Singapur, 6 stycznia 2017 roku

Prof. dr hab. inż. Jacek Mańdziuk
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

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

Niniejsza recenzja została przygotowana na prośbę Zastępcy Dyrektora ds. Naukowych Instytutu Podstaw Informatyki Polskiej Akademii Nauk prof. dr. hab. inż. Wojciecha Penczka wyrażoną w piśmie numer IPI PAN-RN-55/2016 z dnia 2 listopada 2016 roku.

Tematyka rozprawy

Przedstawiona do recenzji rozprawa dotyczy zagadnienia klasyfikacji wieloskalowej danych definiowanego jako umiejętność wykorzystania danych pochodzących z różnych poziomów opisu rozważanego zjawiska (źródła danych).

Podejście wielkoskalowe do opisu danych zakłada różną granulację informacji na różnych poziomach opisu, zarówno pod względem ilościowym (liczba atrybutów danych) jak i jakościowym (faktyczna istotność opisu danej skali w kontekście skuteczności budowanego klasyfikatora).

Podjęte w rozprawie zagadnienie integracji danych pochodzących z wielu skal opisu ma istotne znaczenie z punktu widzenia praktycznego stosowania metod analizy, przetwarzania oraz klasyfikacji danych opisujących problemy i zjawiska występujące w różnorodnych obszarach badawczych. Domeną rozważaną w rozprawie są dane biologiczne dotyczące oddziaływań białek, których klasyfikacja stanowi – ze swej natury – istotne wyzwanie dla obecnie stosowanych metod uczenia maszynowego. W badanym zagadnieniu zastosowanie komitetu klasyfikatorów wykorzystującego wielopoziomowy opis danych (czyli

reprezentację wieloskalową problemu) umożliwia uzyskanie istotnej poprawy wyników klasyfikacji w stosunku do podejść jednoskalowych.

Rozprawa, co warto podkreślić, ujmuje zagadnienie klasyfikacji wieloskalowej w sposób kompleksowy, odnosząc się do wszystkich kluczowych etapów tego procesu: wyboru struktury klasyfikatora, właściwej reprezentacji cech, dokonania klasyfikacji oraz oceny jakości uzyskanych wyników.

Hipotezy badawcze

Czytając rozprawę nie znalazłem jawnego sformułowania tzw. tezy doktorskiej czy hipotez badawczych. Brak ten stanowi mankament formalny – rozprawa doktorska, co do zasady, powinna zawierać tezę doktorską i/lub hipotezy badawcze.

Z drugiej strony, zarówno w streszczeniu jak i we wprowadzeniu Autor jednoznacznie definiuje przedmiot oraz cel rozprawy (strona 5, pierwszy akapit), które po odpowiednim przeformułowaniu mogą być uznane za hipotezy badawcze dysertacji. W związku z powyższym, pomimo niespełnienia zwyczajowego wymogu zdefiniowania *explicite* hipotez badawczych, nie czynię z tego braku istotnego zarzutu.

Treść rozprawy

Rozprawa liczy 106 stron, zawiera streszczenie w językach polskim i angielskim i składa się z 6 rozdziałów oraz spisu literatury.

W rozdziale wprowadzającym Autor przybliża zagadnienie rozważane w pracy, tj. problem integracji danych wielkoskalowych w zagadnieniu klasyfikacji, w kontekście budowy efektywnych komitetów klasyfikatorów uwzględniających w procesie klasyfikacji hierarchiczną naturę danych opisujących badane zjawisko.

W rozdziale drugim Autor wprowadza podstawowe definicje wykorzystywane w dalszej części rozprawy oraz przedstawia metodę analizy formalnej oraz prezentacji działania klasyfikatorów – jako ciągu przekształceń danych. W tym kontekście Autor przeprowadza przegląd istniejących metod klasyfikacji, ilustrując graficznie (w oparciu o omawianą formalną metodykę) popularne sposoby konstruowania komitetów klasyfikatorów (*bagging*, *boosting*, *stacked generalization*, sieci neuronowe jednokierunkowe oraz sieci konwolucyjne), a także ich rozszerzenia na przypadek klasyfikacji wieloskalowej.

