STUDIA INFORMATICA 2012
Volume 33 Number 2A (105)

Agnieszka NOWAK-BRZEZIŃSKA, Tomasz JACH
Uniwersytet Śląski, Instytut Informatyki

WYBRANE ASPEKTY WNIOSKOWANIA W SYSTEMACH Z WIEDZĄ NIEPEŁNĄ

Streszczenie. Autorzy proponują użycie metod analizy skupień (grupowania) do szybkiego wyszukiwania, aktywowania reguł i wnioskowania w złożonych bazach wiedzy z wiedzą niepełną. Artykuł porównuje użycie dwóch algorytmów – AHC oraz mAHC, przedstawiona jest również metoda do wyznaczania optymalnej liczby skupień oraz eksperymenty obliczeniowe potwierdzające zdolność zaproponowanego podejścia do wnioskowania z wiedzą niepełną.

Słowa kluczowe: bazy wiedzy, grupowanie, analiza skupień, AHC, wnioskowanie, wiedza niepełna, systemy wspomagania decyzji

THE CHOSEN ASPECTS IN INFERENCE PROCESSES IN DECISION SUPPORT SYSTEMS WITH INCOMPLETE KNOWLEDGE

Summary. The authors propose to use the methods of cluster analysis (clustering) in complex decision support systems with incomplete knowledge. The paper compares using of mAHC and AHC algorithms. The problem of finding the optimal number of clusters is addressed, the experiments confirming the ability of proposed approach to inference within decision support systems with incomplete knowledge are provided.

Keywords: knowledge bases, cluster analysis, clustering, decision support systems, incomplete knowledge, inference, AHC

1. Wprowadzenie

Od wielu lat komputerowe systemy wspomagania decyzji są nieodłącznym elementem skomplikowanych systemów informatycznych. Pryncypium ich działania [1] stanowi zdolność do wyciągania logicznych i poprawnych wniosków przy dysponowaniu zbiorem faktów oraz reguł. Klasyczne wnioskowanie polega na uaktywnianiu reguł, których wszystkie prze-

słanki są spełnione (inaczej: przesłanki są faktami znanymi w systemie). Przez r_i oznaczymy i-tą regułę w systemie odpowiadającą przyjętej w klasycznych SWD postaci klauzuli Horna, gdzie każdy literał z części przesłankowej i decyzyjnej tworzony jest na podstawie zbióru atrybutów A oraz zbioru wartości każdego atrybutu $V_a, a \in A$. Parę (a, v_a) budującą przesłanki i konkluzje reguł będziemy dalej nazywać deskryptorem $(d_i = (a_j, v_{a_j}))$, dzięki czemu regułę r_i możemy przedstawić następująco: $r_i = d_1 \wedge d_2 \wedge ... \wedge d_m \Rightarrow DEC_i$.

Efektywność systemów wspomagania decyzji zależy przede wszystkim od czasu wnioskowania, który z kolei jest tym dłuższy, im więcej reguł zawiera analizowana baza wiedzy. Fakt, że reguły mogą mieć zmienną liczbę przesłanek również wpływa na czas ich analizy. Złożoność budowy reguł też ma tu istotne znaczenie. Dopuszcza się bowiem tzw. reguły złożone, w których wybrany literał (para (atrybut, wartość)), występujący w części konkluzyjnej jednej z reguł, stanowi część przesłankową w innej regule. W procesie wnioskowania wprzód analizowane są wszystkie reguły i te, których przesłanki mają pokrycie w zbiorze faktów, zostają uaktywnione, a konkluzje tych reguł zostają dopisane jako nowe fakty do bazy wiedzy. Generowanie zbyt dużej liczby nowych faktów – oprócz oczywistej zalety w formie poszerzonej wiedzy w badanym zakresie – ma także wadę w postaci trudności interpretacji dużej ilości nowej wiedzy. Przeanalizujmy prostą bazę reguł:

$r1: (pogoda, słłoneczn\ ) \wedge (temperatura, wysoka) \wedge (wiatr, słłab\ ) \Rightarrow (spacer, tak),$

$r2: (pogoda, słłoneczn\ ) \wedge (temperatura, niska) \wedge (wiatr, słłab\ ) \Rightarrow (spacer, nie),$

$r3: (pogoda, pochmurna) \wedge (temperatura, wysoka) \wedge (wiatr, silny) \Rightarrow (spacer, tak),$

$r4: (pogoda, deszcz) \Rightarrow (weź\ parasol, tak),$

$Fakty: (pogoda, słłoneczn\ ), (temperatura, niska), (wiatr, silny).$

Zgodnie z regułami logiki klasycznej, aby którakolwiek reguła mogła być uaktywniona, wszystkie jej przesłanki muszą być spełnione [2]. Jednak w tym przypadku żadna z reguł nie będzie uaktywniona. Należy również zauważyć, że nie mamy do czynienia z prostym problemem klasyfikacji, ponieważ wartości decyzji mogą dotyczyć całkowicie różnych dziedzin (np. wszystkie systemy przemysłowego sterowania procesem produkcji składają się nie tylko z decyzji „wstrzymać produkcję" albo „wznowić produkcję", ale także „wolna linia produkcyjna nr 1", co może stać się przesłanką do innej reguły zapisanej w systemie).

