Dr hab. inż. Krzysztof Krawiec, prof. nadzw.
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2
60-965 Poznań

Poznań, 14.01.2017

Recenzja rozprawy doktorskiej mgr inż. Juliana Zubka pt. "Metody integracji wieloskalowej informacji w sztucznych systemach uczących się"

1 Tematyka rozprawy

Rozprawa mgr Zubka dotyczy oceny złożoności zadań uczenia maszynowego, w szczególności w powiązaniu z klasyfikatorami złożonymi i klasyfikacją wieloskalową. W uczeniu maszynowym punktem wyjścia dla rozważań teoretycznych i praktyki są dane uczące, zazwyczaj stanowiące pewną próbę w wielowymiarowej przestrzeni kartezjańskiej. Charakterystyka tej próby, w szczególności jej rozkład, determinuje w nietrywialny sposób skuteczność procesu uczenia. Mgr Zubek poświęcił swoją rozprawę metodom opisu trudności takich danych, pokazując jak otrzymywane w ten sposób charakterystyki mogą wspomagać i usprawniać proces uczenia. Przyczynki te osadził w zagadnieniach wieloskalowych, gdzie dane uczące reprezentowane są na wielu poziomach abstrakcji.

Tematyka rozprawy umiejscawia ją jednoznacznie w obszarze uczenia maszynowego (ang. machine learning), jednego z najprężniej rozwijających się obecnie działów informatyki, lokowanego na przecięciu ze statystyką, i silnie powiązanego ze sztuczną inteligencją (*artificial intelligence*) i inteligencją obliczeniową (*computational intelligence*), także powszechnie traktowanych jako działy dyscypliny informatyka. Wątki podjęte przez Autora w rozprawie uważam za aktualne i istotne zarówno z teoretycznego jak i praktycznego punktu widzenia.

2 Ocena wkładu oryginalnego

Za główne oryginalne przyczynki Autora uważam następujące elementy pracy:

1. Autorską, interesującą perspektywę na modele hierarchiczne w uczeniu maszynowym (Rozdział I) oraz koncepcję/interpretację pojęcia klasyfikatora wieloskalowego (Rozdział II).

2. Propozycję czytelnej i spójnej definicji i notacji (w tym notacji graficznej) procesów transformacji danych, agregacji danych, uczenia i testowania klasyfikatorów, w tym klasyfikatorów złożonych (Rozdział II), oraz prezentację w tej notacji wybranych istotnych schematów konstrukcji klasyfikatorów, w szczególności klasyfikatorów złożonych typu bagging, boosting i stacked generalization, a także wybranych architektur sztucznych sieci neuronowych oraz schematów wieloskalowych.

3. Oryginalną metodę 'bezmodelowej' analizy i ilościowej oceny złożoności danych uczących, w tym miarę opartą na odległości Hellingera, w tym jej warunkową odmianę uwzględniającą zmienną zależną.

4. Wnikliwą analizę teoretyczną i empiryczną wyżej wymienionych miar, na problemach syntetycznych i rzeczywistych, ilustrującą jej czułość na obecność nadmiarowych atrybutów, stopień zależności pomiędzy atrybutami, obecność obserwacji odstających, oraz porównanie ich z innymi miarami złożoności.

5. Propozycje zastosowań proponowanych miar do przewidywania skuteczności klasyfikatorów i przycinania (redukcji) rozmiaru danych uczących.

6. Przekonujące wyniki uzyskane przez proponowane algorytmy na problemach testowych (benchmarkach).

7. Przyczynki do metodyki testowania klasyfikatorów wieloskalowych, w tym przemyślane algorytmy losowania przykładów negatywnych oraz przekonujący wynik eksperymentalny (Rozdział IV)

8. Interesujące studium przypadku zastosowania opracowanej przez Autora metodyki w wieloskalowym modelowaniu i klasyfikacji oddziaływań białek na dwóch poziomach abstrakcji oraz szerszą wizję zastosowań podejść wieloskalowych do takich problemów, z wyodrębnieniem wielu skal przestrzennych i czasowych (początek Rozdziału V), solidnym porównaniem z metodami referencyjnymi (poprzedzonym starannym strojeniem metody), prowadzącą do bardzo dobrych i interesujących (także z praktycznego punktu widzenia) wyników.

Przyczynki te pozwalają mi stwierdzić iż rozprawa mgr inż. Zubka wnosi znaczący i oryginalny wkład w rozwój teorii i praktyki uczenia maszynowego, a zatem także do informatyki i powiązanych dyscyplin.

