Kraków 29.09.2022

prof. dr hab. inż. Grzegorz J. Nalepa
Instytut Informatyki Stosowanej
Wydział Fizyki, Astronomii i Informatyki Stosowanej
Uniwersytet Jagielloński

Tytuł rozprawy: Indukcja reguł akcji na podstawie metody sekwencyjnego pokrywania
Autor rozprawy: Paweł Matyszok
Promotorzy rozprawy: dr hab. Marek Sikora, prof. Pol. Śl., dr inż. Łukasz Wróbel
Dziedzina: nauki techniczne
Dyscyplina: informatyka techniczna i telekomunikacja

1 Wstęp

Rozprawa doktorska mgra Pawła Matyszoka wpisuje się w niezwykle istotny w informatyce nurt metod eksploracji danych (ang. *data mining, DM*). W ostatnich dwóch dekadach widać ogromny wzrost zainteresowania tą dziedziną, podyktowany m.in. drastycznym zwiększeniem się dostępności danych w związku z rewolucją *Big Data*, dużym zapotrzebowaniem na użycie i rozwijanie metod DM, oraz w końcu bardzo dużymi postępami w obszarze uczenia maszynowego (ang. *machine learning, ML*), gdzie opracowuje się algorytmy stanowiące trzon DM. Natomiast w ostatnich kilku latach gorąco dyskutuje się wyzwania związane z tzw. objaśnialnością i interpretowalnością metod sztucznej inteligencji (ang. *artificial intelligence, AI*) – tematykę tę określa się ogólnie terminem XAI (ang. *eXplainable AI*). W ramach tych dyskusji podkreśla się m.in. wartość tych metod ML, które są w większym stopniu zrozumiałe dla ekspertów dziedzinowych i praktyków ML. Ma to szczególne znaczenie w przypadku podejść nienadzorowanych do uczenia, w przypadku których często mówimy, że proces DM ma charakter odkrywania wiedzy (ang. *knowledge discovery, KD*).

Dalsza część recenzji odpowiada strukturalnie wytycznym przekazanym przez Radę Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej.

2 Cel, zakres, teza i charakter rozprawy

Jedną z ważnych metod opisywania pojęć w DM są reguły. Jest to jedna z najczęściej wykorzystywanych metod reprezentacji oraz przetwarzania wiedzy w postaci symbolicznej i wraz pokrewnymi formalnie drzewami i tablicami decyzyjnymi lokuje się w centrum inżynierii wiedzy. Odkrywanie reguł w podejściach DM pozwala na formułowanie wiedzy w postaci zrozumiałej dla człowieka, a same algorytmy indukcji drzew i reguł były jednymi z pierwszych algorytmów KD.

Autora rozprawy interesuje specyficzny przypadek użycia reguł w DM do analizy zmiany wartości atrybutu decyzyjnego w regule. Służące do tego celu i rozważane w pracy reguły akcji (ang. *action rules*) (zaproponowane w 2000r. przez Rasia i Wieczorkowską) mogą w szerszym sensie służyć jako rekomendacja pozwalająca zmieniać dane, tak aby uzyskać konkretne decyzje, np. wynik klasyfikacji.

Za cel rozprawy Autor stawia sobie opracowanie algorytmu indukcji reguł akcji opierającego się o tzw. paradygmat sekwencyjnego pokrywania i uwzględniającego wiodącą rolę przykładów klasy źródłowej lub docelowej w procesie indukcji. Projektowany algorytm ma nadawać się do zastosowań w problemach klasyfikacyjnych i regresyjnych. Sam proces indukcji ma być kontrolowalny przez użytkownika. Jako drugi

cel Autor wskazuje sformułowanie kolejnego algorytmu analizy indukowanych reguł, który pozwalałby na rekomendację reguł akcji dostosowanych do konkretnego przykładu.
We wstępie do pracy Autor formułuje następującą tezę rozprawy:

Zastosowanie w odkrywaniu reguł akcji paradygmatu sekwencyjnego pokrywania i odpowiednio dobranych kryteriów sterowania algorytmem indukcji reguł pozwala na uzyskanie modeli o dobrych zdolnościach prognostycznych i objaśniających.

