STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Agnieszka NOWAK-BRZEZIŃSKA
Uniwersytet Śląski, Instytut Informatyki

EKSPLORACJA WIEDZY A EFEKTYWNOŚĆ SYSTEMÓW WSPOMAGANIA DECYZJI

Streszczenie. Artykuł przedstawia wyniki analizy efektywności systemów wspomagania decyzji, a zwłaszcza wpływu metod eksploracji wiedzy na tę efektywność. Okazuje się, że metody analizy skupień pozwalają w sposób znaczący tę efektywność poprawić. W artykule można znaleźć porównanie efektywności procesów wnioskowania, realizowanych w przypadkach baz wiedzy o różnej reprezentacji: klasycznej oraz wzbogaconej analizą skupień bądź tzw. częściowymi regułami decyzyjnymi. Zbadano także skuteczność różnych miar podobieństwa (względnie odległości) w kontekście ich wpływu na efektywność systemów wspomagania decyzji, będących przedmiotem analizy niniejszego artykułu.

Słowa kluczowe: analiza skupień, regułowe bazy wiedzy, częściowe reguły decyzyjne, efektywność procesów wnioskowania

DATA MINING AND AN EFFICIENCY OF DECISION SUPPORT SYSTEMS

Summary. The paper presents the results of the experiments in which an efficiency of decision support systems is analyzed. During the research it was noted that the most influenced on such efficiency is the method of representation of knowledge base. Following approaches are considered due to experiments: rules knowledge base with clusters of decision rules, rules knowledge base with partial decision rules and rules knowledge base with clusters of partial decision rules. The efficiency of choosing the best similarity measure is also analyzed.

Keywords: cluster analysis, rules knowledge bases, efficiency, partial decision rules

1. Wprowadzenie

Systemy wspomagania decyzji, a zwłaszcza ich efektywność są celem licznych badań naukowych, realizowanych w obszarze sztucznej inteligencji. Efektywność klasycznego systemu wspomagania decyzji zależy od tego, jaka jest efektywność procesów wnioskowania, które stanowią tzw. rdzeń każdego takiego systemu. Proces wnioskowania zależy od przyjętej metody wnioskowania i daje się opisać następująco. W przypadku wnioskowania sterowanego danymi (ang. *forward chaining*) zbiór wszystkich reguł jest przeszukiwany w celu znalezienia reguły bądź podzbioru reguł do uaktywnienia. Zakładamy, że każda reguła decyzyjna jest wyrażeniem implikacji, której poprzednik i następnik są złożeniem logicznym dwójek <atrybut, wartość>. Reguły takie mogą być wprost podane przez eksperta dziedzinowego bądź wygenerowane automatycznie, stosując jeden z możliwych algorytmów indukcji reguł. W procesie wnioskowania uaktywniana jest reguła, której albo wszystkie przesłanki są prawdziwe, a efektem jest wygenerowanie nowej wiedzy w systemie (konkluzja reguły uaktywnionej staje się nowym faktem w bazie wiedzy), albo (gdy wybierzemy metodę wnioskowania sterowanego celem (ang. *backward chaining*)) znalezienie i następnie uaktywnienie reguły bądź reguł, których konkluzja pokrywa się z celem wnioskowania. Wówczas – jeśli wszystkie przesłanki danej reguły uznane zostaną za prawdziwe (będą faktami w bazie wiedzy) – cel wnioskowania zostanie potwierdzony i wyprowadzony jako nowa wiedzy w systemie. Jeśli jednak przynajmniej jedna z przesłanek takiej reguły nie będzie faktem w bazie wiedzy, wówczas staje się ona podcelem wnioskowania i dla jej potwierdzenia proces wnioskowania uruchamiany jest tak samo, jak dla hipotezy głównej. Jak można zatem wywnioskować z powyższego opisu, efektywność procesów wnioskowania będzie wysoka wtedy, gdy możliwe będzie uaktywnianie reguł, a więc jak najwięcej przesłanek będzie dało się pokryć faktami w bazie wiedzy. Dodatkowo z pewnością na tę efektywność będzie wpływać rozmiar takiej bazy wiedzy, a więc liczba reguł w bazie wiedzy i struktura reprezentacji reguł. Klasyczne metody wnioskowania okazują się nieefektywne ze względu na zbyt długi czas pracy oraz zbyt dużo generowanych zbędnych informacji [4]. Konieczne wydaje się zastosowanie modyfikacji formy dziedzinowej bazy wiedzy, a przez to także i procesów wnioskowania, które ograniczą te niedogodności, nie odbierając jednocześnie żadnej ze znanych zalet tej reprezentacji wiedzy. Z tego względu w artykule zostaną przeanalizowane następujące metody reprezentacji reguł decyzyjnych w bazach wiedzy:

1) regułowa baza wiedzy ze skupieniami reguł decyzyjnych,
2) regułowa baza wiedzy z częściowymi regułami decyzyjnymi,
3) regułowa baza wiedzy ze skupieniami częściowych reguł decyzyjnych.

