą uchwały nr 55/04/RDND03/2021-2024,z dnia 21 kwietnia 2021 roku powołała mnie na recenzenta.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Obecnie mamy do czynienia z wielką ilością danych, szczególnie danych napływających regularnie w postaci na przykład notowań z rynków finansowych, danych gospodarczych, danych o środowisku naturalnym, danych z kosmosu, czy danych z sieci komputerowych. Potrzebne są zatem metody analizujące w sposób przyrostowy takie dane. Ponadto ważne jest aby analiza taka odbywała się w jak najkrótszym czasie, przy ograniczonym zapotrzebowaniu na pamięć oraz z możliwością adaptacji do zmieniającej się charakterystyki danych. Doktorant postanowił rozwiązać wybrane problemy istniejących algorytmów do klasyfikacji danych strumieniowych, formułując tezę pracy „Istnieją metody wykorzystujące zarówno wstępne przetwarzanie danych, jak i metody selekcji klasyfikatorów, które przewyższają jakość predykcji znanych z literatury metod stosowanych w klasyfikacji danych trudnych." Ogólnym celem pracy było opracowanie algorytmów klasyfikacji oraz selekcji klasyfikatorów w zespołach dla danych niezbalansowanych i ograniczonym dostępie do etykiet wraz z implementacją w języku Python.

2. Zawartość rozprawy

Recenzowana praca mgra inż. Pawła Zyblewskiego składa się ze spisu skrótów, symboli, streszczeń, sześciu rozdziałów oraz bibliografii. Dokument liczy 207 stron.

Pierwszy rozdział jest krótkim wprowadzeniem do problemu niezbalansowanych danych i strumieni danych oraz ich współistnienia. Omówiona jest geneza uczenia zespołowego wynikająca z różnorodności, niezależności, decentralizacji i agregacji poszczególnych decyzji. Rzadkość podejmowania jednocześnie badań nad danymi niezbalansowanymi i strumieniowymi była motywacją do podjęcia tematyki recenzowanej pracy doktorskiej. Dalej wymienione jest osiem celów szczegółowych oraz omówienie struktury rozprawy.

Rozdział 2 jest wprowadzeniem do problemu rozpoznawania wzorców (pattern recognition) począwszy od opisu uczenia indukcyjnego, poprzez podstawowe pojęcia aż do wybranych typów klasyfikatorów. Następnie omawiane są zespoły klasyfikatorów: komponenty tworzące zespół, miary poziomu różnorodności, metody wymuszające różnorodność, metody kombinacji odpowiedzi, metody selekcji klasyfikatorów, dynamiczne i statyczne, pruning i wybrane metody z literatury. Dalej omawiane są trudne dane w postaci niezbalansowanej oraz metody ich wstępnego przetwarzania i miary dokładności. W tabeli 2.3 zebrane jest 41 statycznych niezbalansowanych zbiorów danych, a następnie opisano sposób przeprowadzania eksperymentów na statycznych danych. Tu Doktorant przechodzi do opisu strumieni danych, czyli danych pojawiających się sekwencyjnie o mogącej zmieniać się dystrybucji, zmiennej prędkości pojawiania się, często niemożności ich pełnego etykietowania, a analizowane często przy ograniczonych zasobach obliczeniowych. Omówiono metody oceny klasyfikatorów do danych strumieniowych. Dalsza część tego rozdziału dotyczy niezbalansowanych strumieni danych, w tym ze zmieniającym się w czasie stopniem niezbalansowania. Autor pisze o trudności z dostępem do rzeczywistych, wymagających danych strumieniowych. Wynika to na przykład ze zbytniej prostoty dostępnych danych, wstępnego podziału kolejności przychodzących danych na klasy, braku dryfu lub zbyt oczywistego dryfu czy braku niezbalansowania. Autor wyselekcjonował pięć najbardziej odpowiednich zbiorów danych, poddał je procesowi binaryzacji (sprowadził do zagadnienia dwóch klas) i podzielił na zbalansowane sekwencje. Dalej opisane są generatory danych syntetycznych oraz autorski generator trudnych danych strumieniowych. Przedstawiono również metody radzenia sobie z danymi częściowo etykietowanymi. Ostatni podrozdział przedstawia autorski moduł języka Python wykorzystujący API biblioteki scikit-learn służący do generowania trudnych strumieni danych. Pozwala wprowadzać dowolny typ zjawiska concept drift oraz niezbalansowania.