Ponadto w pracy sformułowane są tezy pomocnicze (numeracja GJN):

a) Metodyka wymieniona w głównej tezie pracy może być zastosowana do indukcji reguł akcji dla danych klasyfikacyjnych i opisujących problemy regresyjne.
b) W zależności od sposobu prowadzenia indukcji (rozważane podejścia to indukcja z punktu widzenia klasy źródłowej i indukcja z punktu widzenia klasy docelowej) uzyskujemy różne zbiory reguł, o różnych zdolnościach predykcyjnych i opisowych.
c) Agregacja wiedzy zawartej w wyznaczonych regułach akcji pozwala na opracowanie algorytmu rekomendacyjnego, zdolnego do wskazania dla zadanego obiektu w jaki sposób należy zmienić wartości jego atrybutów, aby osiągnąć zamierzoną wartość atrybutu decyzyjnego.

Mając na uwadze cel, tezę pracy, oraz przyjętą metodologię pracy, w której ewaluacja ma charakter eksperymentalny, sama rozprawa ma charakter koncepcyjno-eksperymentalny. W wyżej wymienionym we wstępie do recenzji kontekście, rozprawa p. Matyszoka dobrze lokuje się w bieżącej tematyce DM z uwzględnieniem perspektywy XAI.

3 Zawartość rozprawy

Rozprawa składa się z sześciu rozdziałów, z których pierwszy stanowi wstęp, gdzie Autor zawarł genezę, motywację, tezę i cele pracy, a ostatni zawiera krótkie podsumowanie. Praca ma objętość 172 stron, z czego bibliografia obejmuje 111 pozycji bibliograficznych na 13 stronach.

Rozdział 2. „Podstawowe definicje i metody indukcji reguł klasyfikacyjnych", o objętości 43 str., jest wprowadzeniem koncepcyjnym, terminologicznym, a przy tym przeglądem literatury z obszarów DM związanych z tematyką pracy. Obejmuje krótkie omówienie sposobów reprezentacji reguł i oceny działania klasyfikatorów regułowych. Następnie Autor omawia wybrane algorytmy indukcji reguł klasyfikacyjnych, same reguły akcji i metody ich indukcji.

W rozdziale 3. „Indukcja reguł akcji", o objętości 20 str, omówiona jest indukcja reguł akcji z zastosowaniem paradygmatu sekwencyjnego pokrywania, a także dyskutowany jest wpływ kierunku indukcji na postać reguł, oraz wskazane są inne rodzaje reguł akcji (tj. nie dotyczące zadania klasyfikacji). W rozdziale tym Autor wprowadza pierwsze z algorytmów których opracowanie wskazał jako cel pracy, tj. „Pokryciowy algorytm indukcji reguł akcji" (Algorytm 8, str. 56), oraz wykorzystywane w nim jako podprocedury „Specjalizacja reguły akcji" (Algorytm 9, str. 58), „Przycinanie reguły akcji" (Algorytm 10, str. 62), nadrzędny względem Alg. 8, algorytm „Dwukierunkowa indukcja reguł akcji" (Algorytm 11, str. 65) i w końcu pomocniczą procedurę „Odwracanie reguł akcji" (Algorytm 12, str. 67). Uzupełnieniem dyskusji jest Algorytm 13 Specjalizacja reguły akcji regresji na str. 74.

Rozdział 4. „Rekomendacje na podstawie zbiorów reguł akcji. Weryfikacja reguł akcji i rekomendacji.", o objętości 27 str. dotyczy opisu realizacji drugiego z celów rozprawy. Autor podaje metodę budowy rekomendacji reguł akcji, a także schemat (s. 90) procesu weryfikacji reguł akcji i rekomendacji (które również mają postać reguł akcji).

Część ewaluacyjna pracy jest zawarta w rozdziale 5. „Eksperymenty" o objętości 53 str. Rozdział rozpoczyna się od opisu ogólnego schematu weryfikacji eksperymentalnej. Następnie Autor przechodzi do opisu wybranych testów statystycznych w ocenie jakości algorytmów DM. Właściwa część ewaluacyjna składa się z dwóch fragmentów opisujących indukcję reguł akcji w zadaniach klasyfikacji i regresji. Ewaluacja jest przeprowadzona w oparciu o zbiory danych z popularnego repozytorium UCI.