2. Baza wiedzy ze skupieniami (pełnych) reguł decyzyjnych

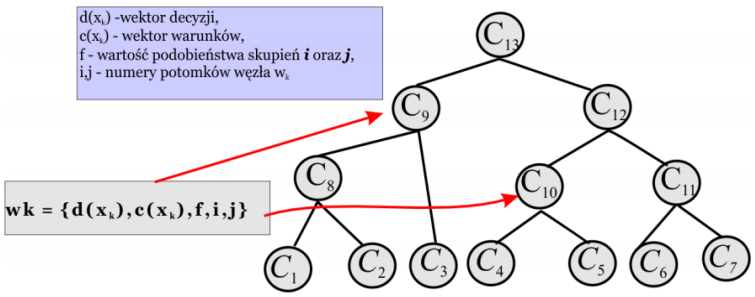
Rozmiar bazy wiedzy w znaczący sposób wpływa na czas wnioskowania. Z tego względu potrzebne zdają się być takie modyfikacje istniejącej struktury dziedzinowej bazy wiedzy, które ograniczą liczbę reguł, analizowaną w procesie wnioskowania, w celu znalezienia reguł do uaktywnienia. W pracy [9] autorka przedstawiła model tzw. złożonej bazy wiedzy (ang. *composited knowledge base*), w którym to modelu zbiór reguł jest prezentowany w postaci struktury skupień reguł podobnych do siebie bądź to po części przesłankowej, bądź konkluzyjnej. Analiza skupień (ang. *cluster analysis*) pozwala na grupowanie reguł podobnych w skupienia [1,2]. Spośród wielu możliwych do zastosowania algorytmów analizy skupień, algorytmy aglomeracyjne z grupy algorytmów hierarchicznych pozwalają na naturalne łączenie w grupy tych obiektów (reguł), które są do siebie najbardziej podobne w danym kroku algorytmu. Utworzona hierarchia skupień mówi nam dodatkowo jak bardzo podobne są do siebie połączone elementy (o tym decyduje poziom w hierarchii). Struktura hierarchiczna, którą uzyskać można poprzez zastosowanie algorytmu *AHC,* prezentowanego w pracach [3, 7, 9] oraz *mAHC* w pracy [9] ma formę drzewa binarnego, w którym każdy węzeł niebędący liściem (reprezentujący skupienie reguł) ma zawsze dwóch potomków. Dla takich typów struktur znane są efektywne techniki ich przeszukiwania, pozwalające ograniczyć złożoność obliczeniową z liniowej do logarytmicznej. Dlatego też rozwiązanie to będzie się nadawało do dużych zbiorów danych, im więcej bowiem będzie reguł w bazie wiedzy, tym większa korzyść z zastosowania takiego modelu bazy wiedzy zamiast klasycznej formy.

2.1. Model hierarchicznej bazy wiedzy

Mając dane: X jako zbiór reguł w bazie wiedzy oraz funkcję podobieństwa F_{sim}, związaną z tym zbiorem reguł X, możemy zbudować hierarchicznie zorganizowany model bazy wiedzy. W takim modelu, n-reguł: $X = \{x_1,..,x_n\}$, gdzie każda reguła x_i opisana jest przy użyciu atrybutów ze zbioru A i wartości V ($V = U_{a \in A}$, U_a jest zbiorem wartości atrybutu a) jest reprezentowane przez: (i) funkcję podobieństwa F_{sim}: $X \times X \rightarrow [0..1]$ oraz (ii) etykietowane drzewo binarne Tree = $\{w_1, \ldots, w_{2n-1}\} = \bigcup_{k=1}^{2n-1} w_k$, utworzone przez grupowanie reguł ze zbioru X stosując funkcję podobieństwa (patrz rys. 1). Każdy węzeł w takim drzewie jest reprezentowany przez piątkę $\{d(x_k),c(x_k),f,i,j\}$, gdzie: $c(x_i) \in V_1 \times V_2 \times \ldots \times V_m$ jest wektorem reprezentujących lewą część reguły x_k, a $d(x_i) \in V_1 \times V_2 \times \ldots \times V_m$, wektorem odpowiadającym prawej stronie reguły. Funkcja $f = F_{sim} : X \times X \rightarrow [0..1]$ jest wartością podobieństwa między dwoma łączonymi elementami (regułami bądź ich skupieniami) x_i i x_j. Elementy i oraz j są numerami potomków aktualnie tworzonego skupienia.

2.2. Wnioskowanie dla skupień reguł decyzyjnych

Proces wnioskowania systemu wspomagania decyzji, w którym baza wiedzy to tak naprawdę struktura skupień reguł podobnych do siebie rożni się znacząco od procesu wnioskowania dla klasycznej bazy wiedzy, gdzie bazę wiedzy stanowi lista wszystkich reguł. W proponowanym podejściu ma miejsce przeszukiwanie struktury skupień reguł technikami opisanymi w pracach [5, 6, 9], stosującymi podejście zbliżone do idei przeszukiwania połówkowego o złożoności O(log n). W wyniku takiego przeszukiwania zostanie znaleziona reguła najbardziej podobna do podanych danych wejściowych (faktów bądź hipotezy głównej) i następuje proces uaktywnienia tejże reguły.