Rozdział 3 prezentuje trzy autorskie metody klasyfikacji niezbalansowanych danych. Klasyfikatory będące członkami zespołu klasyfikatorów są wybierane za pomocą zaproponowanej jednowymiarowej miary różnorodności H.

Pierwszym zaprezentowanym algorytmem jest nazwany bardzo ogólnie Diversity Ensemble Pruning (DEP). Zaproponowane jest wspomniane kryterium H mierzące jak usunięcie danego członka wpływa na różnorodność całego zespołu. Przestrzeń kryterium H jest grupowana za pomocą algorytmu k-środków. Nie jest chyba jasne jak dobrać liczbę grup c. Algorytmowi towarzyszy analiza złożoności obliczeniowej. Eksperymenty wykonane na 41. zbiorach danych porównały zaproponowany meta-algorytm na popularnych algorytmach klasyfikacji kNN, CART i GNB. Wykazały, że zaproponowany algorytm doboru członków zespołu znacząco ulepsza jego działanie oraz może zastąpić wstępne przetwarzanie danych.

Następnie Autor proponuje dwa rozszerzenia algorytmu Multistage Majority Voting Organization (MOMV) z 2002 roku. Oparte są one na wspomnianej już mierze różnorodności *H*. Pierwsza modyfikacja, nazwana Two-step Majority Voting Organization (TSMV) tworzy dwustopniowe głosowanie większością bezwzględną. Druga modyfikacja Random Sampling Multistage Organization (RSMO) wprowadza przed pierwszym krokiem głosowania losowanie ze zwracaniem. Eksperymenty miały na celu sprawdzenie, czy dwuetapowe metody maja lepszą dokładność niż zwykłe zespoły klasyfikatorów lub wstępne przetwarzanie danych. Wyniki eksperymentów pokazały, że obie metody ulepszyły wyniki na danych niezbalansowanych oraz mogą zastąpić wstępne przetwarzanie danych.

Dalej w rozdziale 3 zaproponowane są dwa algorytmy dynamicznej selekcji klasyfikatorów oparte na metodzie Stratified Bagging. Eksperymenty wykazały, że dwie zaproponowane metody selekcji klasyfikatorów uzyskują lepsze rezultaty niż standardowe metody zespołowe działające w połączeniu ze wstępnym przetwarzaniem danych. Szczególnie w przypadku wersji z ważoną odległością Euklidesową można było zaobserwować lepsze działanie na niezbalansowanych zbiorach.

Rozdział 4 dotyczy metod klasyfikacji niezbalansowanych strumieni danych. Pierwszym algorytmem jest algorytm Minority Driven Ensemble. Algorytm działa na silnie niezbalansowanych danych, w których występuje zjawisko concept drift, cechując się przy tym małą złożonością obliczeniową. Zespół składa się ze stałej liczby klasyfikatorów *k*NN, w każdym kroku ten z najmniejszą wartością Balanced Accuracy Score jest usuwany, podobnie klasyfikatory z wartością BAC poniżej pewnego progu są usuwane. Klasyfikatory trenowane są danymi pozbawionymi obiektów odstających. Eksperymenty potwierdziły, że zaproponowany algorytm działa na wysoce niezbalansowanych danych lepiej od klasycznych metod dynamicznej selekcji zespołów oraz że jest odporny ma zaszumienie etykiet i zjawisko concept drift.