3 Ocena treści rozprawy

Podejścia opisywane w pracy są dobrze umotywowane zarówno teoretycznie jak i praktycznie. W szczególności w odniesieniu do podejść wieloskalowych, wiele konwencjonal-

nych metod uczenia maszynowego zakłada jednopoziomową, płaską (zazwyczaj tabelaryczną) reprezentację danych. Jednak problemy napotykane obecnie w praktyce analizy danych i uczenia maszynowego coraz częściej odbiegają od tego schematu, np. oferując dane jednocześnie na wielu poziomach abstrakcji, np. związanej z rozdzielczością przestrzenną (w rozpoznawaniu obrazów) czy czasową (w analizie przebiegów czasowych), ale nie tylko, tj. także wielopoziomową w bardziej abstrakcyjnym sensie, tak jak w przedstawianych przez Autora problemach wieloskalowego modelowania oddziaływań białkowych odnośnie sekwencji, struktury drugorzędowej i trzeciorzędowej. Efektywne rozwiązywanie takich złożonych problemów wymaga nowego aparatu pojęciowego i metodyki. Rozprawa mgr inż. Zubka jest interesującą i oryginalną propozycją w tym obszarze.

Poza wyżej wymienionym głównym przyczynkiem do problemów wieloskalowych, praca zawiera wiele innych komponentów których obecność szczególnie doceniam, w szczególności oryginalną metodę oceny złożoności problemu, nietrywialną metodykę oceny klasyfikacji wieloskalowej, czy propozycję interesującej notacji graficznej. Praca raportuje także wiele dobrze skonstruowanych eksperymentów obliczeniowych (część z nich przeprowadzonych na niemałej próbie ponad 80 problemów), z których kilka potwierdza hipotezy postawione przez Autora i prowadzi do obiektywnie bardzo dobrych wyników, przewyższających metody referencyjne stosowane we wcześniejszych pracach.

Praca prezentuje rodzinę nowych, nietrywialnych i dobrze uzasadnionych metod. Opisane badania przeprowadzone zostały w sposób metodologicznie poprawny, a ich realizacja wymagała znaczących umiejętności naukowych i praktycznych. Wyniki są zachęcające i dają nadzieję na owocną kontynuację.

Dobre wrażenie z lektury pracy nie powstrzymuje mnie jednak przed sformułowaniem kilku uwag o charakterze polemicznym.

W częściach koncepcyjnych i teoretycznych praca prezentuje wiele nietrywialnych i interesujących pojęć i formalizmów, w których analizie Autor wykazał się znaczną biegłością. W tym kontekście zaskakująca jest praktycznie całkowita nieobecność analizy statystycznej otrzymanych przez niego wyników empirycznych. Co prawda w niektórych tabelach (np. V.4) wartościom średnim towarzyszą przedziały ufności (lub odchylenia standardowe), jednak dziś, w dobie powszechnej dostępności pakietów statystycznych, gdzie przeprowadzenie analizy np. testem ANOVA czy Friedmana sprowadza się do wydania jednego polecenia, nie poddanie tych wyników dalszej interpretacji statystycznej uważam za zauważalne niedociągnięcie.

Interpretacja wyników w sekcji Analiza jakości przewidywań klasyfikatorów w Rozdziale III jest nadmiernie optymistyczna. Korelacje prezentowane w Tabeli III.5 są obiektywnie bardzo małe; zwyczajowo wartości współczynnika korelacji liniowej Pearsona w przedziale $(-0.3, 0.3)$ interpretuje się jako brak korelacji. Spodziewam się że Autor nie dyskutuje istotności statystycznej tych wyników, ponieważ prawdopodobnie taka istotność tam nie ma miejsca. W tym kontekście nieco zaskakuje że Autor ucieka się do analizy względnej skuteczności (Tabela III.6), podczas gdy moim zdaniem dużo bardziej uzasadnione i potencjalnie obiecujące byłoby przeanalizowanie innych form korelacji, np. współczynników korelacji rangowej Spearmana czy Kendalla, które nie czynią założenia o liniowym charakterze współzmienności.

Powyższe uwagi dotyczą jednak jedynie wybranych fragmentów pracy, a prezentowane wyniki w zdecydowanej większości nie budzą zastrzeżeń i są interesujące.