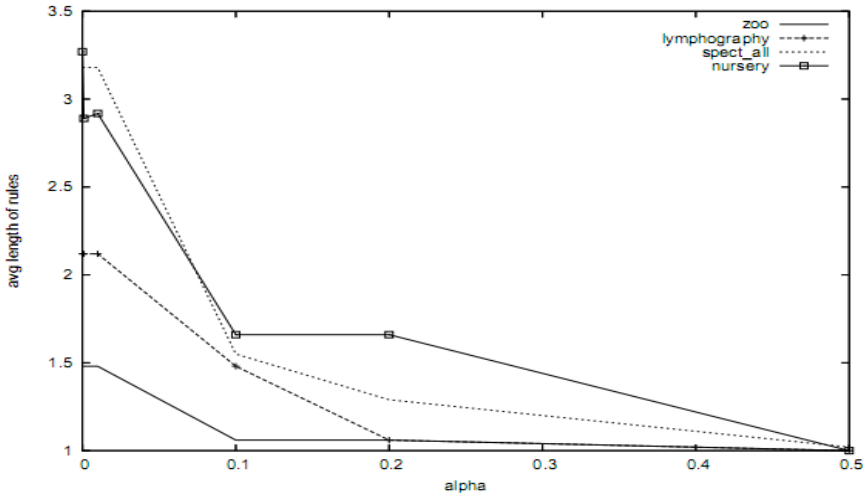


Rys. 1. Model hierarchicznej bazy wiedzy
Fig. 1. Hierarchical model of knowledge base

3. Baza wiedzy z częściowymi regułami decyzyjnymi

Jak już wcześniej wspomniano, na efektywność procesu wnioskowania wpływa złożoność bazy wiedzy powodowana nie tylko przez dużą liczbę reguł, ale także przez długość reguł, a co się z tym wiąże – liczbę przesłanek w regułach. Im więcej przesłanek dana reguła zawiera, tym dłuższy jest proces analizy ich prawdziwości, niezbędny gdy chcemy taką regułę uaktywnić. Warto w tym momencie podkreślić także fakt, że z punktu widzenia użytkownika systemu wspomagania decyzji liczba przesłanek analizowanych powinna być możliwie mała; a zatem preferowane są krótkie reguły. W pracy [11] zaproponowano algorytm generowania tzw. częściowych reguł decyzyjnych (ang. *partial decsion rules*). Dokładne reguły decyzyjne mogą być przeuczone, tzn. zbyt mocno dopasowane do istniejących przykładów danych w zbiorze uczącym lub zbyt mocno zależne od szumu informacyjnego. Natomiast częściowe reguły decyzyjne zawierają mniej atrybutów i te, które oddzielają od danego wiersza prawie wszystkie inne wiersze z inną decyzją. Poziomem pokrycia zbioru steruje użytkownik, a sam algorytm opiera się na strategii zachłannej, by wygenerować takie reguły. Jak się można domyślać, baza wiedzy zawierająca takie reguły będzie analizowana znacznie szybciej, a szansa na uaktywnienie większej liczby reguł niż dla reguł w klasycznej bazie wiedzy jest tu spora (mniej przesłanek w regule to

mniej warunków, które muszą być spełnione, by regułę można było uaktywnić). Rysunek 2 przedstawia prostą zależności między długością reguły a wartością parametru α, sterującego algorytmem generowania częściowych reguł decyzyjnych. Łatwo zauważyć, że im większa jest wartość parametru α tym krótsze są reguły. Zwiększenie wartości tego parametru powoduje, że zmniejszamy wymóg pokrycia zbioru reguł, a więc zezwalamy na to, że pewien zbiór reguł ma przypisaną daną decyzję przy czym zakładamy, że z reguły w takim zbiorze tylko w (1- α)% ta decyzja jest spełniona. Analiza prac [7,8] pozwala zauważyć, że np. w przypadku zbioru *zoo.kb*, w którym każda reguła (obiekt w zbiorze) ma długość 16 deskryptorów, liczba przesłanek przy zastosowaniu algorytmu generowania częściowych reguł decyzyjnych zostaje ograniczona do 4 deskryptorów, w pesymistycznym przypadku i w sytuacji gdy parametr α jest bliski 0 (a nawet do 2, gdy wartość α jest równa 0.2 czy 0.5). W przypadku zbioru *spect_all.kb* skraca się liczba przesłanek w regułach z 23 do nawet 5.



Rys. 2. Parametr α a długość reguły
Fig. 2. The α-parameter and a rule length

3.1. Wnioskowanie dla częściowych reguł decyzyjnych

Proces wnioskowania dla bazy wiedzy z częściowymi regułami decyzyjnymi tym się różni od wnioskowania w klasycznej bazie wiedzy, że każda reguła ma po prostu znacznie mniejszą liczbę warunków. Z jednej strony skraca to czas potrzebny na analizę każdej reguły, a z drugiej sprawia, że więcej reguł można wówczas uaktywnić, a więc większe są szanse na wyprowadzenie nowej wiedzy z takiego systemu, tym samym większe są szanse zakończenia procesu wnioskowania sukcesem. Zakładając, że parametr τ_p określa czas sprawdzania prawdziwości przesłanki p danej reguły w bazie wiedzy, porównanie klasycznego podejścia oraz wnioskowania dla częściowych reguł decyzyjnych przedstawia tabela 1. Dla każdego analizowanego zbioru (spośród trzech zbiorów: *zoo*, *spect_all* oraz *lymphography*) przedstawiono następujące parametry: liczba atrybutów w regule (n_{atr}), maksymalna długość reguły (rl_{max}),

wartość parametru α (α), liczba reguł w bazie wiedzy (N_R), czas wnioskowania dla bazy wiedzy z częściowymi regułami decyzyjnymi (τ_α) oraz czas wnioskowania dla klasycznej bazy wiedzy (τ).