4 Ocena rozprawy

Ocenę pracy warto rozpocząć od jej relacji do obecnie prowadzonych na świecie badań z obszaru DM. Jak wskazuję w dwóch pierwszych częściach recenzji, badania podjęte przez Doktoranta są niewątpliwie aktualne i dobrze ulokowane tematycznie na tle prac z obszaru DM, KD i w końcu XAI. Natomiast skupiając się na zawartości i układzie rozprawy, warto przyjrzeć się podjętym szczegółowo problemom badawczym i tezie pracy przytoczonej w części 2. recenzji. W mojej opinii teza jest sformułowana poprawnie, a koncepcja odkrywania reguł akcji w paradygmacie sekwencyjnego pokrywania jest wg mojej wiedzy oryginalna. Główna część tezy i tez pomocniczych a-c jest wykazana w stopniu zadowalającym w rozdziałach 3-5.

To co zwraca uwagę w konstrukcji nie tylko samej rozprawy, ale przedstawionej w niej argumentacji, to brak jasno wyartykułowanych problemów badawczych których rozwiązania podejmuje się Autor. Zamiast tego podjął On decyzję o sformułowaniu złożonej tezy i dodatkowo wskazania celów pracy. Jest to podejście spotykane w rozprawach i akceptowalne, ale w mojej opinii mniej przejrzyste. Ponadto można się zastanowić, na ile wszystkie elementy zawarte w tezie są równie istotne. W mojej opinii najważniejsza jest teza a, dotycząca możliwości rozwiązywania zadań klasyfikacyjnych i regresyjnych.

Każda rozprawa doktorska powinna być oparta o gruntownie przeprowadzone badanie stanu literatury z dziedziny w której jest tematycznie ulokowana. Doktorant wyniki swoich badań literaturowych zawarł głównie w rozdziale 2, choć odniesienia do powiązanych podejść badawczych można też znaleźć w rozdziałach 3 i 4. Przegląd obejmuje krótkie wprowadzenie do zapisu reguł, ich oceny w zadaniach klasyfikacyjnych i ich indukcji. Następnie Autor wprowadza wybrane pojęcia dotyczące reguł akcji i ich indukcji. Przegląd jest przeprowadzony zadowalająco pod kątem dalszej pracy z regułami akcji.

Trudno jednak oprzeć się wrażeniu, że zakres analizowanej literatury jest relatywnie wąski i silnie zorientowany na szczegółowe problemy indukcji reguł akcji. Użycie reguł w AI jest zagadnieniem niezwykle szerokim i jako takie wykracza poza zakres pracy. Tym niemniej, pisząc o regułach decyzyjnych warto było skonsultować pierwszą część rozdziału z literaturą z obszaru systemów wnioskujących. Reguły w dziedzinie DM też są wykorzystywane na kilka sposobów, w tym do budowy klasyfikatorów, rozwiązywania zadań regresji, oraz odkrywania wiedzy w paradygmacie nienadzorowanym za pomocą reguł asocjacyjnych. Autor o tym wspomina, jednak brakuje syntetycznego zestawienia stosowanych metod w postaci, która umożliwiłaby później ewaluację wyników pracy na tle istniejących, czasem zbliżonych rozwiązań.

