Warszawa, 29.3.2014

prof. dr hab. inż. Mieczysław A. Kłopotek
profesor zwyczajny PAN
Instytut Podstaw informatyki
Polskiej Akademii Nauk
w Warszawie

Recenzja rozprawy doktorskiej mgr. Tomasza Jacha
pt.: "Optymalizacja procesów wnioskowania z wiedzą niepełną"

1. Uwagi wstępne

Niniejsza recenzja powstała na zlecenie Rady Wydziału Informatyki i Nauk o Materiałach Uniwersytetu Śląskiego na podstawie uchwały Rady z dnia 10.12.2013.

Przedmiotowa rozprawa dotyczy obecnego w informatyce od dość dawna zagadnienia wnioskowania regułowego w systemach ekspertowych, w szczególności kwestii szybkości tegoż wnioskowania w bazach wiedzy liczących tysiące reguł wnioskowania, obarczonych potencjalnie niepewnością. Autor stawia tezę, iż przyspieszenie tego procesu można osiągnąć poprzez analizę skupień reguł oraz uwzględnienie niepełności wiedzy.

Przedłożona praca składa się z 9 rozdziałów (o objętości ok.190 stron) oraz bibliografii obejmującej 170 pozycji, w tym 15 prac ze współautorstwem mgr. Jach.

Rozdział pierwszy wprowadza w tematykę rozprawy. Drugi przedstawia rys historyczny rozwoju systemów wspomagania decyzji (SWD) pozycjonując hierarchiczną reprezentację wiedzy, która jest przedmiotem zainteresowania kandydata. Rozdział trzeci lokuje zagadnienie reprezentacji wiedzy w kontekście metod wnioskowania w SWD. W czwartym rozdziale omówiono wybrane metody reprezentacji wiedzy ze wskazaniem na korzyści wynikające z użycia

współczynników CF, które stanowią motywację dla wykorzystywanych przez Autora współczynników IF.

Rozdział piaty jest poświęcony pozyskiwaniu hierarchicznej aranżacji reguł w bazie wiedzy poprzez analizę skupień oraz wykorzystaniu tejże do szybkiego wyszukiwania reguł, które można aktywować przy zadanym stanie bazy faktów.

Rozdział szósty omawia sposoby oceny efektywności metod konstrukcji hierarchicznych baz regułowych w wymiarze jakości skupień, jakości procesu wnioskowania w SWD oraz złożoności obliczeniowej wnioskowania.

Rozdział siódmy prezentuje system generacji hierarchicznej bazy wiedzy i jej oceny autorstwa kandydata, zaś rozdział ósmy opisuje wyniki badań eksperymentalnych nad metodą kandydata.

Rozdział dziewiąty stanowi podsumowanie rozprawy.

Za własny wkład Autora należy uznać rozdziały 5.3 (metodyka grupowania reguł), rozdziały 7 i 8 oraz częściowo 6.3 (analiza złożoności algorytmów autorstwa kandydata).

2. Dyskusja

Przedłożona praca ma charakter konceptualno-eksperymentalny. Bazuje na zdawałoby się oczywistym pomyśle skrócenia liniowego czasu poszukiwania reguł poprzez ich hierarchiczną organizację. Jednakże szczegóły już nie są tak oczywiste, a wszelkie pomysły w tym zakresie wymagają eksperymentalnej walidacji, co widać na przykładzie przeprowadzonych eksperymentów, gdzie np. 2 spośród 4 rozważanych metod wyboru tzw. reprezentanta grupy kompletnie się nie sprawdziły. Dlatego też za podstawowy wkład naukowy przedłożonej rozprawy należy uznać – obok zaproponowanego algorytmu mAHC, metody współczynników pewności IF, czy różnych szczegółowych rozwiązań algorytmu grupującego jak dobór miar podobieństwa, dobór punktu odcięcia itd. – wieloaspektowe badanie proponowanych rozwiązań i empiryczny dobór wybranych parametrów algorytmu.

Do podstawowych problemów, z którymi musiał zmierzyć się Autor, jest zagadnienie ryzyka pominięcia stosowalnych reguł, indukowanego przez hierarchiczne wnioskowanie czy też wnioskowanie oparte o reprezentanta. Takie pominięcie może być wynikiem właśnie procesu analizy skupień i/lub nieadekwatności opisu skupień, czy to płaskich czy to hierarchicznych. Autor proponuje w tym miejscu niestandardowe rozwiązanie, polegające na potraktowaniu takiego wydarzenia jako przypadku dla wnioskowania z wiedzą niepewną, i takie wnioskowanie stosuje w zaimplementowanym przez siebie systemie. Czy jest to rozwiązanie słuszne? Autor nie podjął się dyskusji tego tematu – a szkoda. Wydaje mi się bowiem, że jeśli proces grupowania jest przeprowadzony w miarę poprawnie, to zdarzenia pominięcia reguły mogą być może świadczyć o pewnej sporadyczności wygenerowanej reguły, nie pasowania jej do ogólnego wydźwięku pozostałej bazy wiedzy. Ale taka hipoteza wymagałaby naturalnie stosownych badań. Należy jednakże w tym miejscu podkreślić, iż Autor przeprowadził stosowne badania stopnia precyzji wygenerowanego systemu reguł w porównaniu z abstrakcyjną metodą dokładną.