W przypadku gdy żadna z reguł nie może zostać uaktywniona, pojawia się niepożądana sytuacja impasu: system nie jest w stanie ani podjąć, ani wspomóc użytkownika żadnymi nowymi wnioskami, mając do dyspozycji zebrane obserwacje. Dzięki zastosowanym algorytmom analizy skupień [3] reguły w bazie wiedzy są grupowane przy wzięciu pod uwagę kryterium podobieństwa, co w efekcie pozwala znacznie przyspieszyć wyszukiwanie reguł nawet w dużych bazach wiedzy. Proponowany przez autorów system jest w stanie: szybko odnaleźć reguły, które mają największe pokrycie w bazie faktów (1), oraz oznaczyć do uak-

tywnienia znalezione reguły (2). W ten sposób użytkownik otrzymuje dodatkową wiedzę z systemu, przy jednoczesnej informacji, iż wiedza ta nie jest całkowicie pewna w sensie logicznym. Podcel (1) jest istotny ze względu na fakt coraz większej liczby reguł oraz znacznego zwiększania ich stopnia skomplikowania w zastosowaniach praktycznych. Klasyczne algorytmy wnioskowania zarówno wprzód (sterowane faktami), jak i wstecz (sterowane celem) stają się nieefektywne pod względem złożoności czasowej, gdy muszą przeszukać całą bazę wiedzy, reguła po regule. Proponowane podejście, wykorzystujące wiedzę o reprezentantach grup reguł podobnych do siebie, w bardzo krótkim czasie znajduje grupę najbardziej podobną do szukanych informacji (podanych faktów) i tylko tę grupę analizuje w procesie wnioskowania.

Zaproponowane podejście jest podobne do teorii współczynników CF [4]; rozszerza ją o możliwość szybkiego wyszukania reguł oraz brak konieczności określania wartości współczynników CF dla wszystkich przesłanek. Alternatywne podejście można spotkać w systemach bazujących na zbiorach rozmytych [5] oraz przybliżonych [6].

Autorzy proponują użycie pojęcia niepełności wiedzy dla reguł, których nie wszystkie przesłanki są spełnione. W proponowanym systemie takie reguły będą mogły być uaktywnione, przy czym zostanie określony stopień ich pokrycia, pozwalający ocenić na ile możemy ufać danej regule.

2. Proponowane rozwiązanie

Autorzy proponują użycie mechanizmów analizy skupień w celu grupowania reguł w skupienia oraz wyszukiwanie reguł metodą pnia najbardziej obiecującego [7]. Dodatkowo proponuje się uaktywnianie także reguł, których nie wszystkie przesłanki mają swoje pokrycie w bazie faktów. Więcej uwagi temu tematowi autorzy poświęcili w pracy [8]. Dla lepszego zrozumienia kolejnych rozważań w dalszej części przedstawiony będzie jedynie ogólny schemat pracy systemu.

2.1. Grupowanie reguł

Grupowanie reguł realizowane jest dzięki algorytmom hierarchicznym analizy skupień AHC oraz Agnes [3]. Szczegółowy sposób działania algorytmów prezentowany jest w poprzedniej pracy autorów [8], dalej przedstawiony zostanie jedynie zarys i ogólne założenia.

W pierwszym kroku algorytmu analizowane jest podobieństwo reguł. W tym celu generowana jest kwadratowa macierz podobieństwa. Na przecięciu i-tego wiersza oraz j-tej kolumny znajduje się wartość określająca wzajemne podobieństwo dwóch reguł mierzonych przy użyciu miar podobieństwa szczegółowo omówionych w pracy [8]. Reguły najbardziej

podobne tworzą skupienie, dzięki czemu w kolejnych krokach analizowane jest podobieństwo nie tylko reguł, ale również skupień reguł do innych reguł i skupień. Proces powtarzany jest do otrzymania pełnej hierarchii grup, którą można przedstawić graficznie w formie dendrogramu [9].