O wartości rozprawy przede wszystkim świadczą uzyskane przez Doktoranta wyniki koncepcyjne zawarte w rozdziałach 3 i 4, poddane ewaluacji w rozdziale 5. Przedstawione algorytmy i metody uważam za wartościowe z punku widzenia DM i mające potencjał do zastosowań. Należy tu zwrócić uwagę na fakt, że podczas gdy indukcja reguł akcji i związane z nią zadania pomocnicze są opisane w rozdziale 3. za pomocą sformalizowanych algorytmów (8-13) to drugi z obszarów, czyli rekomendacja właściwych reguł i ich weryfikacja jest opisana relatywnie ogólnie i bez żadnej formalizacji (pomocniczą rolę pełni schematy na s. 90). Jest to zaskakujący brak w rozprawie, uniemożliwiający pełną ocenę poprawności zaproponowanych w rozdziale 4. rozwiązań. Tym bardziej brak jest opisu implementacji wyników, np. w postaci dodatku do rozprawy. Autor jedynie wspomina w podsumowaniu, że implementacja istnieje. Jeszcze istotniejszym wyzwaniem jest szersza ewaluacja przydatności opracowanych algorytmów i metod. We wstępie do rozdziału 4. Autor wskazuje problemy z oceną użyteczności reguł akcji w praktyce. Niestety w rozprawie trudno znaleźć jakieś rozwiązania tych problemów. Przeprowadzona w rozdziale 5 ewaluacja eksperymentalna wprawdzie wskazuje na poprawne (tu: zgodne z założeniami) działanie zaproponowanych algorytmów i metod, ale nie pomaga w ocenie ich użyteczności. W pracy nie ma informacji czy Autor próbował użyć swoich rozwiązań do rozwiązania praktycznych problemów, czy tylko ograniczył się do ich ewaluacji na publicznych zbiorach danych, które wprawdzie pochodzą z różnych dziedzin, ale najczęściej są poddane specyficznej obróbce ułatwiającej ich użycie.

Powyższe uwagi mają wpływ na ocenę pozycji naukowej rozprawy w stosunku do stanu wiedzy reprezentowanych przez literaturę światową i znaczenia uzyskanych wyników. Wprawdzie wskazane

powyżej niedostatki metodologiczne nie podważają poprawności uzyskania wyników, ale zmniejszają wpływ tych wyników na rozwijane na świecie metody DM. Uważam, że opisane w rozprawie wyniki Doktoranta mają potencjał by wzbudzić zainteresowanie społeczności naukowej, ale wymagałyby dalszego opracowania. Oczywiście oceniam zawartość rozprawy i nie odnoszę się tu do publikacji [83].

W mojej opinii, jakkolwiek Autor w sposób ogólnie poprawny przedstawia uzyskane wyniki, to sama konstrukcja pracy i jasność wywodu, pozostawia nieco do życzenia. Można odnieść wrażenie, że rozdziały 2, 3 i 4 powstawały w różnych okresach czasu a ich treść nie do końca została uspójniona. Wprowadzane w rozdziale 2 pojęcia i notacje nie zawsze są później wykorzystywane w sposób konsekwentny, a zawarty tam opis pokrewnych metod nie posłużył Autorowi do próby oceny jakościowej uzyskanych wyników na tle istniejącego stanu wiedzy. Na przykład kluczowe dla pracy pojęcie „reguły decyzyjnej" czasem jest używane w sposób ogólny, a czasem w sensie „reguły w zadaniu klasyfikacji". Niektóre podrozdziały wydają się niedopracowane, lub niepotrzebne, co zaburza wywód. Na przykład w rozdziale 3 pojawia się punkt 3.3. „Inne rodzaje reguł akcji", gdzie Doktorant najpierw opisuje reguły akcji w zadaniu regresji wprowadzając uzupełniający Algorytm 13, a następnie dodaje punkt „Reguły akcji w analizie przeżycia", w którym nie są opisywane żadne konkretne wyniki, a jedynie możliwość poszerzenia wcześniejszych wyników, na co jest raczej miejsce w podsumowaniu rozprawy. Rozdział powinien się skupiać na opisywaniu uzyskanych wyników w nawiązaniu do sformułowanych tez i celów pracy. Same opisy części algorytmów i metod są ilustrowane przykładami, co można uznać za zaletę. Z drugiej strony, przykłady te są abstrakcyjne i nie pokrywają się z tymi które zostały użyte w rozdziałach 2.

Ponadto sposób opisu wyników w rozdziale 4 istotnie odbiega od standardów uprzednio przyjętych w rozdziale 3. Podczas gdy Doktorant opisuje metodę indukcji reguł w postaci sformalizowanej jako algorytmy, to już metoda budowania rekomendacji jest opisana w postaci swobodnej i miejscami mało precyzyjnej wypowiedzi. Uniemożliwia to dokładną ocenę metody. Autor wskazuje wprawdzie, że rekomendacje są indukowane (4.2.2), ale brakuje konkretnego zdefiniowania jak ten proces wygląda i w jaki sposób korzysta z algorytmów z rozdziału 3. Warto również zwrócić uwagę, że Autor we wstępie deklaruje, że opracuje algorytm, co sugeruje sformalizowany sposób opisu. Elementarną próbą uspójnienia treści tak sformułowanych rozdziałów mogłyby być syntetyczne podsumowania, wskazujące zwięźle co zostało zrobione i jak uzyskane wyniki odnoszą się do założeń przyjętych we wstępie, tez i celów pracy.

