Warszawa, 15.05.2015

Prof. Dr hab. Henryk Rybinski
Instytut Informatyki Politechniki Warszawskiej
hrb@ii.pw.edu.pl

Recenzja rozprawy doktorskiej
Dariusza Brzezinskiego
p/t. „Block-Based And Online Ensembles For
Concept-Drifting Data Streams

Dane ogólne

Przedstawiona do recenzji rozprawa jest napisana w języku angielskim. Składa się z siedmiu rozdziałów, aneksów (a) zawierającego skrypty eksperymentów i (b) listy publikacji autora liczącej 6 pozycji, ze wskazaniem rozdziału rozprawy, a także bibliografii (175 pozycje). Objętość rozprawy - 154 strony.

1. Problem badawczy i jego znaczenie

1.1 Problem rozważany w rozprawie

Tematyka rozprawy ściśle wiąże się z dziedziną eksploracji danych. Dziedzina ta jest od dawna w obszarze zainteresowań statystyków. Również w dziedzinie sztucznej inteligencji zagadnienie to jest od dawna przedmiotem wielu badań, w szczególności w kontekście akwizycji wiedzy, uczenia się, a także metod wnioskowania indukcyjnego przez analizę przykładów. Od ponad 20-tu lat trwa gwałtowny rozwój badań nad nowymi metodami eksploracji danych, głównie w zastosowaniach do baz danych. Klasyczne już dzisiaj zadania eksploracji danych, takie jak odkrywanie reguł asocjacyjnych, grupowanie czy klasyfikacja, dotyczą przede wszystkim statycznych zbiorów danych. Badania nad eksploracją danych strumieniowych dotyczą odkrywania wiedzy związanej z procesami - wiąże się to z dynamiką zbiorów danych. Tematyka ta pojawia się już w badaniach pod koniec ubiegłego stulecia. Jednak podstawowym ograniczeniem badań strumieniowych przez długi okres wydają się być przede wszystkim możliwości techniczne sprzętu – głównie ograniczenia pamięci i szybkość przetwarzania. Z rozwojem technologii obliczeniowych badania w zakresie eksploracji strumienia danych nabierają w ostatnich kilku latach istotnego znaczenia. Tematyka rozprawy wiąże się z tym właśnie obszarem badań. Dokładniej, rozprawa dotyczy badań nad możliwościami powiązania metod blokowych eksploracji strumienia danych z metodami *on-*

line. Główną motywacją doktoranta jest fakt, że w literaturze przedmiotu klasy algorytmów bazujących na metodach blokowych oraz algorytmów *on-line* rozważane są oddzielnie, natomiast często konieczne jest powiązanie tych metod, aby możliwe było odkrywanie wiedzy w zakresie charakterystyk strumienia danych, zarówno z punktu widzenia metod blokowych (dających wiedzę o zmianach strumienia w dłuższych okresach czasu), jak też metod *on line* (dających wiedzę o tendencjach zmian w strumieniu danych w krótkich odcinkach czasu).

Celem, jaki postawił sobie doktorant, jest opracowanie nowej metody uczenia przyrostowego na potrzeby analizy strumienia danych, w którym mogą występować różnego rodzaju tendencje zmian własności (autor używa tu pojęcia stosowanego w literaturze anglosaskiej, mianowicie *concept drift*). Zasadniczą hipotezą, jaką stawia we wstępie autor jest przyjęcie, że nowa metoda może pracować zarówno blokowo, jak też *on-line*. Badania autora koncentrują się na skonstruowaniu takiej właśnie metody oraz na zweryfikowaniu jej parametrów w porównaniu z metodami istniejącymi pod kątem dokładności działania, wykorzystania pamięci oraz czasów przetwarzania.