Osobno należy zauważyć fakt, iż praca prezentuje dość obszerny przegląd metod wnioskowania eksperckiego, metod reprezentacji i akwizycji wiedzy, rozwoju tych metod, a także wykorzysujących je systemów. Potwierdza to niewątpliwie, iż dokonane przez Autora wybory i zaproponowane metody poprzedzone były wnikliwą analizą aktualnego stanu wiedzy.

Tym nie mniej mym obowiązkiem jako recenzenta jest wskazanie na dość liczne niedociągnięcia przy redakcji pracy, szczególnie w wymiarze stosowania formalizmów oraz projektu eksperymentów.

I tak na stronie 8 w 4 linii od dołu Autor winien doprecyzować, że chodzi mu o iloczyn *logiczny* deskryptorów.

Na stronie 14 w 6 linii od dołu autor wprowadza funkcję informacji f nie precyzując, co miałaby ona obliczać. Gdyby autor ograniczył się do deskryptorów atrybut=wartość w regułach, nie budziłaby ona mych wątpliwości. Natomiast dopuszczenie relacji mniejszości, większości itd. Oznacza możliwość wystąpień ia

tego samego atrybutu z innymi wartościami, co czyni funkcję niejednoznaczną (czyli przestaje to być funkcja).

Na stronie 26 Autor ponownie definiuje funkcję informacyjną f, tym razem myląc argumenty - winno być f(x,a), a nie f(a,x).

Na stronie 28, opisując tabelę 2.4 Autor pisze w sposób nieprecyzyjny (2 linia) "Jednocześnie zostały usunięte atrybuty rozróżniające obiekty wchodzące w skład tej samej klasy decyzyjnej". W tabeli 2.4 atrybut k został usunięty tylko w niektórych miejscach, a nie w całej tabeli.

Na stronie 32 w definicji mSWD Autor pomija definicję f_sim, a winien podać, czym różni się ten symbol od F_sim.

Na stronie 60 Autor pisze kilkukrotnie o "prawdopodobieństwie" w teorii Dempstera-Shafera. Tymczasem kilku autorów niezależnie wykazało, że w trakcie wnioskowania Bel i Pl nie zachowują się jak (granice przedziałów) prawdopodobieństwa. Ich zachowanie jest bliższe wnioskowaniu w teorii zbiorów przybliżonych Pawlaka.

Na stronie 110 Autor definiuje pojęcie podobieństwa jako liczby z zakresu [0,1]. Tymczasem on sam na stronie 119 wprowadza pojęcie "prostego podobieństwa" (simpleSimilarity), które ten warunek narusza. Ponadto na stronie 110 podaje przykład funkcji podobieństwa jako odwrotności odległości, co jest w sumie pojęciem źle zdefiniowanym, bo dla tego samego obiektu (d(x,x)=0) podobieństwo sięgnie nieskończoności.

Na stronie 111 Autor nieprecyzyjnie definiuje odległość. Powinno być: d(x,y)=0 wtedy i tylko wtedy gdy x=y. .

Na stronie 120, gdzie Autor opisuje swą metodę grupowania skupień, nie precyzuje on, jakie kryterium odległości między skupieniami stosuje. Pisze o "wcześniej omówionych". Czy są to wszystkie metody z np. tablicy 5.7? Wtedy jest problem braku precyzji, gdyż tylko dla części określił on sposób korzystania z podobieństwa (de facto tylko dla dwóch pierwszych pozycji).

Na stronie 123 widzimy dość kontrowersyjną metodę obliczania optymalnej liczby skupień T. Jeśli wykorzystamy simpleSimilarity ze strony 119, to T będzie liczbą

mieszaną - dla odtwarzalności badań z rozdziału 8 musielibyśmy wiedzieć, czy np. zaokrąglano w górę, w dół, czy w inny sposób. Dla weightedSimilarity ze strony 119 tak określone T traci sens, bo weightedSimilarity jest z zakresu [0,1], więc T wyniesie albo 0 albo 1, czyli nie ma różnicy mAHC i AHC w tym przypadku. Brakuje mi również rozumowego uzasadnienia doboru T z zakresu 1-simpleSimilarity.

