Warszawa, 20.07.2019

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

Recenzja rozprawy doktorskiej mgr Rafała Latkowskiego
p/t. „Wnioskowanie z danych z Brakującymi Wartościami Atrybutów"

Dane ogólne

Na przedstawioną do recenzji rozprawę składa się zbiór ośmiu publikacji w języku angielskim, ponadto autoreferat oraz konspekt w języku polskim. Przedstawiony zbiór publikacji 4 publikacje autorskie oraz 4 współautorskie (z tych czterech w jednej doktorant jest na pierwszym miejscu):

1. Latkowski, Rafał. "Application of data decomposition to incomplete information systems." *Intelligent Information Systems 2002*. Physica, Heidelberg, 2002. 321-330.
2. **Latkowski, Rafał**. "On decomposition for incomplete data." *Fundamenta Informaticae* 54.1 (2003): 1-16.
3. Latkowski, Rafał, and Michał Mikołajczyk. "Data decomposition and decision rule joining for classification of data with missing values." *Transactions on Rough Sets I*. Springer, Berlin, Heidelberg, 2004. 299-320.
4. Latkowski, Rafal. "High computational complexity of the decision tree induction with many missing attribute values." *Proceedings of CS&P*. 2003.
5. **Latkowski, Rafał.** "Flexible indiscernibility relations for missing attribute values." *Fundamenta informaticae* 67.1-3 (2005): 131-147.
6. Bazan, Jan G., Rafał Latkowski, and Marcin Szczuka. "Missing template decomposition method and its implementation in rough set exploration system." *International Conference on Rough Sets and Current Trends in Computing*. Springer, Berlin, Heidelberg, 2006.
7. Bazan, Jan G., Rafał Latkowski, and Marcin Szczuka. "DIXER–distributed executor for rough set exploration system." *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*. Springer, Berlin, Heidelberg, 2005.

8. Wojna, Arkadiusz, and Rafał Latkowski. "RSESLIB 3: Library of rough set and machine learning methods with extensible architecture." *Transactions on Rough Sets XXI*. Springer, Berlin, Heidelberg, 2019. 301-323.

Zadeklarowany przez autora udział w każdej z prac współautorskich jest powyżej 50%, średnio wynosi ok. 60%. W zbiorze przedstawionych prac dwie są w czasopismach z listy JCR, obie autorskie. Łączna objętość wszystkich publikacji wynosi ponad 100 stron. Ponadto do rozprawy dołączono oświadczenia współautorów o ich udziale - dotyczy

1. Problem badawczy i jego znaczenie

Tematyka rozprawy ściśle wiąże się z metodami odkrywania wiedzy ze zbiorów danych. Dziedzina ta jest od dawna w obszarze zainteresowań statystyków, nieco krócej w obszarze zainteresowań naukowców zajmujących się metodami sztucznej inteligencji, w szczególności metodami uczenia maszynowego. Zagadnienie to ma ogromne znaczenie, zarówno praktyczne jak też teoretyczne, w szczególności w kontekście systemów akwizycji wiedzy, maszynowego uczenia, a także metod wnioskowania indukcyjnego przez analizę przykładów. Dlatego też od blisko 30-u lat daje się zaobserwować gwałtowny wzrost zainteresowań badaniami nad nowymi metodami analizy danych. Zainteresowania te dotyczą nie tylko środowisk akademickich, ale także przemysłowych laboratoriów badawczych. Wynika to przede wszystkim z zapotrzebowania na narzędzia w dziedzinie analizy dużych zasobów informacyjnych.

Wynika to przede wszystkim z zapotrzebowania na narzędzia w dziedzinie analizy dużych zasobów informacyjnych. Dziedzina eksploracji wiedzy w dużych zasobach informacyjnych wyznacza najważniejsze kierunki badań w takich dziedzinach sztucznej inteligencji jak metody uczenia maszynowego, 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 tym nurcie badań. Przytoczone publikacje dotyczą przede wszystkim problemów związanych z wnioskowaniem z niepełnych danych. Autor koncentruje się na zagadnieniach wnioskowania w ramach teorii zbiorów przybliżonych, stawiając sobie jako zadanie udoskonalenie znanych metod w tym zakresie. Dotyczy to metod indukcji klasyfikatorów, które nie wymagają uzupełniania brakujących danych

Tematyka rozważana przez doktoranta jest niezwykle ważna i niewątpliwie jest godna rozprawy doktorskiej.

1.2 Cel rozprawy

Zasadniczym celem pracy było opracowanie nowych metod wnioskowania na podstawie niepełnych danych przy zastosowaniu metod zbiorów przybliżonych. Celem dodatkowym było stworzenie praktycznych narzędzi, pozwalających realizować eksperymenty na zbiorach niepełnych danych w taki sposób, aby usprawnić efektywność obliczeń.

