Prof. dr hab. Henryk Rybinski
Instytut Informatyki PW
Warszawa, Nowowiejska 17

Warszawa, 2017-09-15

Recenzja
rozprawy doktorskiej mgr inż. Sylwii Buregwy-Czumy
pt. " metody stosowania wiedzy dziedzinowej do poprawiania jakości klasyfikatorów".

1. Wstęp

Przedstawiona do recenzji rozprawa składa się z dziesięciu rozdziałów, bibliografii (163 pozycje, w tym 9 autorstwa bądź współautorstwa doktorantki), oraz dwóch dodatków. Objętość rozprawy - 223 strony w tym 176 stron tekstu podstawowego. W swojej opinii przedstawię ogólną charakterystykę rozprawy, a następnie przedstawię swoje uwagi ogólne i szczegółowe.

2. Charakterystyka ogólna

2.1 Dziedzina

Obszarem badań zaprezentowanych w rozprawie jest dziedzina eksploracji danych (ang. *data mining*). Zagadnienie analizy danych nie jest nową dziedziną. Jest to od dawna obszar zainteresowań statystyków. W dziedzinie sztucznej inteligencji zagadnienie eksploracji danych jest także od wielu lat przedmiotem intensywnych badań, w szczególności w kontekście akwizycji wiedzy oraz uczenia maszynowego, a także metod wnioskowania indukcyjnego przez analizę przykładów. Klasyczne propozycje z dziedziny statystyki w wielu praktycznych zastosowaniach eksploracji danych zawodzą w zetknięciu ze skalą zbiorów danych powstających w efekcie działania klasycznych systemów informacyjnych. Dlatego też od kilkunastu lat daje się zaobserwować gwałtowny wzrost zainteresowań badaniami nad nowymi metodami

eksploracji danych. Dziedzina eksploracji wiedzy w dużych zasobach informacyjnych wyznacza najważniejsze kierunki badań w takich dziedzinach sztucznej inteligencji jak metody uczenia maszynowego, czy budowa klasyfikatorów. Już w 1959 roku Arthur Samuel zdefiniował dziedzinę uczenia maszynowego jako "*...field of study that gives computers the ability to learn without being explicitly programmed*". Większość prowadzonych prac w tym zakresie koncentruje się na konstrukcji wydajnych algorytmów, często specjalizowanych pod kątem wybranych własności.

Opiniowana rozprawa leży w nurcie badań związanych z metodami klasyfikacji.

Główne zagadnienie rozważane w rozprawie leży na pograniczu informatyki i medycyny. Z punktu widzenia metod informatycznych dotyczy budowy klasyfikatorów z obszaru medycyny. Natomiast z medycznego punktu widzenia, rozprawa koncentruje się na zagadnieniu rozpoznawania istotnych zwężeń (stenoz) tętnic wieńcowych w chorobie niedokrwiennej serca i podjęcia decyzji czy należy wykonać zabieg udrażniania naczyń na podstawie danych klinicznych oraz wyników badania Holtera. W szczególności celem, jaki postawiła sobie doktorantka, jest opracowanie metod budowy klasyfikatorów na tyle dokładnych, aby możliwe było stosowanie tych klasyfikatorów w praktyce.

Podsumowując, uważam, że tematyka rozważana przez doktorantkę jest ważna i jest godna rozprawy doktorskiej.

2.2 Konstrukcja rozprawy

W konstrukcji pracy wyróżnić można następujące zasadnicze części:

1. motywacja badań, omówienie problemów badawczych poruszanych w pracy (rozdziały 1-3),
2. prezentacja opracowanych przez doktorantkę nowych metod budowania klasyfikatorów (rozdziały 4 - 8);
3. eksperymenty – w rozdziale 9 przedstawiony jest opis eksperymentów natomiast w rozdziale 10 zawarto ich podsumowanie oraz omówienie dalszych kierunków badań.