Samo podsumowanie rozprawy wydaje się mało dopracowane i tylko częściowo spełnia swoją rolę. W tej części rozprawy powinna się znaleźć syntetyczna rekapitulacja argumentacji i uzyskanych wyników ze wskazaniem w jaki sposób zostały osiągnięte cele pracy i została wykazana teza. Ponadto podsumowanie zaczyna się od nieco zaskakującego zdania „W niniejszej pracy przedstawiono ogólny zarys algorytmu indukcji reguł akcji z zastosowaniem paradygmatu sekwencyjnego pokrywania." Wskazane we wstępie dwa cele pracy to opracowanie dwóch algorytmów dla indukcji i rekomendacji reguł akcji. Akurat najważniejszy algorytm – indukcji, a właściwie grupa algorytmów, jest opisany relatywnie precyzyjnie w rozdziale 3. Cześć rekomendacyjna jest opisana znacznie mniej precyzyjnie i być może do niej trafniej odnosi się mało fortunny „zarys". Dalsze uwagi szczegółowe zawarte są w kolejnej części recenzji.

Praca jest zredagowana ogólnie poprawnie. Szczegółowe usterki wskazuję w dalszej części recenzji.

Pomimo wskazanych powyżej niedostatków co do konstrukcji pracy, uważam osiągnięte wyniki za oryginalne i wartościowe, a moja jej ocena jest pozytywna.

5 Uwagi dyskusyjne i słabe strony rozprawy

W nawiązaniu do przedstawionej powyżej oceny rozprawy, za najważniejsze słabe strony rozprawy należy w mojej opinii uznać:

a) brak należytej spójności w opisie uzyskanych wyników – brak jest formalizacji metod rozdziale 4 i występują niespójności z definicjami wprowadzonymi z rozdziale 2,
b) ograniczoną ewaluację uzyskanych wyników – Autor nie podejmuje próby charakterystyki efektywności czy złożoności algorytmów w rozdziale 3, a eksperymentalna ewaluacja, która przeprowadzono w rozdziale 5 nie zawiera żadnej praktycznej informacji o implementacji używanych algorytmów, środowisku testowym, etc. Ponadto brak jest porównania jakości indukowanych reguł do tych otrzymanych za pomocą innych algorytmów.
c) nie w pełni przejrzysta struktura pracy – treść podrozdziałów nie zawsze stanowi spójną treść. Brakuje precyzyjnego wskazania w których miejscach – zdaniem Autora – zrealizowane są założone we wstępie cele. Wywnioskowanie tego podczas lektury pracy, nie zawsze jest łatwe, a niektóre fragmenty tekstu wydają się zbędne.

Poniżej formułuję szereg uwag i pytań szczegółowych.