Zagadnienie uczenia przyrostowego ma ogromne znaczenie w dziedzinie eksploracji danych. Jest ono przedmiotem badań w wielu ośrodkach. Autor dobrze określił swoje cele badawcze, motywując ich wagę oraz analizując możliwości ich praktycznego zastosowania. Motywując swoje badania autor odniósł się do rozwiązań znanych z literatury (wstępnie w p.1.1, dokładniej w rozdziale 2). Tematyka rozważana przez doktoranta jest ważna z praktycznego, jak też teoretycznego punktu widzenia i jest godna rozprawy doktorskiej.

1.2 Charakter rozprawy i znaczenie praktyczne badań

Praca dowodzi solidnego warsztatu teoretycznego doktoranta, jednocześnie badania autora są wsparte implementacjami oraz eksperymentami wskazującymi na

1. równie solidny warsztat doktoranta w zakresie metod i narzędzi eksploracji danych
2. możliwości praktycznego zastosowania opracowanych algorytmów.

W szczególności możliwości praktycznego wykorzystania są przekonywująco pokazane w rozdziałach 5 i 6 – część eksperymentów była prowadzona na danych rzeczywistych.

2. Wkład autora

Wkład autora jest imponujący – obejmuje on szereg ważnych elementów związanych z uczeniem maszynowym na potrzeby analizy strumienia danych. Wyróżnić tu można trzy zasadnicze problemy, którymi autor się zajmuje, i dla których proponuje interesujące rozwiązanie:

1. Uczenie przyrostowe bazujące na blokowej analizie danych;
2. Związki pomiędzy zbiorami klasyfikatorów blokowych i „*on-line*";
3. Analiza strumienia danych ze zmiennym w czasie rozkładem.

Ad. 1

Ważnym osiągnięciem autora jest wprowadzenie elementów uczenia przyrostowego w zespołach klasyfikatorów bazujących na blokowej analizie danych (*ensembles*). W szczególności autor rozwija nowy algorytm AUE (*Accuracy Updated Ensemble*), który charakteryzuje się wyższą efektywnością predykcji (uśrednionej) w stosunku do znanych algorytmów uczenia adaptacyjnego w zastosowaniu do strumieni danych charakteryzujących się różnorodną zmiennością (od gwałtownej poprzez stopniową, do powtarzalnej). NB, ta część rozprawy była publikowana przez autora i promotora w roku 2011 w materiałach konferencyjnych (LNCS) konferencji *Hybrid Artificial Intelligence Systems* oraz w roku 2014 w prestiżowym czasopiśmie *IEEE Transactions on Neural Networks and Learning Systems*. Potwierdzeniem wagi osiągnięcia doktoranta dla dziedziny eksploracji danych strumieniowych jest cytowalność obu prac, odpowiednio 38 (praca z roku 2011) i 32 (praca z roku 2014).

Ad. 2

Równie istotnym osiągnięciem jest zbadanie związków pomiędzy zespołami klasyfikatorów działającymi blokowo i *on-line*. Problemowi temu są poświęcone rozdziały 4 i 5. Autor zbadał, czy możliwe jest przekształcenie zespołu klasyfikatorów działających blokowo w zespół klasyfikatorów *on-line*. W oparciu o tę analizę autor wprowadza funkcję, która oblicza składowe klasyfikatory w trybie *on-line* w miarę pojawiania się w strumieniu nowych przykładów. Funkcja stanowi podstawę działania nowego algorytmu *Online Accuracy Updated Ensemble* (OAUE). I w tym przypadku wagę uzyskanych wyników potwierdzają publikacje – jedna w czasopiśmie *Information Sciences*, druga w materiałach prestiżowej

konferencji ECML PKDD.

Ad. 3
Autor pokazał, że w przypadku strumienia z ewoluującymi własnościami i ze zmiennym niezrównoważonym rozkładem klas klasyczne metody oceny klasyfikatorów są nieefektywne. W takich przypadkach metryka AUC sekwencyjnych predykcji (*prequential AUC*) zaproponowana przez autora jest właściwsza dla oceny klasyfikatorów.