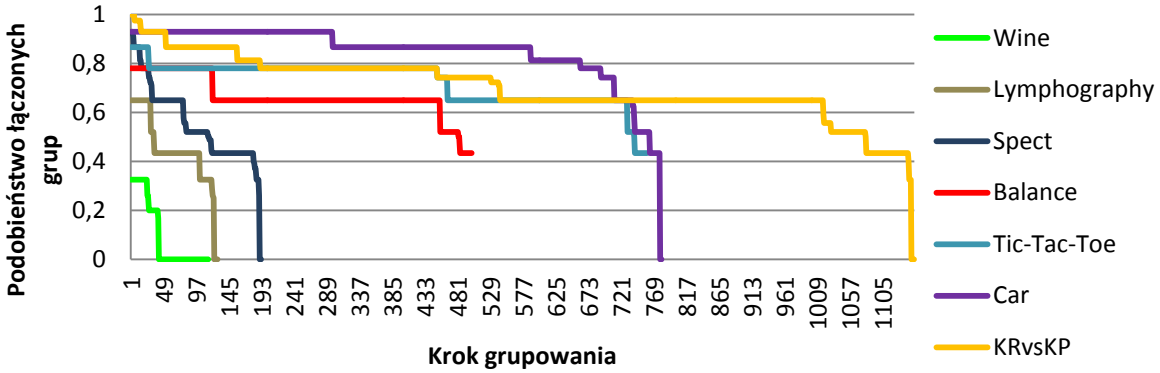
2.2. Wyszukiwanie reguł do uaktywnienia

Proces wnioskowania rozpoczyna się od znalezienia w bazie wiedzy reguł do uaktywnienia (stąd proces nazywany jest procesem wyszukiwania). We wcześniejszych pracach w procesie wyszukiwania ustalano konkretną liczbę grup w systemie, wyznaczano reprezentantów tychże grup oraz porównywano zbiór faktów z reprezentantami grup. To rodziło (sygnalizowany wcześniej) problem z wyznaczeniem optymalnej liczby skupień. Proponowane w tym artykule podejście pozwala ów problem rozwiązać.

2.2.1. Wyszukiwanie strukturalne

Proponowanym rozwiązaniem jest obserwacja wzajemnego podobieństwa grup łączonych w konkretnym kroku grupowania. Algorytm przedstawia się następująco:

1. Ustalenie parametrów grupowania: wybór miary podobieństwa, metody łączenia skupień, liczby skupień.
2. Rozpoczęcie algorytmu grupowania.
3. Zapamiętanie maksymalnej wartości podobieństwa dwóch reguł z pierwszego kroku algorytmu grupowania (simMax).
4. Wyznaczenie wartości progowej współczynnika podobieństwa dwóch skupień, będącej iloczynem simMax oraz wartości podanej przez użytkownika, określającej, kiedy zakończyć grupowanie.
5. Wykonywanie algorytmu grupowania dopóki łączone skupienia mają wartość podobieństwa większą od wartości progowej.



Rys. 1. Podobieństwo grup łączonych w poszczególnych krokach algorytmu grupowania
Fig. 1. Similarity between groups merged in each step of the clustering algorithm

Rys. 1 przedstawia wartość podobieństwa grupowanych reguł w kolejnych krokach grupowania. Widać wyraźnie, że w pewnym momencie jakość grupowania drastycznie spada. Autorzy sądzą, że jest to punkt, w którym łączone ze sobą skupienia reguł są już dość mało do siebie podobne. Jest to sygnał do zaprzestania grupowania. Eksperymenty przeprowadzono dla baz pobranych z Machine Learning Repository [10]. Autorzy wybrali bazy o różnym stopniu skomplikowania (różna liczba reguł czy atrybutów opisujących reguły), aby ustalić optymalną wartość parametru współczynnika progowego.

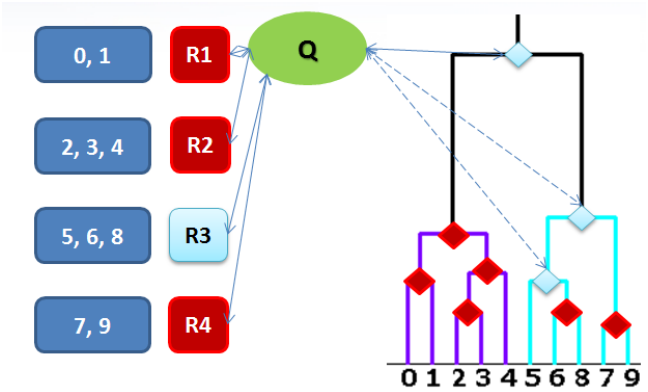
W wyniku eksperymentów najlepsze rezultaty osiąga się w przypadku, gdy współczynnik progowy ma wartość 0,85 * simMax (patrz rys. 1).