2.3 Charakter rozprawy

Praca ma przede wszystkim charakter praktyczny. Zawartość rozprawy dowodzi znajomości metod budowania klasyfikatorów. Badania autorki są wsparte implementacjami oraz eksperymentami, te zaś wskazują na

1. głęboką wiedzę dziedzinową w zakresie medycyny oraz
2. możliwości praktycznego zastosowania opracowanych algorytmów.

Możliwości praktycznego wykorzystania opracowanych rozwiązań są przekonywująco pokazane w rozdziale 9.

3. Wkład doktorantki

Wkład doktorantki oceniam pozytywnie – obejmuje on szereg ważnych elementów związanych z uczeniem maszynowym i klasyfikatorami. Wyróżnić tu można opracowanie metod wykorzystania wiedzy dziedzinowej do poprawy jakości klasyfikatorów. Autorka proponuje i implementuje cztery metody:

1. zdefiniowanie cech w oparciu o dane temporalne
2. zaproponowanie nowej miary jakości podziałów w węzłach drzewa decyzyjnego
3. Wprowadzenie cięć weryfikujących podział i zaproponowanie metody budowy drzewa z wykorzystaniem tych cięć;
4. wprowadzenie miary odległości semantycznej w ontologii i opracowanie na tej podstawie nowej metody budowania drzewa klasyfikującego.

Ponadto przeprowadza badania eksperymentalne na zaimplementowanej hurtowni danych wykazując przewagę opracowanych metod nad metodami standardowymi.

4. Poprawność

Pozytywnie oceniam przyjętą przez autorkę metodologię badań. Autorka dokonuje krytycznej analizy stanu badań w dziedzinie metod budowy klasyfikatorów, a następnie w oparciu o tę analizę proponuje szereg własnych rozwiązań, po czym przeprowadza badania eksperymentalne. Badania te stanowią istotny element rozprawy.

Uzyskane eksperymentalnie wyniki potwierdzają skuteczność proponowanych rozwiązań.

Warunkiem prawidłowego przeprowadzenia eksperymentów było stworzenie odpowiedniego warsztatu badawczego. Na bazie uzyskanych danych rzeczywistych w zakresie badań urządzeniem Holtera autorka stworzyła hurtownię danych.

Zawarta w pracy dyskusja wyników jest przeprowadzona w sposób jasny i czytelny.

5. Inne uwagi

Praca jest napisana starannie i dobrym językiem. Układ pracy jest logiczny i podporządkowany tezie rozprawy. Zamieszczenie w rozprawie stosowanej notacji oraz wykaz rysunków i tabel, ułatwia poruszania się po tekście. Również wprowadzenie dodatków czyni pracę bardziej przejrzystą.

Moje krytyczne uwagi przedstawiam poniżej.

Uwagi ogólne:

W wielu miejscach opisy i przykłady są zbyt rozwlekłe, czasem zastępują formalne definicje. Np. na str. 24-25 zamiast wprowadzić formalną definicję reguły, autorka omawia koncepcję reguły za pomocą przykładu. Podobnie na str. 26 nieformalnie rozpisuje się na temat „różnych miar", zamiast zdefiniować najważniejsze, np. wsparcie i zaufanie. Na stronie 95 autorka pisze „k należący do liczb naturalnych" zamiast np. $k \in \mathcal{N}$.

Uważam, że rozdziały 4-8 powinny tworzyć jeden wspólny rozdział poświęcony metodom. Jest to uzasadnione zarówno treścią jak też ich długością. W szczególności, w istniejącym układzie są one bardzo krótkie (na przykład Rozdział 4 liczy 8 stron, Rozdział 5 - 6 stron). Rozdział poświęcony opracowanym metodom mógłby mieć strukturę analogiczną do tej, jaką ma rozdział poświęcony eksperymentom wszystkich opracowanych alorytmów.

W wielu miejscach autorka używa pojęcia „metoda zachłanna" (rozdziały 5, 6), brakuje natomiast definicji tego pojęcia (choćby zgrubnej).