1. Terminologia związana z eksploracją danych nie jest używana konsekwentnie: dotyczy to m.in. zamiennego używania pojęć „zadanie klasyfikacji/regresji" i „problem klasyfikacji/regresji".
2. Autor nie podaje na bazie jakiej literatury wprowadza w rozdziale 2.1 definicję „reguły decyzyjnej". Z dalszej części pracy można wnioskować, że utożsamia ją z regułą klasyfikującą, co w mojej opinii nie jest uzasadnione bez jawnej definicji. Podobnie nieco swobodnie traktowane są definicje atrybutów. Reguły decyzyjne są pojęciem bardziej ogólnym i w literaturze można spotkać znacznie bogatszą ich składnię, niż zakłada Autor.
3. Jeśli chodzi o algorytm/y indukcji reguł opisane w rozdziale 3, nasuwa się pytanie w jakim stopniu Doktorant przeprowadzał analizę ich złożoności.
4. W pracy należałoby jawnie opisać algorytm indukcji rekomendacji (4.2) i wskazać jego relację do bazowych algorytmów indukcji reguł.
5. Istotną usterką jest brak numeracji linii w algorytmie 13 na str. 74, podczas Autora jawnie do tej numeracji się odnosi na str. 72.
6. W definicji 2. na str. 12 podana jest składnia wyrażenia elementarnego i rozważane w nim relacje, wskazane za pomocą konkretnych symboli. Na str. 70 Doktorant pisze „relacja między wartościami konkluzji (...) może być dowolnie zadana przez użytkownika, który nie musi ograniczać się do relacji «większy», «mniejszy» czy «różny»." Składnia opisu relacji powinna być spójna.
7. Na str. 83 pojawia się zapis „Warunki elementarne budujące W_a można przedstawić w formie a > v lub a <= v, gdzie v jest pewną wartością z dziedziny atrybutu a.". Jak ten zapis odnosi się do przyjętej wcześniej składni?
8. W podsumowaniu (s. 158) Doktorant słusznie zauważa „w szczególności nie zostało wykonane porównanie jakości modeli akcyjnych wygenerowanych przez inne, dostępne metody" – jest to istotne ograniczenie ewaluacji. Nasuwa się pytanie o wskazanie procedury eksperymentalnej, która pozwoliłaby na takie porównanie.
9. W pracy nie wskazano, czy opracowane algorytmy były używane na innych danych niż te pochodzące ze wskazanego w rozdziale 5 repozytoriom UCI. Innymi słowy, w jakim stopniu Autor miał okazję do użycia wyników rozprawy do rozwiązania rzeczywistych problemów.
10. Powtarzają się drobne (i często występujące w wielu rozprawach) usterki redakcyjne dotyczące błędów łamania – występują tzw. sieroty i bękarty – oraz używania dywizu zamiast pauzy.
11. Występuje szereg braków i usterek w składzie bibliografii, np. w pozycjach 12,14, 20, 21, 22, 43, 58, 72, 75, 78, 105.

Powyższe słabe strony rozprawy i uwagi krytyczne nie podważają mojej pozytywnej oceny rozprawy.

6 Działalność publikacyjna

W podsumowaniu (rozdział 6.) Autor rozprawy wskazuje, że wczesne wersje opisywanych w rozprawie algorytmów zostały przedstawione w dwóch artykułach. Pierwszy z nich to artykuł [57] na konferencji „BDAS 2018: Beyond Databases, Architectures and Structures. Facing the Challenges of Data Proliferation and Growing Variety", oraz w artykule [58] na konferencji „ISCIS 2018: Computer and Information Sciences", materiały obu konferencji były publikowane w tomach w serii Springer CCIS. Doktorant jest pierwszym autorem obu publikacji. Algorytm rekomendacji i sposób weryfikacji modeli akcyjnych opisane zostały w artykule [83] w czasopiśmie *Information Sciences* (200pkt MNiE) opublikowanym w roku 2022. Doktorant jest drugim Autorem pracy, a w podsumowaniu rozprawy nie opisuje szczegółowo swojego do niej wkładu merytorycznego. Wszystkie artykuły są opublikowane z promotorami rozprawy.

W bazie DBLP można ponadto znaleźć jeszcze jeden artykuł konferencyjny z roku 2017, którego drugim autorem jest Doktorant, oraz wcześniejszą wersje pracy [83] zamieszczoną w serwisie ArXiv.org. Doktorant nie ma założonego profilu w serwisie Google Scholar.

W mojej opinii, mając na uwadze okres pracy nad rozprawą, aktywność publikacyjna nie jest duża.

7 Podsumowanie i wniosek końcowy

Recenzowana rozprawa doktorska zawiera szereg oryginalnych wyników, w aktywnie rozwijanym na świecie i ważnym dla AI i informatyki obszarze eksploracji danych. Autorowi z powodzeniem udało się zrealizować postawione cele badawcze oraz wykazać prawdziwość sformułowanej w rozprawie tezy. Doktorant wykazał się umiejętnościami opisu problemów koncepcyjnych, w tym formułowania algorytmów, a także praktycznymi umiejętnościami eksperymentalnej ewaluacji opracowanych przez siebie metod.

Podsumowując, stwierdzam, że recenzowana przez mnie rozprawa spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i wnoszę o dopuszczenie Pana mgra Pawła Matyszoka do dalszych etapów przewodu doktorskiego, w tym publicznej obrony.