4 Ocena redakcji rozprawy

Praca, przygotowana w języku polskim, zaprojektowana jest i zredagowana bardzo starannie. W zdecydowanej większości przypadków rozdziały i sekcje układają się w logiczny przewód. Powtórzenia są stosunkowo nieliczne. Szczególnie doceniam styl który cechuje się zwięzłością i zazwyczaj wczesną prezentacją pojęć i koncepcji, a dopiero następnie ich dyskusją.

Niedociągnięć w redakcji pracy jest niewiele. Najbardziej rzucił mi się w oczy brak jawnej listy przyczynków (zbliżonej do tej przygotowanej przeze mnie w p. 2 niniejszej recenzji), która moim zdaniem powinna być niezbędnym komponentem Wprowadzenia. Autor co prawda opisuje pracę zarówno w Streszczeniu jak i Wprowadzeniu, jednak jest to po prostu zapowiedź treści, bez szczególnie istotnego w tym kontekście rozróżnienia na treści oryginalne i zaczerpnięte z innych źródeł. Konfrontacja z listą przyczynków sformułowaną przez autora jest bardzo istotną częścią dyskursu naukowego, a w szczególności procesu recenzowania.

W zakresie kompozycji pracy, rozważyłbym włączenie Rozdziału IV jako sekcji do Rozdziału V.

Nieco zaskakuje brak numerów sekcji i podsekcji, co momentami utrudnia nawigację po materiale. Podobnie nietypowe jest zaniechania numerowania formuł. Konsekwencją przyjęcia tych konwencji jest stosunkowo rzadkie wiązanie przez Autora poszczególnych wątków poprzez odnośniki do innych sekcji, rozdziałów, czy formuł.

Wywody formalne prowadzone są poprawnie i czytelnie; drobnym odstępstwem od tej reguły jest sekcja „Trudność zadania klasyfikacji", w której równoległe odnoszenie się do (w gruncie rzeczy) tych samych wielkości przy pomocy notacji probabilistycznej (P) oraz funkcji gęstości (f, g) niekoniecznie było pomocne. W tej samej sekcji oczekiwałbym użycia tradycyjnego symbolu dla wartości oczekiwanej ($\mathbb{E}$), m.in. ponieważ zwykły symbol E pojawił się już w innej roli stronę wcześniej (na s. 37). Konwencje stosowanych oznaczeń czasami zmieniają się pomiędzy rozdziałami: np. symbol D stosowany w Rozdziale IV do oznaczenia dziedziny problemu (s. 63) był wcześniej używany w Rozdziale III do oznaczenia próby uczącej (s. 33).

W dobie gdy nawet ułamki procenta mogą być statystycznie istotne lub przekładać się na krytyczne różnice w zastosowaniach praktycznych, prezentowanie trafności klasyfikowania ('Skuteczność') w Tabeli III.8 z dokładnością do dwóch miejsc po przecinku, czyli jednego procenta, wydaje się zdecydowanie niewystarczające.

W pracy dopatrzyłem się zaledwie kilku literówek i niedociągnięć językowych. Na przykład uważam że w miejsce terminu *zmienna/rozkład kategorialny* naturalnej używać jest terminu *zmienna/rozkład nominalna* (s. 47). Skrót MLPPI na s. 85 nie został wcześniej wprowadzony ani rozwinięty

Te nieznaczne niedociągnięcia nie utrudniają jednak lektury rozprawy i nie wpływają na moją finalną ocenę.

5 Konkluzja końcowa

Wymienione powyżej uwagi polemiczne odnośnie treści i prezentacji pracy nie podważają głównych konkluzji rozprawy i mojej wysokiej jej oceny, zarówno od strony merytorycznej jak i prezentacyjnej. Uważam że cele postawione przez Autora pracy zostały osiągnięte, podparte silnymi wynikami empirycznymi, oraz przedstawione w interesujący sposób. Usystematyzowane i dobrze sformalizowane podejście do wieloskalowych zadań uczenia maszynowego zaproponowane w rozprawie może znacząco przyczynić się do efektywnego stosowania algorytmów uczących się w złożonych problemach rzeczywistych, gdzie coraz częściej dane nie mają jednopoziomowej, czysto tabelarycznej reprezentacji.

Wobec powyższych obserwacji, stwierdzam że rozprawa mgr inż. Zubka wyraźnie wykracza ponad poziom przeciętny i **spełnia z nawiązką warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony**.