Rozdzial zredagowany jest bardzo dobrze. Nie mam istotnych uwag do jego treści, poza dwiema sugestiami: jedna z podstawowych definicji wykorzystywanych w rozprawie, mianowicie określnie *stopnia związania danych* powinna być zilustrowana konkretnymi przykładami, które ułatwiłyby zrozumienie intuicji stanowiących jej podstawę. W oparciu o podane przykłady warto byłoby także przeprowadzić dyskusję adekwatności czy uniwersalności przyjętej definicji oraz jej potencjalnych ograniczeń w konkretnych kontekstach realizacyjnych. Druga uwaga dotyczy następującego stwierdzenia umieszczonego na str. 9: „Zewnętrzna rzeczywistość jest ciągła, niepodzielna i dynamiczna". Szczerze mówiąc nie do końca rozumiem znaczenie tego zdania w rozważanym kontekście. Model otoczenia (podobnie jak model zjawiska) są często konsekwencją przyjętej formy opisu i nie muszą w sposób bezwzględny posiadać wymienionych wyżej cech.

Kolejny rozdział omawia zagadnienie złożoności danych (dokładniej: próbki danych wejściowych) rozumianej w rozprawie jako stopień „gęstości informacji zawartej w zbiorze danych". W celu zbadania tak rozumianej złożoności danych Doktorant przedstawia autorską metodę jej szacowania w oparciu o próbkowanie podzbiorów zbioru danych różnych rozmiarów oraz obliczanie odległości Hellingera pomiędzy rozkładem prawdopodobieństwa indukowanym przez pełen zbiór danych a rozkładem indukowanym przez próbkowany podzbiór. W oparciu o średnią odległość Hellingera względem rozmiaru wybieranego podzbioru wykreślana jest, zaproponowana przez Autora, tzw. krzywa złożoności, która umożliwia ocenę stopnia reprezentatywności danego podzbioru, a w efekcie może służyć do doboru rozmiaru próbki uczącej. W dalszej części rozdziału Autor analizuje teoretyczne własności przyjętej metody oraz demonstruje praktyczne możliwości jej wykorzystania do scharakteryzowania rozwiązywanego problemu klasyfikacyjnego oraz do analizy skuteczności konkretnych algorytmów klasyfikacyjnych.
Omawiany rozdział przedstawia najważniejsze samodzielne wyniki Autora, stąd jego ocena ma kluczowe znaczenie z punktu widzenia oceny całości rozprawy. W moim przekonaniu, pomimo pewnych upraszczających założeń (dotyczących niezależności atrybutów danych czy równomiernego podziału danych pomiędzy klasy) opisana w rozprawie metoda doboru rozmiaru próbki danych jest interesującą propozycją alternatywną w stosunku do istniejących podejść. Na szczególne wyróżnienie zasługuje pogłębiona analiza eksperymentalna własności zaproponowanej metody w oparciu o wygenerowane zbiory syntetyczne ilustrujące zarówno jej zalety jak i potencjalne ograniczenia. Jednym z istotnych wniosków płynących w przeprowadzonej analizy jest zależność pomiędzy wzajemnym położeniem krzywej złożoności oraz krzywej uczenia (definiowanej w standardowy sposób) a złożonością rozważanego problemu (rozumianą jako minimalny wymagany rozmiar zbioru uczącego). W ostatniej części rozdziału Autor porównuje zaproponowaną przez siebie metodę doboru zbioru uczącego z klasycznymi podejściami realizującymi różne warianty podziału.

Zawartość omawianego rozdziału oceniam wysoko. Metoda badania złożoności próbki danych zaproponowana przez Autora, z uwagi na niezależność od wykorzystywanego modelu klasyfikatora, ma charakter uniwersalny i może być potencjalnie stosowana w różnorodnych kontekstach i problemach badawczych. Metoda, w naturalny sposób, może zostać rozszerzona do przypadku, w którym dane zbierane są na bieżąco (nauka online).
Obok niewątpliwych zalet wymienionych powyżej, lektura tego fragmentu pracy pozostawia także pewien niedosyt. Po pierwsze, zabrakło w nim pogłębionej dyskusji dotyczącej doboru niektórych kluczowych parametrów metody, np. wpływu parametru wygładzania h na jakość otrzymywanych wyników. Szkoda także, że Autor nie rozwinął wątku dotyczącego możliwości detekcji obserwacji odstających w oparciu o wariancję krzywej złożoności i nie zaproponował jakiejś heurystycznej metody ich szacowania w praktyce (sama informacja o zwiększonej wariancji nie jest wystarczająca bez stosownego punktu odniesienia).
Z uwag o mniejszym znaczeniu – charakterystyka zbioru Glass w tabeli III4 jest błędna, prawidłowa liczba atrybutów to 10, a liczba klas 6.