2.2.2. Wyszukiwanie hierarchiczne

Algorytm AHC generuje pełne drzewo reguł [11], co pozwala szybko wyszukiwać reguły, porównując w każdym kroku zbiór faktów do reprezentantów poddrzew lewego i prawego aż do zejścia do poziomu liści. Formalnie, jeśli przez D będziemy rozumieć zbiór deskryptorów (par atrybut-wartość), f będzie funkcją podobieństwa, która dwóm regułom (skupieniom reguł) przyporządkowuje wartość podobieństwa, a k_i, l_i będą węzłami łączonymi, wtedy każda grupa w_i będzie definiowana jako $w_i = (D_i, f, k_i, l_i)$, gdzie:

$$D_i = \{d_i, \ldots, d_m\}, f : X \times X \to R \,|\, [0 \ldots 1].$$

Analogicznie do idei, na której opiera się system SMART Saltona [7], metoda węzła najbardziej obiecującego rozpoczyna wyszukiwanie od korzenia drzewa. Następnie w każdym kroku zbiór faktów Q porównywany jest z reprezentantami poddrzew prawego i lewego aktualnie rozpatrywanego węzła, co przedstawia rys. 2. Do dalszej analizy wybierana jest ścieżka o większej wartości podobieństwa faktów do grupy reguł. Proces kończy się w momencie dotarcia do liścia oznaczającego konkretną regułę w bazie wiedzy.



Rys. 2. Wyszukiwanie strukturalne (z lewej) i wyszukiwanie hierarchiczne (z prawej)
Fig. 2. Structural search (left) and hierarchical search (right)

Problemem w implementacji metody węzła najbardziej obiecującego jest wyznaczenie wartości podobieństwa zbioru faktów do poszczególnych węzłów. W tym celu autorzy pro-

ponują trzy różne podejścia: metodę pokrycia deskryptorowego, metodę pokrycia atrybutowego oraz podejście hybrydowe.

Najbardziej intuicyjną miarą jest wyznaczenie liczby deskryptorów występujących zarówno w zbiorze faktów, jak i w poszczególnych węzłach zgodnie ze wzorem:

$$f_d(k,l) = card(d_k \cap d_l),$$

gdzie d_l oraz d_k to zbiory deskryptorów węzłów odpowiednio l i k.

Takie podejście, nazwane metodą pokrycia deskryptorowego, faworyzuje jednak węzły zawierające dużą liczbę powtarzających się, częstych deskryptorów w systemie. Co więcej, w kontekście wiedzy niepełnej już informacja o wspólnych atrybutach występujących w obu grupach powinna być brana pod uwagę w wyliczaniu podobieństw (np. ze względu na niedoskonałości pomiaru, wartości puste itp.). Zwłaszcza w bazach medycznych, często niepoddanych poprawnej dyskretyzacji, informacje o wykonaniu danego badania (bez znajomości wyniku) będą mogły być wykorzystane do dalszego rozróżniania grup. Niestety bez znajomości relacji pomiędzy wartościami poszczególnych atrybutów nie jest możliwe wyprowadzenie funkcji podobieństwa, uwzględniającej, które wartości atrybutów są sobie bliższe niż inne, innymi słowy – dla większości atrybutów będących cechami nominalnymi (jakościowymi). Dlatego też autorzy proponują ogólne podejście, właściwe niezależnie od rodzaju danych.

Drugim ze sposobów określania miary podobieństwa jest metoda pokrycia atrybutowego, która przy wyznaczaniu podobieństwa reguł (bądź ich skupień) bierze pod uwagę tylko informacje o wspólnych atrybutach. Przy zachowaniu konwencji oznaczeń omówionych w poprzednim rozdziale i gdy zbiory a_k oraz a_l oznaczać będą zbiory atrybutów reguł występujących w poszczególnych grupach, miara pokrycia atrybutowego będzie wyznaczana w sposób następujący:

$$f_a(k,l) = card(a_k \cap a_l).$$

Ze względu na duże podobieństwo reguł do siebie oraz stosunkowo liczne zbiory wartości poszczególnych atrybutów autorzy postanowili wykorzystywać w obliczaniu podobieństwa dwóch węzłów tylko informacje o wspólnych atrybutach. Dzięki temu można będzie wyróżnić, być może spójne, grupy reguł.

Trzecim zaproponowanym podejściem będzie połączenie dwóch poprzednich sposobów w metodzie hybrydowej:

$$f_h(k,l) = card(d_k \cap d_l) \cdot \mathrm{C}_1 + card(a_k \cap a_l) \cdot \mathrm{C}_2,$$

gdzie C_1 oraz C_2 to dodatnie czynniki stopniujące sumujące się do wartości 1.