Tabela 1

Czasy wnioskowania dla klasycznego podejścia oraz częściowych reguł decyzyjnych

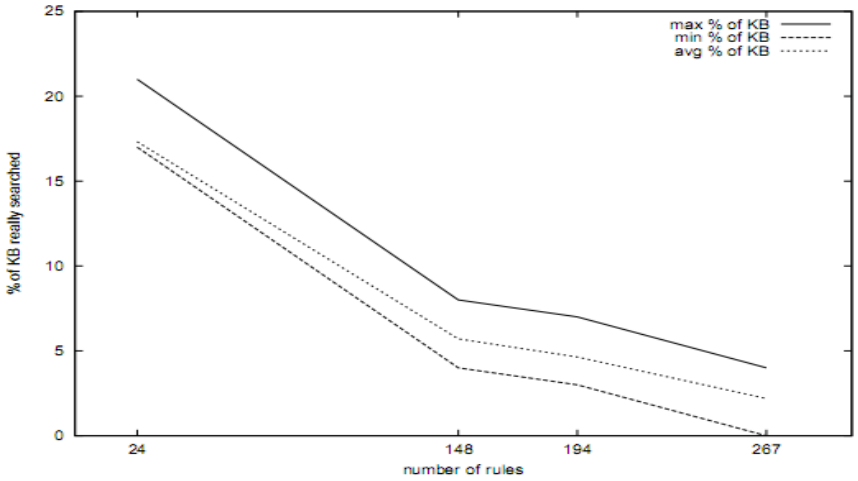
set	$\mathbf{n_{atr}}$	$\mathbf{rl_{max}}$	α	$\mathbf{N_R}$	τ_α	τ
Zoo	16+1	4	0	101	$101*4*\tau_p$	$101*\mathbf{16}*\tau_p$
Zoo	16+1	4	0.001	101	$101*4*\tau_p$	$101*\mathbf{16}*\tau_p$
Zoo	16+1	4	0.01	101	$101*4*\tau_p$	$101*\mathbf{16}*\tau_p$
Zoo	16+1	2	0.1	101	$101*2*\tau_p$	$101*\mathbf{16}*\tau_p$
Zoo	16+1	2	0.2	101	$101*2*\tau_p$	$101*\mathbf{16}*\tau_p$
Zoo	16+1	1	0.5	101	$101*\mathbf{1}*\tau_p$	$101*\mathbf{16}*\tau_p$
lympho	18+1	4	0.01	148	$148*4*\tau_p$	$148*\mathbf{18}*\tau_p$
lympho	18+1	2	0.1	148	$148*2*\tau_p$	$148*\mathbf{18}*\tau_p$
lympho	18+1	2	0.2	148	$148*2*\tau_p$	$148*\mathbf{18}*\tau_p$
lympho	18+1	1	0.5	148	$148*\mathbf{1}*\tau_p$	$148*\mathbf{18}*\tau_p$
spect all	23+1	10	0	267	$267*10*\tau_p$	$267*\mathbf{23}*\tau_p$
spect all	23+1	10	0.001	267	$267*10*\tau_p$	$267*\mathbf{23}*\tau_p$
spect all	23+1	10	0.01	267	$267*10*\tau_p$	$267*\mathbf{23}*\tau_p$
spect all	23+1	7	0.1	267	$267*7*\tau_p$	$267*\mathbf{23}*\tau_p$
spect all	23+1	5	0.2	267	$267*5*\tau_p$	$267*\mathbf{23}*\tau_p$
spect all	23+1	2	0.5	267	$267*2*\tau_p$	$267*\mathbf{23}*\tau_p$

Analizując pierwszy wiersz, a więc zbiór *zoo.kb* z parametrem α=0, można łatwo zauważyć, że zastosowanie częściowych reguł decyzyjnych zamiast pełnych reguł pozwala skrócić czas wnioskowania o 75% ($101*4*\tau_p$ zamiast $101*16*\tau_p$). Oczywiście zysk ten zmienia się w zależności od zastosowanego zbioru, raz zyskujemy więcej, czasem mniej – generalnie zysk jest z przedziału od 8% do 43%. Oczywiście nadal ma miejsce przegląd zupełny całej bazy wiedzy w procesie wnioskowania, różnica polega jednak na tym, że proces analizy każdej reguły jest szybszy niż ma to miejsce w metodzie klasycznej.