Rozdział czwarty odnosi się do zagadnienia konstrukcji zbiorów treningowego i testowego w przypadku klasyfikatorów wielkoskalowych. Zagadnienie to, z racji na istniejącą wewnętrzną hierarchię danych oraz zależności pomiędzy poszczególnymi skalami opisu wymaga, w przypadku wieloskalowym, szczególnej uwagi. W szczególności, jak pokazuje Autor, proste zastosowanie metod jednoskalowych dla danych o wielu skalach opisu może prowadzić do sztucznego zawyżenia wyniku klasyfikacji.
Rozważając problem właściwego podziału zbioru danych na dane treningowe i testowe w kontekście istnienia wielu skal, Autor wykorzystuje reprezentację zależności pomiędzy danymi w postaci hipergrafu i sprowadza problem do znalezienia ustalonej z góry liczby k składowych wierzchołków hipergrafu. W tym celu proponowany jest autorski algorytm zrównoważonego podziału zbioru wierzchołków nazwany „wet za wet", w którym składowe rozszerzane są naprzemiennie w sposób zachłanny, którego skuteczność jest weryfikowana w dalszej części rozprawy w zagadnieniu oddziaływania par białek.
Omawiana część pracy zawiera, podobnie jak rozdział poprzedni, istotne wyniki samodzielne Doktoranta adresujące bardzo istotny, a w praktyce często traktowany pobieżnie, problem właściwej ewaluacji narzędzia klasyfikacyjnego. Zaproponowana metoda okazała się skuteczna dla problemu oddziaływań białek, niemniej jej zastosowanie wykracza poza tę domenę. Szkoda, że Autor nie wymienił innych przykładów jej potencjalnych zastosowań.
Wracając do przykładu rozważanego w pracy, wyjaśnienia wymaga długość reprezentacji pary białek (w postaci ciągów częstości wystąpień aminokwasów) wynosząca 44 (a nie 40 jak sugerowałaby liczba istniejących aminokwasów).

Kolejny rozdział dysertacji zawiera podstawowe wyniki eksperymentalne dotyczące zastosowania wieloskalowego schematu klasyfikacji w oparciu o komitety klasyfikatorów w zagadnieniu oddziaływania par białek. W tym celu wykorzystywane są wieloskalowe schematy klasyfikacji (zaprezentowane w rozdziale II), autorska metoda analizy złożoności danych (opisana w rozdziale III) oraz autorska procedura ewaluacji klasyfikatora w oparciu o zrównoważony zbiór danych (zaprezentowana w rozdziale IV). Rozważany schemat klasyfikacji wykorzystuje dwie skale opisu: poziom sekwencji aminokwasów oraz poziom struktury białkowej. Szczegółowa analiza uzyskanych wyników eksperymentalnych wskazuje na przewagę klasyfikatora wielkoskalowego nad podejściami jednoskalowymi. Z drugiej strony, jak stwierdza Autor, uzyskane wyniki wciąż nie są w pełni satysfakcjonujące i planowane są dalsze prace mające na celu poprawę skuteczności metody wieloskalowej w badanym zagadnieniu.
Pomimo, że uzyskane wyniki nie pozwalają na stwierdzenie, że problem oddziaływania białek został satysfakcjonująco rozwiązany (w przypadku dwóch rozważanych w pracy organizmów), przewyższają one rezultaty klasyfikatorów jednoskalowych i stanowią postęp na drodze do uzyskania (w przyszłości) bardziej satysfakcjonujących rozwiązań. Entuzjastyczną ocenę uzyskanych wyników wstrzymuje dość istotna, moim zdaniem, specyfika rozważanego problemu i konieczność znaczącej adaptacji (parametryzacji) metody w przypadku rozwiązywania innych problemów. W szczególności transformacja z reprezentacji macierzowej do postaci wektora przedstawiona na stronie 81 jest bardzo specyficzna i słabo uzasadniona w rozprawie. Ogólnie rzecz ujmując, zabrakło w tym miejscu rozważań na temat wykorzystania zaproponowanych przez Doktoranta metod w obszarach innych niż rozważane w rozprawie zagadnienie oddziaływania par białek.