Autorzy sądzą, że podejście to wykorzysta zalety zarówno dokładności metody z pokryciem deskryptorów, jak i dodatkowych informacji rozróżniających reguły pomiędzy sobą. Współczynniki stopniujące służą zwiększaniu lub zmniejszaniu ważności części deskrypto-

rowej i atrybutowej. W rozważaniach sprawdzono dwa przypadki: w jednym znacznie większą ważność otrzymuje część deskryptorowa, w drugim – część atrybutowa.

Przykładowo dla dwóch węzłów: $k : d_k = \{(A,1),(A,1),(A,2),(B,1),(B,1),(C,1)\}$ oraz $l : d_l = \{(A,2),(A,2),(B,1),(B,1),(B,1),(C,1)\}$ i zbioru faktów $Q = \{(A,2),(C,1)\}$ odpowiednie wartości podobieństw przedstawiają się następująco:

- $f_d(k,Q) = 2;\ f_d(l,Q) = 3,$
- $f_a(k,Q) = 4;\ f_a(l,Q) = 3,$
- dla C_1=0,75 oraz C_2=0,25 $f_h(k,Q) = 2{,}5;\ f_h(l,Q) = 3,$
- dla C_1=0,25 oraz C_2=0,75 $f_h(k,Q) = 3{,}5;\ f_h(l,Q) = 3.$

Widać wyraźnie, że podejście hybrydowe pozwala na uwzględnienie również współwystępowania atrybutów w zbiorze faktów oraz analizowanej regule i być może przyczyni się do wyboru optymalnej ścieżki w warunkach dużego podobieństwa do siebie reprezentantów węzłów, reprezentujących obie ścieżki.

3. Eksperymenty obliczeniowe

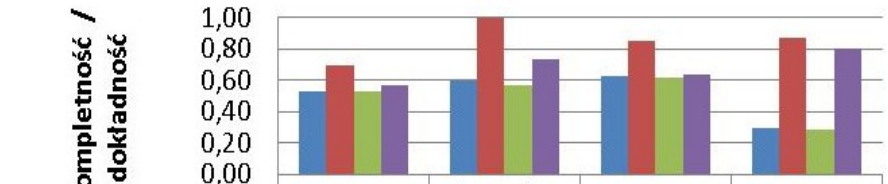
3.1. Wyszukiwania hierarchiczne i strukturalne

W celu porównania przedstawionych podejść autorzy zaimplementowali dwa hierarchiczne algorytmy grupowania: omówiony wcześniej AHC, korzystający z pełnego drzewa hierarchicznego reguł i wyszukiwania, oraz mAHC, wykorzystujący wcześniej omówioną technikę wyznaczania optymalnej liczby skupień i wyznaczający reprezentantów każdej z grup. Wyniki przedstawia rys. 3.

Oba algorytmy grupowania testowano dla czterech wybranych baz danych z repozytorium MLR. Dla każdej bazy wiedzy przygotowano osobno po 10 zestawów faktów losowo wybranych spośród przesłanek i konkluzji reguł faktycznie zapisanych w tych bazach. Obliczono wartości kompletności rozumianej jako stosunek liczby wspólnych deskryptorów, występujących zarówno w zbiorze faktów, jak i w reprezentancie (mAHC) lub w otrzymanej grupie (AHC), do liczby wszystkich deskryptorów opisujących zbiór faktów i reprezentanta (mAHC) lub grupę (AHC).

Niestety wyszukiwanie z użyciem hierarchii daje stosunkowo gorsze rezultaty od podejścia strukturalnego. Autorzy sugerują dalsze prace nad udoskonaleniem wyszukiwania, w szczególności omówione w dalszej części artykułu.

Autorzy przeanalizowali również czas wyszukiwania reguł do uaktywnienia dla obu algorytmów. Wyniki prezentuje rys. 4.



	Wine	Lympho graphy	Balance	Abalone
Kompletność AHC	0,53	0,60	0,63	0,29
Kompletność mAHC	0,70	1,00	0,85	0,87
Dokładność AHC	0,53	0,57	0,62	0,29
Dokładność mAHC	0,57	0,74	0,64	0,80

Rys. 3. Jakość wyszukiwań hierarchicznego i strukturalnego
Fig. 3. The quality of hierarchical and structural search

	Wine	Lymphogra phy	Balance	Abalone
Czas (norm) AHC	0,26	0,11	0,05	0,47
Czas (norm) mAHC	1	1	1	1

Rys. 4. Czas wyszukiwania z użyciem algorytmów AHC i mAHC
Fig. 4. Clustering time using AHC and mAHC