4. Baza wiedzy ze skupieniami częściowych reguł decyzyjnych

Gdy baza wiedzy ma strukturę skupień reguł podobnych do siebie, w procesie wnioskowania wystarczy przejrzeć tylko grupę o największym stopniu podobieństwa ze zbiorem podanych faktów. Czas wnioskowania zostaje znacznie zmniejszony (kilkudziesięciokrotnie), nie traci przy tym na efektywności (kompletności i dokładności) wnioskowania. Zauważono, że przy dużych wektorach (długich regułach) niektóre miary błędnie szacują podobieństwo między regułami. Na przykład miara cosinus preferuje długie reguły decyzyjne, lecz inne miary zachowują się odwrotnie. Przy długich regułach zdarza się „zabłądzić" w strukturze skupień reguł i do uaktywnienia wybrać regułę nierelewantną. Zależność między rozmiarem bazy wiedzy a procentem bazy wiedzy faktycznie przeglądanym przedstawia rys. 3. Łatwo

zauważyć, że im większa jest baza wiedzy, tym mniejszy jest obszar bazy wiedzy, który jest faktycznie analizowany w procesie wnioskowania.



Rys. 3. Zależność między liczbą reguł a % BW faktycznie przeglądanym
Fig. 3. Rules' length and a % of knowledge based searched

4.1. Wnioskowanie dla skupień częściowych reguł decyzyjnych

Częściowe reguły decyzyjne podlegające grupowaniu utworzą skupienia o krótkim opisie, łatwo interpretowalne. W procesie wnioskowania niwelujemy problem preferowania przez niektóre miary reguł długich o tych samych przesłankach co zbiór podanych faktów. Czas wnioskowania znacznie się skraca (reguły są krótsze, a wiec nie tylko czas wyszukiwania reguł, ale i wnioskowania (uaktywnienia wybranych reguł) się zmniejsza). Tabela 2 przedstawia wyniki analizy tych samych, co poprzednio zbiorów danych, gdy częściowe reguły decyzyjne zostaną zgrupowane algorytmem AHC, omówionym już w rozdziale 2 niniejszego artykułu.

Tabela 2
Parametry wnioskowania dla skupień częściowych reguł decyzyjnych

set	rl_{avg}	N_R	N_n	N_{data}	N_{rr}	N_{Rrs}	Precision	%KB
spect_all_0	3.18	267	16	1	38	133	0	3%
spect_all_0	3.18	267	12	9	6	218	1	2.25%
spect_all_0	3.18	267	6	9	4	116	**0**	1.12%
spect_all_001	3.18	267	24	1	38	195	**0**	4.5%
spect_all_001	3.18	267	6	2	11	260	**0**	1.12%
spect_all_001	3.18	267	6	3	6	260	**0**	1.12%
spect_all_001	3.18	267	24	9	19	264	**0**	4.5%
spect_all_001	3.18	267	12	9	6	218	1	2.25%
Zoo_0	1.48	101	20	1	41	45	1	10.05%
Zoo_0	1.48	101	14	2	2	98	1	7%
Zoo_0	1.48	101	12	3	1	90	1	6%
Zoo_001	1.48	101	20	1	41	45	1	10%
Zoo_001	1.48	101	10	2	4	40	1	5%
Zoo_001	1.48	101	12	4	1	52	1	6%

Dla każdego analizowanego zbioru (spośród trzech zbiorów: *zoo*, *spect_all* oraz *lymphography*) przedstawiono następujące parametry: średnia długość reguły (rl_{avg}), liczba reguł w bazie wiedzy (N_R), liczba węzłów (skupień utworzonych w bazie wiedzy) (N_n), liczba danych wprowadzonych jako fakty w procesie wnioskowania (N_{data}), liczba reguł relewantnych względem podanego zbioru faktów (N_{rr}), liczba reguł faktycznie przeglądanych (N_{Rrs}), wartość parametru dokładności (Precision) oraz % BW faktycznie przeglądany w procesie wnioskowania (%KB).

5. Analiza efektywności systemów wspomagania decyzji stosujących metody eksploracji wiedzy – eksperymenty

Tabela 3

Parametry wnioskowania dla proponowanych koncepcji baz wiedzy

zbiór	n_{atr}	rl_{avg}	rl_{min}	rl_{max}	α	N_R	N_n	D	%KB	τ_α	$(\tau_{AS+\alpha})$	τ
Zoo	16+1	1.48	1	4	0	101	12	1	6.03%	$12*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
Zoo	16+1	1.48	1	4	0.001	101	10	1	5.03%	$10*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
Zoo	16+1	1.48	1	4	0.01	101	12	1	6.03%	$12*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
Zoo	16+1	1.059	1	2	0.1	101	8	1	4.02%	$8*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
Zoo	16+1	1.059	1	2	0.2	101	10	1	5.03%	$10*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
Zoo	16+1	1	1	1	0.5	101	14	1	7.03%	$14*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
Lympho	18+1	2.12	1	4	0.01	148	12	1	4.07%	$12*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
lympho	18+1	1.48	1	2	0.1	148	18	1	6.1%	$18*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
lympho	18+1	1.06	1	2	0.2	148	14	1	4.75%	$14*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
lympho	18+1	1	1	1	0.5	148	26	1	8.81%	$26*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
spect all	23+1	3.18	1	10	0	267	12	1	1.13%	$12*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
spect all	23+1	3.18	1	10	0.001	267	12	1	1.5%	$12*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
spect all	23+1	3.18	1	10	0.01	267	8	1	1.88%	$8*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
spect all	23+1	1.55	1	7	0.1	267	26	1	3%	$26*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
spect all	23+1	1.29	1	5	0.2	267	16	1	1.5%	$16*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$
spect all	23+1	1.022	1	2	0.5	267	8	1	1%	$8*1.48*\tau_p$	$101*1.48*\tau_p$	$101*4*\tau_p$