Uwagi szczegółowe

Na stronie 18 autorka pisze
„Ostatnio w literaturze..."
odwołując się do pozycji [48] i [173], przy czym jedna z tych pozycji jest z roku 2007, zaś druga z 2004, więc stwierdzenie, że „ostatnio" nie ma uzasadnienia.

Na stronie 22 autorka kategoryzuje symboliczne metody reprezentacji wiedzy w sposób niezbyt precyzyjny – w szczególności ontologie należałoby zaliczyć przede wszystkim do metod bazujących na zastosowaniach logiki (np. OWL).

Str. 26 –autorka pisze, że pojęcia „ontologia" używa się od lat 60-tych, Dobrze byłoby wskazać na literaturę. Wg. mojej wiedzy w kontekście sztucznej inteligencji McCarthy używa tego pojęcia w roku 1980.

Str. 26 – w zastosowaniach informatycznych celem ontologii jest przede wszystkim uzgadnianie znaczeń dla systemów, a nie aby „wiedza była z łatwością przetwarzana przez człowieka".

Podrozdział 2.4 – tytuł „Definicja wiedzy dziedzinowej" jest trochę na wyrost, ponieważ nie ma tu formalnych definicji a jedynie rozważania na temat „wiedzy dziedzinowej. Lepszy tytuł byłby „Wiedza dziedzinowa". Poza tym lepiej unikać pojęć nieostrych, wprowadzać tylko takie, które daje się (formalnie) zdefiniować.

W Rozdziale 5 zabrakło mi dyskusji, jak określać wagi. Jest tutaj też problem, że skoro zależy to od eksperta, to potem należy zweryfikować, czy inaczej ustawione wagi (przez innego eksperta) nie dają lepszych wyników.

Str. 95: Niezbyt zręcznie oznaczana jest funkcja *Disc* – raz jako funkcja jednej zmiennej ($Disc(p)$), a zaraz obok jako funkcja 2-ch zmiennych ($Disc(p_i,p_j)$).

Str. 105: Pełniejszego wyjaśnienia wymaga ustalenie wag pojęć podrzędnych w drzewie pojęć w metodzie IV. Mylące jest wyjaśnienie autorki:

„Wagi zostały dobrane arbitralnie przez eksperta dziedzinowego, w taki sposób, że suma wag pojęć podrzędnych (podklas) odpowiada wadze pojęcia nadrzędnego (nadklasy)."

Mam wrażenie, że rozkład wag podtypów jest odzwierciedleniem statystycznych częstości występowania podtypów w ramach określonego typu, co zresztą sugeruje sumowanie się wag podtypów do 1 (rys. 7.1), a nie do wagi nadtypu.

We wzorze 7.4 są niejasności. Co, jeżeli w zbiorze atrybutów A są zarówno atrybuty numeryczne, jak i symboliczne.

Wzór 8.4 (prawa strona reguły) jest niejasny. We wzorze 8.3 po prawej stronie występuje prawdopodobieństwao, zas w 8.4 funkcja $E(d)$ – czy to też jest prawdopodobieństw? Ponadto, we wzorze 8.5 *dec* występuje raz jako zbiór decyzji (pod sumą), a innym razem jak decyzja (w zapisie $P(dec=d)$

Uwagi terminologiczne

Terminy *top-down* i *bottom-up* należy tłumaczyć "od ogółu do szczegółu" i „od szczegółu do ogółu" a nie „góra-dół" czy „dół-góra"

Uwagi te mają w większości przypadków charakter uwag redakcyjnych i nie zmieniają one mojej pozytywnej ocenę rozprawy.

Podsumowanie

Praca ma oryginalny charakter. Wkład autorki jest istotny. Autorka uzyskała interesujące wyniki o znaczeniu praktycznym. Uważam, że opiniowana praca spełnia wymagania zawarte w obowiązujących przepisach dotyczących rozpraw doktorskich, wnoszę zatem o dopuszczenie mgr inż. Sylwii Buregwa-Czumy do publicznej obrony.