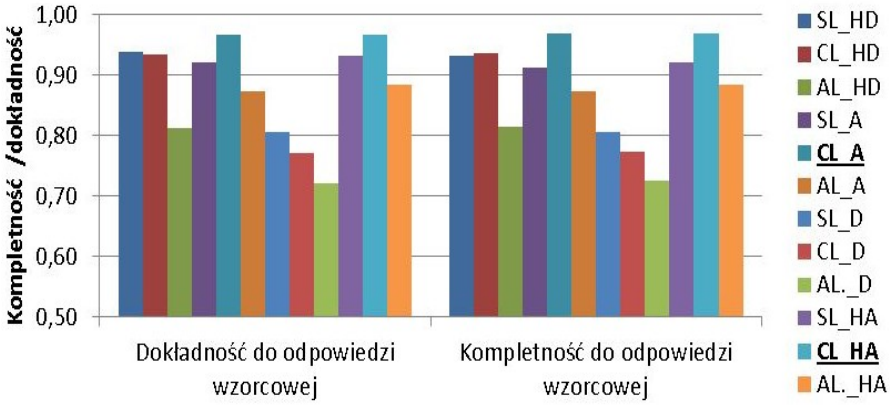
Oczywiste jest, że podejście hierarchiczne będzie znacznie szybsze i hipoteza ta znalazła swoje potwierdzenie w wynikach eksperymentalnych. Co więcej, im baza reguł jest większa, tym widoczny jest większy zysk czasowy dla podejścia hierarchicznego.

3.2. Metoda ścieżki najbardziej obiecującej

W celu sprawdzenia, która z przedstawionych metod da optymalne rezultaty, przeprowadzono eksperymenty obliczeniowe. Na początku przyjmowano, że deskryptory opisujące aktualnie analizowaną regułę stanowią jednocześnie zbiór faktów. Do pełnego systemu, wyznaczonego za pomocą różnej kombinacji metod ścieżki najbardziej obiecującej oraz metody łączenia skupień, zadawano ów zbiór faktów. Odpowiedź systemu traktowano jako odpowiedź wzorcową.

Następnie z bazy wiedzy usuwano tę konkretną regułę i powtarzano proces wyszukiwania reguł dla tego samego zbioru faktów. W kolejnym krokusprawdzano wartości kompletności

i dokładności wyszukiwania dla takiego przypadku. Na wszystkich wykresach przyjęto następujące oznaczenia: SL – Single Linkage, CL – Complete Linkage, AL – Average Linkage, HD – metoda hybrydowa ścieżki najbardziej obiecującej ze zwiększoną wartością współczynnika dla wspólnych deskryptorów, HA – metoda hybrydowa ścieżki najbardziej obiecującej ze zwiększoną wartością współczynnika dla wspólnych atrybutów, A – metoda pokrycia atrybutowego, D – metoda pokrycia deskryptorowego.

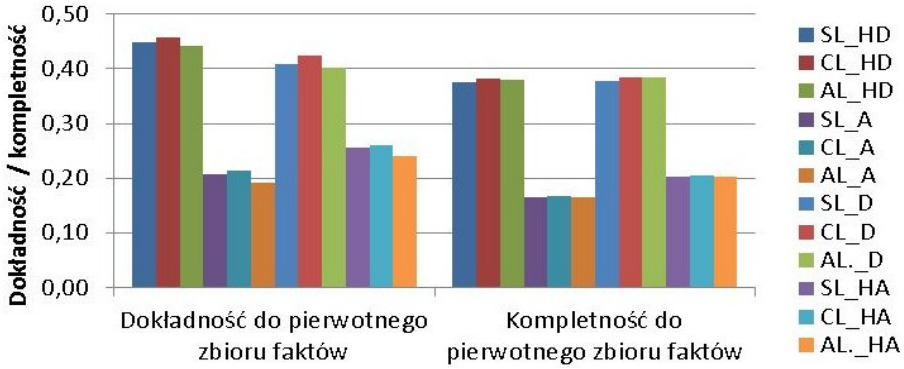


Rys. 5. Eksperymenty dla metody ścieżki najbardziej obiecującej
Fig. 5. Experiments involving the most promissing path

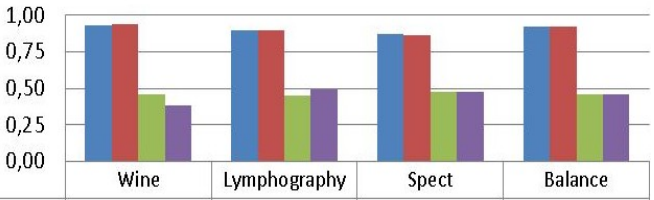
Najlepsze rezultaty uzyskano, gdy algorytm korzystał z metody CL do wiązania skupień. W przypadku oceny dokładności i kompletności w stosunku do najlepszej grupy wyniki wszystkich przedstawionych podejść prezentują zbliżone rezultaty (najlepsza jest metoda hybrydowa z większą wagą dla wartości wspólnych atrybutów oraz metody pokrycia atrybutowego). Wydaje się, że podejście rozróżniające grupy daje lepsze rezultaty w stosunku do tradycyjnego podejścia pokrycia deskryptorowego.