W eksperymentach zastosowano następujące zbiory danych z repozytorium [12]: *Shuttle, Zoo, Lenses, Lymphography, Monks*(test set)*, Monks*(training set)*, Balance, Soybean* oraz *Spect_all*. Ze względu na ograniczenia niniejszego artykułu przedstawione zostaną wyniki tylko dla zbiorów*: zoo, spect_all* oraz *Lymphography*. Warto jednak w tym miejscu zaznaczyć, że wyniki uzyskane dla pominiętych zbiorów pokrywają się z tymi, przedstawionymi w pracy [12]. Analizie podlegają: (i) czas wnioskowania w klasycznym systemie wspomagania decyzji (τ), (ii) czas wnioskowania w systemie wspomagania decyzji, z zastosowaniem częściowych reguł decyzyjnych (τ_α), (iii) czas wnioskowania dla skupień częściowych reguł decyzyjnych ($\tau_{AS+\alpha}$). Tabela 3 przedstawia wyniki eksperymentów. Dla każdego zbioru możliwe jest porównanie następujących parametrów: liczba atrybutów w regule (n_{atr}), średnia długość reguły (rl_{avg}), minimalna długość reguły (rl_{min}) czy maksymalna długość reguły (rl_{max}),

wartość parametru α (α), liczba reguł w bazie wiedzy (N_R), liczba węzłów (skupień utworzonych w bazie wiedzy) (N_n), wartość parametru dokładności (D), rozumianego jako zdolność systemu do nieuaktywnienia reguły nierelewantnej, % BW faktycznie przeglądany w procesie wnioskowania (%KB), czas wnioskowania dla bazy wiedzy z częściowymi regułami decyzyjnymi (τ_α), czas wnioskowania dla bazy wiedzy ze skupieniami częściowych reguł decyzyjnych ($\tau_{AS+\alpha}$) oraz czas wnioskowania dla klasycznej bazy wiedzy (τ), tp zaś to czas sprawdzania prawdziwości przesłanki *p* danej reguły w BW.

Najważniejsze w analizie wydaje się być to, w jakim stopniu udaje się skrócić czas wnioskowania, gdy zamiast klasycznej bazy wiedzy zastosujemy którąś z proponowanych metod eksploracji wiedzy: skupienia reguł decyzyjnych, częściowe reguły decyzyjne bądź skupienia częściowych reguł decyzyjnych. Wyniki takiej analizy przedstawia tabela 4. Weźmy pod uwagę np. wiersz pierwszy, z którego wynika, że dla zbioru *zoo* z częściowymi regułami decyzyjnymi (gdy średnia długość reguły zakładała 4 atrybuty zamiast przeglądać wszystkie 101 reguł (gdzie w klasycznej bazie wiedzy każda reguła ma 16 przesłanek)) skracamy czas potrzebny na analizę prawdziwości przesłanek o 75%, gdy zastosujemy skupienia reguł decyzyjnych: o ok. 82%, a gdy zastosujemy skupienia częściowych reguł decyzyjnych skrócimy ten czas o 97%. Wszystko dzięki temu, że przeglądamy jedynie 12 węzłów w drzewie (parametr N_{Data}). Niewątpliwie zatem zastosowania którejkolwiek z przedstawionych koncepcji pozwoli zwiększyć efektywność systemu wspomagania decyzji, w którym baza wiedzy zostanie odpowiednio zmodyfikowana.

Tabela 4

Procent bazy wiedzy przeglądany w procesie wnioskowania

zbiór	Rl_{max}	N_R	N_{Data}	%KB-τ_α	%KB-($\tau_{AS+\alpha}$)	%KB-τ
Zoo	4	101	12	25,00%	2,97%	11,88%
Zoo	4	101	10	25,00%	2,48%	9,90%
Zoo	4	101	12	25,00%	2,97%	11,88%
Zoo	2	101	8	18,75%	1,49%	7,92%
Zoo	2	101	10	18,75%	1,86%	9,90%
Zoo	1	101	14	6,25%	0,87%	13,86%
Lympho	4	148	12	22,22%	1,80%	8,11%
lympho	2	148	18	11,11%	1,35%	12,16%
lympho	2	148	14	11,11%	1,05%	9,46%
lympho	1	148	26	5,56%	0,98%	17,57%
spect all	10	267	12	43,48%	1,95%	4,49%
spect all	10	267	12	43,48%	1,30%	3,00%
spect all	10	267	8	43,48%	1,30%	3,00%
spect all	7	267	26	30,43%	2,96%	9,74%
spect all	5	267	16	21,74%	1,30%	5,99%
spect all	2	267	8	8,70%	0,26%	3,00%

6. Wpływ miary podobieństwa na efektywność procesów wnioskowania w systemie wspomagania decyzji z zastosowaniem skupień częściowych reguł decyzyjnych