1.3 Charakter rozprawy i znaczenie praktyczne badań

Większość publikacji łączy w sobie charakter teoretyczny z eksperymentalnym, pokazując zarówno solidny teoretyczny warsztat autora, jak też warsztat praktyczny, pozwalający autorowi nie tylko realizować badania eksperymentalne, ale także rozwijać narzędzia do przeprowadzenia tych badań. Propozycje autora są wsparte zaawansowanymi implementacjami oraz eksperymentami. Wskazują one na

1. solidny warsztat doktoranta w zakresie narzędzi uczenia maszynowego;
2. duży potencjał praktycznego wdrażania opracowanych algorytmów;

2. Wkład doktoranta

Wkład autora obejmuje szereg ważnych elementów związanych z uczeniem maszynowym na niekompletnych danych w kontekście zbiorów przybliżonych:

1. W pracach [1-2] autor prezentuje opracowaną przez siebie metodę indukcji klasyfikatorów poprzez dekompozycję danych na regularne obszary danych kompletnych oraz łączenie wniosków częściowych uzyskanych z klasyfikatorów utworzonych dla tych obszarów. Autor pokazał eksperymentalnie wyższość opracowanego podejścia w stosunku do metod zbiorów przybliżonych bez stosowania dekompozycji, jak tez w stosunku do metody C4.5. Metoda ta w połączeniu z algorytmem skracania i łączeniu reguł decyzyjnych [3] okazała się jeszcze bardziej dokładna, a dodatek pozwala zredukować liczbę reguł klasyfikatora;
2. W [4] autor pokazał problemy ze złożonością obliczeniową algorytmów indukcji drzew decyzyjnych w przypadku, gdy rośnie liczba brakujących wartości atrybutów w przypadku stosowania strategii rozdzielania niekompletnego obiektu do wszystkich węzłów „dzieci" danego węzła.

3. Autor wprowadził w [5] uogólnione relacje nierozróżnialności. Zabieg ten pozwala definiować szeroką klasę możliwych znaczeń dla brakujących wartości, w szczególności relacje ograniczone atrybutowo i relacje ograniczone deskryptorem. Przeprowadzone eksperymenty pokazały, że dzięki doborowi odpowiednich znaczeń uzyskuje się poprawę dokładności klasyfikatora, aczkolwiek poprawa ta często nie jest statystycznie istotna z uwagi na dużą wariancję.
4. Istotnym wkładem doktoranta jest rozwój oprogramowania RSESLIB, realizowanego w latach 2005-2019 (w zbiorze publikacji ujętych w rozprawie są to prace [6-8]). W ramach tych prac autor podejmował się próby uwzględnienia w oprogramowaniu pojęć związanych z indukcją i dopasowaniem do obiektów w przypadku relacji nierozróżnialności. Zasadniczym osiągnięciem jest implementacja oprogramowania dla obliczeń rozproszonych z wykorzystaniem paradygmatu sieci komputerów (*grid computing*).

Osiągnięcia te oceniam pozytywnie. Wartość opracowanych metod została także potwierdzona cytowaniami w bazach Google Scholar i Scopus (liczba cytowań odpowiednio 156 i 95)

3. Wiedza kandydata

W rozprawie doktorantka wykazuje się dużą wiedzą, przede wszystkim w zakresie uczenia maszynowego, w tym, m.in. w metodach analizy danych bazujących na zbiorach przybliżonych.

4. Inne uwagi

Pracę stanowi 8 publikacji, w tej liczbie znajdują się dwa artykuły z lity JCR, dwie prace w podserii *Transactions on Rough Sets* serii LNCS, pozostałe są opublikowane w materiałach konferencji o zasięgu międzynarodowym. Uważam, że przedstawione artykuły stanowią dorobek, który jest dobrą podstawą do rozprawy doktorskiej.

Moja zasadnicza uwaga wiąże się z brakiem wprowadzenia do zaprezentowanych publikacji. Wprawdzie doktorant dołączył do rozprawy autoreferat i konspekt, które w pewnym zakresie wypełniają lukę, jednak moim zdaniem są to nieco powierzchowne teksty. W szczególności brakuje mi przeglądu równoległych algorytmów w dziedzinie zbiorów przybliżonych i ustosunkowania się do nich. Zabrakło mi na przykład odpowiedzi na pytanie jakie są różnice pomiędzy proponowaną implementacją z wykorzystaniem sieci

komputerów a wcześniej proponowaną implementacją proponowaną przez Roberta Susmagę (1998) i następnie przez T. Strąkowskiego (2008)

1. Susmaga R. (1998) Parallel Computation of Reducts. In: Polkowski L., Skowron A. (eds) Rough Sets and Current Trends in Computing. RSCTC 1998. Lecture Notes in Computer Science, vol 1424. Springer, Berlin, Heidelberg
2. Strąkowski, T. (2008). *Parallel algorithms for Rough Sets Theory* (Doctoral dissertation, The Institute of Computer Science).

Powyższe uwagi nie umniejszają mojej pozytywnej opinii o rozprawie.

5. Podsumowanie

Pozytywnie oceniam przedstawione w doktoracie artykuły i wkład doktoranta. Rozprawa spełnia wymagania zawarte w obowiązujących przepisach dotyczących rozpraw doktorskich, wnoszę zatem o dopuszczenie mgra inż. Rafała Łatkowskiego do publicznej obrony.