W drugiej części eksperymentu obliczono kompletność i dokładność odpowiedzi dla ograniczonego systemu w stosunku do zadawanego zbioru faktów. W ten sposób autorzy chcieli zbadać zdolność stworzonego systemu do kompensowania niepełnej wiedzy. Jednakże należy zauważyć, że wartości kompletności i dokładności bliskie maksymalnym nie mogą wystąpić ze względu na fakt usunięcia reguły, która jest optymalną odpowiedzią na zbiór faktów z systemu (jego ograniczenie).

Rysunek 6 jednoznacznie pokazuje, że autorska metoda hybrydowa ze wzmocnieniem części deskryptorowej sprawdza się lepiej, niezależnie nawet od sposobu łączenia skupień. Wartości parametrów efektywności uzyskiwane dla tej metody są blisko dwukrotnie większe od pozostałych rozwiązań. Do dalszych eksperymentów przyjęto metody całkowitego wiązania (CL) oraz ścieżki najbardziej obiecującej ze wzmocnieniem wag dla wspólnych deskryptorów. Ustaliwszy je, autorzy przeprowadzili testy dla większej liczby baz wiedzy. Wyniki prezentuje rys. 7. Bliskie maksymalnym wartości kompletności i dokładności w stosunku do grupy optymalnej potwierdzają słuszność zaproponowanego podejścia w problemie wnioskowania w systemach z wiedzą niepełną.



Rys. 6. Wyniki eksperymentów obliczeniowych
Fig. 6. The results of computational experiments



	Wine	Lymphography	Spect	Balance
Dokładność do odpowiedzi wzorcowej	0,93	0,89	0,87	0,92
Kompletność do odpowiedzi wzorcowej	0,94	0,89	0,87	0,92
Dokładność do pierwotnego zbioru faktów	0,46	0,45	0,47	0,46
Kompletność do pierwotnego zbioru faktów	0,38	0,49	0,47	0,46

Rys. 7. Wyniki eksperymentów metody hybrydowej dla wybranych baz wiedzy
Fig. 7. The results of hybrid method for chosen knowledge bases

4. Wnioski oraz kierunki dalszych badań

Wcześniejsze rezultaty [8] udało się poprawić dzięki zaproponowanemu w niniejszym artykule algorytmowi, pozwalającemu określić właściwą liczbę skupień reguł. Poprzednie eksperymenty musiały opierać się na metodzie wielokrotnych powtórzeń przy różnej liczbie tworzonych grup i wyborze rozwiązania optymalnego. Zaproponowano również model wyszukiwania reguł możliwych do aktywowania przy założeniach o niepełności wiedzy, poczynionych przez autorów i omówionych we wstępie artykułu. Modyfikacja metody ścieżki najbardziej obiecującej pozwoliła na znaczne skrócenie czasu wnioskowania oraz umożliwiła odnalezienie reguł o minimalnej liczbie niespełnionych przesłanek w krótkim czasie. Aktywacja tych reguł przyczyni się do zwiększenia liczby generowanych faktów, a dzięki temu – do wyprowadzenia większej wiedzy z systemu. Wnioskiem z przedstawionych badań jest ustalenie wartości współczynnika progowego dla optymalnej liczby grup na około 85% wartości największego podobieństwa dwóch reguł między sobą. Dzięki temu liczba grup jest