Niewątpliwie miara podobieństwa bądź odległości stosowana, jako kryterium przydziału obiektów (reguł) do grup w znaczący sposób wpływa na efektywność systemów wspomagania decyzji. Poparciem tego wniosku niech będą wyniki eksperymentów i treść pracy [10]. W badaniach użyto miar: odległości euklidesowej oraz miar podobieństwa: *Gowera*, miary cosinusowej oraz nakładania [2 i 9]. Dwie ostatnie miary (jak wykazano) w niektórych przypadkach błędnie wyznaczają wartość podobieństwa, traktując pewne wektory jako identyczne (o czym ma świadczyć wartość 1 takiej miary), podczas gdy wiadomo, że wektory te może są w znacznym stopniu podobne, ale nie identyczne. Miary *Gowera* oraz odległości euklidesowej okazały się najbardziej efektywne w wyznaczaniu tego podobieństwa. Zaletą miary *Gowera* jest fakt, że miara jest na tyle elastyczna, że potrafi się dostosować zarówno do danych ilościowych, jak i jakościowych, a ponadto radzi sobie również z danymi niekompletnymi (gdybyśmy np. nie znali wartości jakiegoś atrybutu w regule, którą porównujemy z inną regułą, której atrybut i wartość tego atrybutu są znane). Dla rozwinięcia analizy postanowiono przeprowadzić eksperymenty z użyciem dodatkowych, specyficznych miar podobieństwa i ostatecznie wybrano: modyfikowaną miarę nakładania, *indeks Tversky'ego* oraz *indeks Jaccarda.*

6.1. Proponowane miary podobieństwa

Modyfikowana miara nakładania tym ma się różnić od miary nakładania (klasycznej, testowanej w poprzednich eksperymentach), że w mianowniku zamiast obliczać pewną wartość minimalną, wybiera wartość maksymalną. To sprawi, że nigdy już dla nieidentycznych wektorów nie uzyskamy wartości 1. Jeśli r_1 i r_2 to analizowane reguły z bazy wiedzy i każda reguła będzie zapisana jako wektor n-elementowych, gdzie wartość r_{1i} oznacza wartość i-tego atrybutu w regule r_1, wówczas modyfikowaną miarę nakładania można będzie zapisać następująco:

$$S(r_1, r_2) = \frac{\sum_{i=1}^{n} \min(r_{1i} \cap r_{2i})}{\max(\sum_{i=1}^{n} r_{1i}, \sum_{i=1}^{n} r_{2i})}$$

Dość ciekawym podejściem wydaje się być tzw. *indeks Tversky'ego*, w którym podobieństwo dwóch reguł r_1 i r_2 wyznaczamy następująco:

$$S(r_1, r_2) = \frac{|r_1 \cap r_2|}{|r_1 \cap r_2| + \alpha |r_1 - r_2| + \beta |r_2 - r_1|}$$

We wzorze pojawiają się dwa parametry α i β, które w ogólnej postaci powinny być o tej samej wartości (α=1/2 i β=1/2), a ich suma powinna się sumować do jedności ($\alpha + \beta$=1). Jeśli któraś z reguł porównywanych r_1 bądź r_2 jest ważniejsza, wówczas wartości α i β są odpowiednio mniejsze bądź większe.

Dużą popularnością cieszy się *miara Jaccarda*, której ogólną postać możemy zapisać w następujący sposób:

$$S(r_1, r_2) = \frac{|r_1 \cap r_2|}{|r_1 \cup r_2|}.$$

6.2. Skuteczność miar podobieństwa

W eksperymentach postanowiono obliczyć jaka była skuteczność wyboru poszczególnych miar podobieństwa bądź odległości w obu kwestiach: grupowania reguł (budujących bazę wiedzy) oraz wyszukiwania reguł do uaktywnienia (w procesie wnioskowania). Wyniki prezentuje rys. 4.

Wyszukiwanie reguł do uaktywnienia (wnioskowanie)

Grupowanie reguł (budowa BW)

	gower	cosinus	overlap	Modyfikacja (overlap)	tversky	Jaccard
Euklides	100%	81%	75%	100%	58%	50%
Gower	100%	88%	75%	100%	44%	50%
Cosinus	100%	81%	75%	100%	44%	50%
Overlap	100%	81%	75%	100%	56%	50%
Modyfikacja (overlap)	100%	94%	85%	100%	81%	88%
Tversky	100%	81%	73%	100%	56%	50%
Jaccard	100%	81%	73%	100%	56%	50%

Rys. 4. Skuteczność miar podobieństwa
Fig. 4. The efficiency of similarity measures

Interpretacja rysunku powinna być następująca. Jeśli weźmiemy pod uwagę miarę podobieństwa *Gowera* do grupowania reguł w bazie wiedzy, to później w procesie wyszukiwania reguł do uaktywnienia tylko w przypadku miar *Tversky'ego* i *Jaccarda* skuteczność (a więc zdolność wyszukania reguły relewantnej) nie przekracza 50%. Z kolei skuteczność tej miary do samego wyszukiwania reguł relewantnych w procesie wnioskowania, niezależnie od metryki wcześniej użytej do grupowania jest równa 100%. Podobnie jest w przypadku modyfikowanej miary nakładania (ang. *overlap*).