3. Poprawność

Wysoko oceniam przyjętą przez autora metodologię badań. Autor dokonuje krytycznej analizy stanu badań w dziedzinie analizy danych strumieniowych, w oparciu o tę analizę proponuje szereg rozwiązań a następnie przeprowadza badania eksperymentalne. Badania te stanowią istotny element rozprawy. Autor na danych rzeczywistych i syntetycznych bada skonstruowane algorytmy i miary, porównując je z propozycjami z literatury. Uzyskane eksperymentalnie wyniki potwierdzają słuszność proponowanej koncepcji analizy danych strumieniowych.

Na wysoką ocenę zasługuje też warsztat autora. Do przeprowadzenia eksperymentów autor wykorzystuje otwarte oprogramowanie Massive On-line Analysis (MOA) i implementuje własne algorytmy, częściowo opisane w Dodatku A rozprawy i w całości dostępne na stronie personalnej autora. Dzięki temu zabiegowi jest możliwość porównania opracowanych metod z metodami znanymi z literatury. Zawarta w pracy dyskusja wyników jest przeprowadzona w sposób jasny i czytelny.

4. Wiedza kandydata

W rozprawie doktorant wykazuje się dużą wiedzą, przede wszystkim w zakresie uczenia maszynowego i eksploracji danych, ale także w dziedzinach sztucznej inteligencji, struktur danych i algorytmiki. Przegląd dziedziny jest zaprezentowany przede wszystkim w rozdziale 2, ale dyskusje literaturowe są prowadzone także w pozostałych rozdziałach.

Autor odwołuje się w pracy do 175 pozycji literaturowych, w tym do pozycji klasycznych (np. [7] z roku 1972), jak też najnowszych prac odnoszących się do wyników w dziedzinie eksploracji danych. Przedstawiona literatura jest kompletna.

5. Inne uwagi

Praca jest napisana starannie i dobrym językiem angielskim (kilka drobnych poprawek zaznaczyłem w tekście). Układ pracy jest logiczny i podporządkowany tezie rozprawy. Zamieszczono stosowaną w rozprawie notację oraz wykaz akronimów.

Moje krytyczne uwagi przedstawiam poniżej. Mają one charakter uwag redakcyjnych i nie zmieniają one mojej wysokiej ocenę rozprawy.

1. Przyjęta w rozprawie hipoteza badawcza jest sformułowana w sposób sztuczny (to często bywa bolączką nawet dobrych doktoratów).

   Zamiast pisać, że
   „*...novel methods for... can be proposed..*",

   należałoby stwierdzić, że
   „*...the proposed methods for... are* [tu lista własności nowych metod, np. *...better suited to analyzing stream data...*]"

2. Jest szereg definicji wprowadzonych w rozdziałach 2 i 6, dotyczących spraw raczej oczywistych, natomiast brakuje nawet pobieżnego wyjaśnienia niektórych zaawansowanych pojęć – w szczególności używana terminologia bywa dosyć hermetyczna, jasna tylko dla znawców siedzących głęboko w dziedzinie. Przykładem są terminy „*prequential*", „krzywa *ROC*".

3. Dla Twierdzenie 2.1 nie ma w pracy formalnego dowodu. Brakuje referencji do publikacji, gdzie ten dowód się znajduje.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami) moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego? (wybierz jedną opcję stawiając znak X)

Zdecydowanie TAK	Raczej TAK	Trudno powiedzieć	Raczej NIE	Zdecydowanie NIE
X				

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka lub Automatyka i Robotyka?

Zdecydowanie TAK	Raczej TAK	Trudno powiedzieć	Raczej NIE	Zdecydowanie NIE
X				

C. Czy kandydat umiejętność samodzielnego prowadzenia pracy naukowej?

Zdecydowanie TAK	Raczej TAK	Trudno powiedzieć	Raczej NIE	Zdecydowanie NIE
X				

Ponadto, biorąc pod uwagę wysoki poziom naukowy rozprawy i osiągnięte wyniki rekomenduję wyróżnienie rozprawy doktorskiej.