Następnym zaproponowanym algorytmem jest Dynamic Ensemble Selection for Imbalanced Stream Classification (DESISC). Nowy klasyfikator jest dodawany dla każdego bloku danych, aż do osiągnięcia maksymalnej założonej ich liczby. Członkowie zespołu są oceniani miarą BAC i najgorszy jest usuwany. Jednocześnie, klasyfikatory z miarą BAC mniejszą niż założony próg są usuwane. Opisowi metody towarzyszy analiza złożoności obliczeniowej. Celem eksperymentów był dobór doświadczalny rozmiaru zespołu oraz współczynnika odpowiedzialnego za usuwanie najgorszych członków. Ponadto metoda została oczywiście porównana z innymi algorytmami selekcji klasyfikatorów i okazała się lepsza niż metody z literatury, wykazując odporność na niezbalansowanie zbioru danych i pojawianie się zjawiska concept drift.

W następnej kolejności, poprzednio omówiona metoda Dynamic Ensemble Selection została rozszerzona o dobór klasyfikatora dla każdego z bloku danych za pomocą strategii bagging. Eksperymenty pokazały, że metoda daje lepszą dokładność niż standardowe metody dynamicznej selekcji klasyfikatorów.

W rozdziale 5 Doktorant zajmuje się problemem braku pełnej informacji o etykietach obiektów w strumieniach danych. Jednym z obszarów zajmujących się takimi danymi jest tzw. aktywne uczenie. Zaproponowany algorytm Budget Active Labeling Strategy (BALS) zakłada, że algorytm otrzymuje poza danym oetykietowanym obiektem, pewną liczbę losowych obiektów z bieżącego bloku danych. Algorytm składa się z trzech komponentów. Pierwszym jest Budget Labeling, gdzie klasyfikator jest uczony dodatkowo małą liczbą losowych obiektów z bieżącego bloku strumienia. Drugim jest moduł aktywnego uczenia ALS. Trzecim jest właściwy algorytm BALS, łączący dwa pierwsze w jedną strategię zwiększającą zdolność do uogólniania algorytmu klasyfikacji przez dodatkową dywersyfikację obiektów mających być ocenianymi przez eksperta w aktywnym uczeniu. Została przeanalizowana złożoność obliczeniowa poszczególnych składników metody. Eksperymenty zostały przeprowadzone na trzech typach syntetycznych strumieni z nagłym, wzrastającym i stopniowym dryfem koncepcji w wersjach powtarzających i nie oraz strumieniach niezbalansowanych. Ponadto metoda została przebadana na pięciu rzeczywistych strumieniach danych.

Rozdział 6 jest podsumowaniem rozprawy, w którym zebrano w jednym miejscu wnioski kończące opisy poszczególnych autorskich metod i eksperymentów. Doktorant podał również wiele pomysłów, które wytyczają wiele kierunków badawczych, np. nowe sposoby obliczania miary *H*, typy głosowania, metody ważenia klasyfikatorów, nowe metody kombinowana odpowiedzi, itp. Doktorant ma również plany na rozbudowę autorskiego oprogramowania. W rozdziale zestawione są również publikacje Autora.

Pracę kończy bibliografia składająca się z 284 aktualnych pozycji.

Ogólnie, zasadnicze i oryginalne rezultaty pracy można podsumować następująco:

- Opracowanie wprowadzenia do tematyki i przeglądu literatury klasyfikacji niezbalansowanych strumieni danych.
- Stworzenie metod selekcji członków zespołu klasyfikatorów dla danych niezbalansowanych,
- Opracowanie algorytmów klasyfikacji wysoce niezbalansowanych strumieni danych,
- Stworzenie hybrydowej metody łączącej Dynamiczną Selekcję Klasyfikatorów z przetwarzaniem wstępnym danych działającej na niezbalansowanych danych strumieniowych,
- Opracowanie metod klasyfikacji strumieni danych z ograniczonym dostępem do etykiet, wraz z zastosowaniem uczenia aktywnego,
- Opracowanie autorskiego oprogramowania w języku Python, egzystującego w otoczeniu istniejącego ekosystemu bibliotek analizy danych i uczenia maszynowego,
- Przeprowadzenie bardzo dobrze zaplanowanych eksperymentów.