stosunkowo duża. Mimo zwiększonego nakładu na wyznaczanie reprezentantów oraz porównywanie zbioru faktów z regułami bądź ich skupieniami proponowane podejście daje wysoką jakość odnajdowanych skupień, a tym samym reguł. Algorytm wyszukiwania hierarchicznego wymaga dalszego dostosowania do specyfiki grupowania reguł w bazie wiedzy. Autorzy spotkali się z tendencją do łańcuchowania grup reguł. Krótki opis każdej reguły oraz ich mała rozróżnialność względem siebie mogą przyczynić się do zaburzenia równomierności dendrogramu (zdarzało się i tak, że w jednym z poddrzew na każdym poziomie była tylko jedna reguła, a w drugim – pozostałe). Po przeanalizowaniu sytuacji zauważono niepokojący fakt małej rozróżnialności wartości w macierzy podobieństwa, budowanej na początku działania algorytmu. Przykładowo dla bazy Abalone liczba komórek macierzy podobieństwa wynosiła 7 138 531, gdzie dla całej bazy występowały tylko 43 różne wartości podobieństwa reguł (a więc wiele było par tak samo podobnych), i o kolejności tworzenia skupień decydowała kolejność pary reguł (lub w późniejszym etapie – skupień reguł) w bazie wiedzy. Dalsze badania będą miały na celu wyeliminowanie tego zjawiska. Optymistycznym wnioskiem jest poprawa algorytmu wyszukiwania reguł metodą ścieżki najbardziej obiecującej za pomocą autorskiej metody hybrydowej z większą wagą dla wspólnych deskryptorów. Wciąż duży problem stanowi ocena jakości tworzonych skupień. Z tego względu autorzy w dalszych pracach skupią się na poszukiwaniu modyfikacji dotychczasowych rozwiązań, jednocześnie wypracowane zostaną mechanizmy informowania użytkownika o możliwościach aktywowania reguł niepewnych, a tym samym – wnioskowania w warunkach wiedzy niepełnej. Planowane jest również wykorzystanie metody współczynników pewności CF oraz innych, omówionych w [1], w celu poprawnego zamodelowania niepewności.

BIBLIOGRAFIA

1. Nowak-Brzezińska A., Simiński R., Jach T., Xięski T.: Towards a practical approach to discover internal dependencies in rule-based knowledge bases. Rough Sets and Knowledge Technology, 2011.
2. Nowak-Brzezińska A., Wakulicz-Deja A.: Analiza efektywności wnioskowania w złożonych bazach wiedzy. Systemy Wspomagania Decyzji, 2007.
3. Jain A., Dubes R.: Algorithms for clustering data. Prentice Hall, New Jersey 1988.
4. Chandru V., Hooker J.: Optimization methods for logical inference. John Wiley & Sons, New York 1999.
5. Zadeh L., Kacprzyk J.: Fuzzy logic for the management of uncertainty. John Wiley & Sons, New York 1992.
6. Pawlak Z.: Rough set approach to knowledge-based decision suport. European Journal of Operational Research, 1997, s. 48÷57.

7. Salton G.: Automatic Information Organization and Retreival. McGraw-Hill, New York, USA 1975.
8. Jach T., Nowak-Brzezińska A.: Wnioskowanie w systemach z wiedzą niepełną. Studia Informatica, Vol. 32, No. 2A(96), Wydawnictwo Politechniki Śląskiej, Gliwice 2011, s. 377÷391.
9. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. Exit, Warszawa 2008.
10. Frank A., Asuncion A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA 2010, http://archive.ics.uci.edu/ml.
11. Kaufman L., Rousseeuw P. J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York 1990.
12. Latkowski R.: Wnioskowanie w oparciu o niekompletny opis obiektów. Praca magisterska, Wydział Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego, Warszawa 2001.
13. Myatt G.: Making Sense of Data A Practical Guide to Exploratory Data Analysis and Data Mining. John Wiley and Sons, Inc., New Jersey 2007.
14. Kumar V., Tan P., Steinbach M.: Introduction to Data Mining. Addison-Wesley, 2006.
15. Geiger D., Heckerman D.: Knowledge representation and inference in similarity networks and Bayesian multinets. Artificial Intelligence, 1996, s. 45÷74.
16. Towell G., Shavlika J.: Knowledge-based artificial neural networks. Artificial Intelligence, 1994, s. 119÷165.
17. Bazan J., Nguyen H. S., Nguyen S. H., Synak P., Wróblewski J.: Rough set algorithms in classification problems. Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, 2000, s. 49÷88.
18. Bazan J., Szczuka M., Wróblewski J.: A new version of rough set exploration system. Third International Conference – RSCTC, 2002, s. 397÷404.

Wpłynęło do Redakcji 8 stycznia 2012 r.

Abstract

The authors propose to use cluster analysing techniques (particularly clustering) to speed-up the process of finding rules to be activated in complex decision support systems with incomplete knowledge. Apart from that, the authors wish to inference within such decision support systems also using rules, of which premises were not fully covered by the facts. In order to achieve that, the AHC algorithm is used. Authors proposed the method to obtain the

optimal number of clusters (Figure 1). The comparison between AHC and mAHC algorithms in context of clustering rules is covered (Figure 2) and experimentally measured (Figures 3, 4). The authors also adapted Salton's most promising path method for a fast lookup of the rules. The parameters of this method were established experimentally (Figures 5, 6) and then further checked with more rule bases (Figure 7).

Adresy

Agnieszka NOWAK: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39,
41-200 Sosnowiec, Polska, agnieszka.nowak@us.edu.pl.
Tomasz JACH: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39,
41-200 Sosnowiec, Polska, tomasz.jach@us.edu.pl.