Ostatni rozdział zawiera podsumowanie wyników przedstawionych w pracy, przypomina podstawowe wnioski płynące z przeprowadzonych badań oraz zarysowuje możliwości kontynuacji prac badawczych w obszarze poruszonych w rozprawie zagadnień.

Rozprawę kończy spis literatury obejmujący około stu pozycji, z których większość opublikowana została w okresie ostatnich 10 lat. Dobór pozycji bibliograficznych oraz sposób posługiwania się zawartymi w cytowanych pracach wynikami potwierdza pogłębioną wiedzę Autora w zakresie szeroko rozumianej problematyki klasyfikacji danych (w tym klasyfikacji wieloskalowej) oraz zastosowań w domenie biologii obliczeniowej.

Oryginalny wkład Autora rozprawy

Oryginalny wkład Autora w ramach rozważanego w rozprawie zagadnienia naukowego dotyczy trzech następujących obszarów:

1. Opracowania metody szacowania złożoności danych (rozumianej jako stopień nasycenia informacją rozważanej próbki danych) w postaci tzw. *krzywej złożoności*, wspomagającej dobór rozmiaru próbki danych oraz odpowiedniego narzędzia klasyfikacyjnego;

2. Zaproponowania nowatorskiej metody testowania klasyfikatorów wielkoskalowych bazującej na zrównoważonym podziale zbioru obserwacji (nazywanej w rozprawie zasadą „wet za wet") oraz wykazanie wyższości tego podejścia nad powszechnie stosowanymi podziałami „losowymi";

3. Wykazanie przewagi rozważanego w rozprawie schematu klasyfikacji wieloskalowej nad podejściem jednoskalowym na przykładzie zagadnienia przewidywania oddziaływań białek, będącego nietrywialnym zagadnieniem klasyfikacyjnym o istotnym znaczeniu praktycznym.

Wymienione wyżej rezultaty badawcze zostały częściowo opublikowane w czasopiśmie *PeerJ* należącym do części A listy ministerialnej.

Konkluzja

Pracę doktorską mgr Juliana Zubka czyta się bardzo dobrze. Autor sprawnie operuje językiem polskim a poruszane zagadnienia przedstawione są w sposób spójny i logiczny, bazując na formalizmie matematycznym. Nieliczne błędy językowe (których nie wymieniam szczegółowo w recenzji) w najmniejszym stopniu nie umniejszają mojej wysokiej oceny pracy od strony redakcyjnej. Liczba błędów literowych czy stylistycznych, które zauważyłem w trakcie czytania bez wątpienia mieści się w „zwyczajowych granicach".
Na szczególne podkreślenie zasługuje warstwa teoretyczna pracy, w szczególności zaproponowane przez Autora podejście do problemu badania złożoności informacyjnej danych oraz sformułowana przez Niego metoda zrównoważonego podziału danych. Oba wymienione wyżej elementy dysertacji odnoszą się do kluczowych aspektów zagadnienia klasyfikacji i mogą być potencjalnie wykorzystane w wielu kontekstach badawczych, wykraczających poza ramy rozprawy.
Warstwa eksperymentalna stanowi słabszą stronę pracy i zdecydowane mogłaby być poszerzona o testy na innych - poza przedstawionym w pracy – problemach. Szersza ewaluacja eksperymentalna niewątpliwie umożliwiłoby pełniejszą ocenę praktycznej skuteczności zaproponowanej metodyki budowy klasyfikatorów wieloskalowych.

Wymienione w recenzji nieliczne uwagi krytyczne mają charakter polemiczny i nie zmniejszają mojej ogólnie wysokiej oceny dysertacji zarówno w warstwie merytorycznej jak i prezentacyjnej. Rozprawa dotyczy aktualnej, istotnej i szeroko rozwijanej tematyki badawczej i stanowi ważny wkład w obszarze rozwoju metod klasyfikacji danych ze szczególnym uwzględnieniem zróżnicowanego poziomu granulacji opisu danych.

W związku z powyższym **stwierdzam, że rozprawa spełnia z nawiązką wymagania stawiane przez odnośną Ustawę i wnoszę o jej przyjęcie oraz dopuszczenie jej Autora, mgr inż. Juliana Zubka do dalszych etapów przewodu doktorskiego.**