Mgr Zyblewski opublikował ponadprzeciętną liczbę prac naukowych: jedną w czasopiśmie za 200 punktów, pięć w czasopismach i konferencjach za 140 punktów, jedną w czasopiśmie za 70 punktów, trzy prace w materiałach konferencji za 20 punktów. Zaprezentowany materiał pokazuje, że Doktorant zrealizował cel pracy.

3. Uwagi krytyczne i wskazówki dotyczące rozprawy

Praca napisana jest schludnie i przejrzyście. Praca obfituje w czytelne rysunki oraz schematy. Eksperymenty są przemyślane i przeprowadzone w standardach istniejących w najlepszych publikacjach z tej tematyki. Ponadto na uwagę zasługuje użycie języka angielskiego na bardzo dobrym poziomie. Poniżej zamieszczam kilka pytań, które zrodziły się w czasie czytania pracy:

1. Czy algorytm Diversity Ensemble Pruning (DEP) nie jest nazwany w sposób zbyt ogólny? W podobny sposób, np. Diversity-based Ensemble Pruning można by nazwać całą rodzinę metod istniejących w literaturze.
2. Jak dobrać w powyższym algorytmie liczbę grup c?
3. Jak dobrać liczbę klasyfikatorów w metodzie Dynamic Ensemble Selection for Imbalanced Stream Classification (DESISC)?

Poniżej wymieniam kilka przykładów znalezionych możliwych błędów językowych i edytorskich nieistotnych oczywiście w odbiorze merytorycznym pracy:

str. 64: … used for clustering in diversity space – brakuje przedimka określonego.

str. 62: Diversity based one-dimensional clustering – brakuje myślnika pomiędzy dwoma pierwszymi słowami,

str. 75: Example of random sampling multistage organization is shown in Figure 3.5 and the pseudocode for its prediction process is presented in Algorithm 3. – chyba powinno być napisane Algorithm 4,

str. 100: Additionally, at each step all models with BAC are lower than 0.5 + α, where α is the algorithm's parameter responsible for the outdated models removing rate, are removed from Π. – najprawdopodobniej niepotrzebne słowo "are" po BAC,

str. 123: Akapit rozpoczyna się od „n general, the order …" – brak litery „I",

str. 152: "… framework may outperform both bath and online-based …" – "batch" zamiast "bath"

str. 153: "This chapter introduces the new method for…" – może lepiej wyglądałoby "a new method",

str. 155: "… the decision boundary doe not exceed …" – brak "s" w "does"

str. 166: complexity of O(log n. – brak nawiasu zamykającego.

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że Pan mgr inż. Paweł Zyblewski w rozprawie doktorskiej „Classifier selection for imbalanced data stream classification":

- Zrealizował cel rozprawy,

- Opracował wprowadzenie do tematyki i dokonał przeglądu literatury klasyfikacji niezbalansowanych strumieni danych.
- Stworzył rodzinę metod selekcji członków zespołu klasyfikatorów dla niezbalansowanych danych statycznych i strumieniowych,
- Opracował algorytmy klasyfikacji wysoce niezbalansowanych strumieni danych,
- Zaprojektował metody klasyfikacji strumieni danych z ograniczonym dostępem do etykiet, wraz z zastosowaniem uczenia aktywnego,
- Stworzył autorskie oprogramowanie w języku Python, egzystujące w otoczeniu istniejącego ekosystemu bibliotek analizy danych i uczenia maszynowego,
- Przeprowadził bardzo dobrze zaplanowanych eksperymentów,
- Wykazał się umiejętnością samodzielnej pracy badawczej oraz pracy w zespole międzynarodowym, znajomością literatury światowej i wiedzą w zakresie uczenia maszynowego, szczególnie metod zespołowych,
- Zadbał o popularyzację wyników swoich badań w wysokopunktowanych wydawnictwach.

Recenzowana praca spełnia wymagania ustawy o tytule i stopniach naukowych w dyscyplinie naukowej Informatyka Techniczna i Telekomunikacja. Wnoszę o jej przyjęcie i dopuszczenie do dalszych etapów postepowania doktorskiego. Jednocześnie ze względu na wysoki poziom naukowy rozprawy, wnioskuję o wyróżnienie rozprawy.