Najpóźniej w punkcie 5.3.4 Autor winien zdefiniować, w jaki sposób opisuje zestawem deskryptorów grupy reguł, gdyż jest to konieczne do rozumienia omawianej tu funkcji podobieństwa węzłów.

W Algorytmie 8 na stronie 128 w linii z "if" brakuje nawiasu zamykającego po f_sim(W, a generalnie f_sim jest tu funkcją jednoargumentową, podczas gdy w tekście poprzedzającym ma dwa argumenty.

Na stronie 129 we wzorze na IF nie podano, co znaczy małe d_i, a dokładniej rzecz ujmując wyjaśnienie (i-ta przesłanka) ma się nijak do podanego dalej przykładu (str.130), z którego można się domyślać, że właściwie to chodzi o regułę R_i, której przesłanką D_i jest pewien zbiór deskryptorów.

Na stronie 167 w pkt.3 Autor pisze o "obu " analizowanych typach reprezentantów, opisanych w rozdziale 5.2.2. Ale w 5.2.2. rozważa się cztery typy, wobec tego nie bardzo wiadomo, o które chodzi. Dopiero w rozdziale nt. eksperymentów Autor nadmienia, że z tych czterech typów wybrano dwa.

W punkcie 7 na tejże stronie autor odwołuje się do definicji miar podobieństwa grup reguł z rozdziału 5.3.4. dla 3 pierwszych miar z tabeli 5.7 ze strony 120. Ale jak wspomnieliśmy, brak jest tam definicji miary podobieństwa dla trzeciej z tych miar (podano tam tylko odległość), co oznacza, że metoda średniego podobieństwa nie jest w pracy zdefiniowana.

Czy na stronie 168 w 6 linii od dołu należy rozumieć, że system sam poszukuje parametru t_prec, a jeśli tak, to jaką metodą? Jeśli nie, to na czym polega tu optymalizacja?

Na początku rozdziału 8 pada informacja, że reguły, przetwarzane w ramach eksperymentów, były wygenerowane systemem LEM2. O ile dobrze pamiętam,

system ten generował reguły jednopoziomowe (wnioskowanie bez tworzenia pośrednich faktów). Czy tylko takie były wykorzystywane w badaniach? Trochę też smuci fakt, że tylko dwie spośród rozważanych baz reguł posiadały ponad 1000 reguł.

Na stronie 176 Autor uzasadnia wybór sposobu ustalania reprezentanta grupy. Jest tam niepokojące stwierdzenie, że cechy dominujące są zbyt podobne między grupami. Czy nie jest to wskazówka, że metoda grupowania nie działa najlepiej?

Na stronie 176 w 5 linii od dołu Autor nadmienia, iż w tabeli 8.2 podano maksymalne uzyskane wyniki. Dobrze byłoby wiedzieć, dla jakiej kombinacji "metod wyznaczania macierzy nierozróżnialności i kryteriów łączenia skupień" je uzyskano.

W opisie testów na stronie 177 brakuje mi wskazania, w jaki sposób wybierano fakty do bazy faktów. Czy testowano wszystkie możliwości, czy je losowano, a jeśli tak, to w jaki sposób?

Tytuł rozdziału 8.2 wprowadza w błąd. Autor bazuje na miarach podobieństwa a nie odległości.

Na str. 179 w linii 11 tekst jest niezrozumiały. Chyba Autor miał na myśli "niż" pisząc "dla".

Na str.185 w 6 linii od dołu Autor pisze, że wyniki dla hierarchii były "minimalnie" gorsze od wyników dla metody z reprezentantem. Patrząc na rysunek 8.5 odnoszę wrażenie, że różnica jest dość radykalna. Być może problem polega na tym, że autor nigdzie nie podaje miar rozrzutu dla prezentowanych wyników.

Przestawione powyżej uwagi nie zmieniają faktu, iż autor opracował i zbadał ciekawe podejście do przyspieszania wyszukiwania reguł w bazie wiedzy i wykonał inspirujące eksperymenty. W/w uwagi winny z jednej strony posłużyć Autorowi do opracowania erraty tam, gdzie widać ewidentne błędy, a z drugiej strony przygotować się do dyskusji kwestii, które wydają się być otwarte.

Uwaga edytorska: po stronie 33 i w kilku innych miejscach występują puste kartki.

3. Wnioski

Wobec wykazania słuszności stawianej w pracy tezy, mimo wspomnianych niedociągnięć, stwierdzam, że **przedłożona praca spełnia wymogi formalne i zwyczajowe stawiane pracom doktorskim i wobec tego wnoszę o dopuszczenie Kandydata do dalszych etapów przewodu.**