7. Podsumowanie

Dopiero połączenie kilku metod eksploracji wiedzy pozwala optymalizować systemy wspomagania decyzji. Hierarchiczna baza wiedzy ze skupieniami reguł podobnych do siebie pozwala realizować zadanie modularyzacji baz wiedzy, a także skraca czas wnioskowania i rozmiar generowanej nowej wiedzy. Przeprowadzone eksperymenty potwierdzają, iż zaproponowane metody wnoszą istotne skrócenie czasu wnioskowania, poprzez redukcję liczby przeszukiwanych reguł (konieczność przeszukania 2% BW), nie osłabiając przy tym (w większości przypadków) parametrów dokładności wyszukiwania. Ostateczna efektywność utworzonych struktur skupień reguł zależy nie tylko od wybranego algorytmu grupowania, ale także od wybranej metryki podobieństwa bądź odległości. Najbardziej efektywna okazała się być miara Gowera oraz modyfikowana miara nakładania (skuteczność 100%). Nasuwa się także pytanie, czy może inny algorytm grupowania wpłynąłby na efektywność utworzonej struktury bazy wiedzy? Niech to pytanie posłuży jako kierunek dalszych badań w tym zakresie.

BIBLIOGRAFIA

1. Kaufman L., Rousseeuw P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley Sons, New York 1990.
2. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT, Warszawa 2005.
3. Nowak A., Wakulicz-Deja A., Bachliński S.: Optimization of Speech Recognition by Clustering of Phones. Fundamenta Informaticae 72, IOS Press, 2006, s. 283÷293.
4. Nowak A., Simiński R., Wakulicz-Deja A.: Towards modular representation of knowledge base. Springer-Verlag Berlin Heidelberg – Advances in Soft Computing, 2006, s. 421÷428.
5. Salton G.: Automatic Information Organization and Retreival. McGraw-Hill, New York 1975.
6. Jardine N., van Rijsbergen C. J.: The use of hierarchic clustering in information retrieval. Inforamtion Storage and Retrieval 7, s. 217÷240.
7. Nowak A., Zielosko B.: Inference Processes on Clustered Partial Decision Rules. Recent Advances In Intelligent Information Systems, Springer-Verlag, Advances In Soft Computing, Academic Publishing House EXIT, 2009, s. 579÷588.
8. Nowak A., Zielosko B.: Clustering of partial decision rules. Advanced In Intelligent and Soft Computing, Man-Machine Interactions, Springer-Verlag, 2009, s. 183÷190.

9. Nowak A.: Złożone bazy wiedzy: struktura i procesy wnioskowania. Rozprawa doktorska, Uniwersytet Śląski, Katowice 2009.
10. Nowak-Brzezińska A., Wakulicz-Deja A.: Wybór miary podobieństwa a efektywność grupowania reguł w złożonych bazach wiedzy. Studia Informatica, Wyd. Pol. Śląskiej, Vol. (31), No. 2A (89), 2010, s. 189÷202.
11. Moshkov M., Piliszczuk M., Zielosko B.: Partial covers, reducts and decision rules in rough sets. Theory and aaplications. Springer Verlag, Studies in Computational Intelligence, Vol. 145, 2008.
12. UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/.

Recenzent: Dr inż. Jacek Frączek

Wpłynęło do Redakcji 26 stycznia 2011 r.

Abstract

The paper presents the results of the experiments in which an efficiency of decision support systems is analyzed. During the research it was noted that the representation of knowledge base mostly influences on such efficiency. Following approaches as the methods for representation of rules in knowledge base are considered due to experiments: rules knowledge base with clusters of decision rules, rules knowledge base with partial decision rules and rules knowledge base with clusters of partial decision rules. If a given decision support system doesn't use modified structure of knowledge base, and consists of large number of rules, the inference process is not efficient. The time of inference process is too long and the number of rules that are fired are too big. It makes the process of interpretation such new knowledge generated as the inference process result too difficult for a given user. Much more effective is using some modifications of knowledge base, for example using partial decision rules. It is a method proposed in [11] where instead of original decision rules, partial decision rules are generated. Such rules are just much shorter than rules in original knowledge base, so the time of analyses all of them is decreased. Another method of knowledge base structure is cluster analysis for rules in original knowledge base. Using clustering method for similar rules it is possible to create groups of rules. Each of such groups consists of representative vector which tells us something important about a given group of rules. Then having a structure of rules clusters with their representative vectors we simplify find a group that are most similar to the query (set of facts, the hypothesis). It lets to decrease the time of finding relevant rules in inference process. The most efficient method is clustering partial decision rules,

which clusters short rules, so the time of finding rules for firing and the time necessary to analyze chosen rule is quite small. The experiments compare the time of inference processes in different knowledge base structures for a given real data: *zoo, spect_all* and *Lymphography*. The efficiency of choosing the best similarity measure is also analyzed. It seems that Gower similarity and modified overlapping measure are the best from the set of following measures: Euclidean, Gower, overlapping, modified overlapping, and indexes of Tverski and Jaccard.

Adres

Agnieszka NOWAK-BRZEZIŃSKA: Uniwersytet Śląski, Instytut Informatyki, Wydział Informatyki i Nauki o Materiałach, ul. Będzińska 39, 41-200 Sosnowiec, Polska, agnieszka.nowak@us.edu.pl.