

METODY WYZNACZANIA WSPÓŁCZYNNIKA NIEPEŁNOŚCI WIEDZY W SYSTEMACH Z WIEDZĄ NIEPEŁNĄ

AGNIESZKA NOWAK-BRZEZIŃSKA, TOMASZ JACH

Uniwersytet Śląski, Instytut Informatyki

Streszczenie

W opracowaniu autorzy proponują użycie nowatorskiego rozwiązania – współczynników niepewności (IF) w celu zamodelowania niepewności wiedzy w Systemach Wspomagania Decyzji. Proponowane rozwiązanie bazuje na znanym z literatury podejściu współczynników pewności (CF). W pracy przedstawione są dotychczasowe rezultaty badań oraz wnioski z przeprowadzonych eksperymentów.

1. Wprowadzenie

W obliczu natłoku danych i informacji oraz znacznie zwiększonemu zapotrzebowaniu na szybką, dokładną i automatyczną analizę tychże, systemy wspomagania decyzji zyskują coraz większe znaczenie w dzisiejszej informatyce. Ich zastosowanie może być bardzo szerokie: od sterowania obiektami przemysłowymi, po wspomaganie lekarza w doborze trafnej diagnozy. Niestety, lawinowo rosnąca ilość przetwarzanych danych stawia coraz większe wyzwania dla projektantów SWD. Zarówno ogrom danych, które należy poddać analizie jak i niejednokrotnie ich niska jakość (braki, błędy, sprzeczności) powoduje, że proces automatycznego wspomaganie decyzji jest niezwykle trudny. Z tego powodu autorzy postanowili wypracować model służący wspomaganie procesu wnioskowania w systemach z wiedzą niepełną. W dalszej części artykułu przedstawione zostanie autorskie rozwiązanie służące poprawnemu zamodelowaniu procesu niepewności wiedzy. Przedstawione zostaną również eksperymenty obliczeniowe oraz podobne rozwiązania, które stanowiły inspiracje w tworzeniu przedstawianego rozwiązania.

1.1. Systemy wspomagania decyzji

Pryncypium działania SWD jest wyciąganie logicznych i poprawnych wniosków z posiadanych danych (reguł oraz faktów). Formalnie, system wspomagania decyzji jest zdefiniowany następująco [6]:

$$SWD = \langle U, A, V, f \rangle$$

U – Skończony, niepusty zbiór reguł (uniwersum)

A – Skończony, niepusty zbiór atrybutów

$$C \cup D = A; C \cap D = \emptyset$$

C – zbiór atrybutów warunkowych

D – zbiór atrybutów decyzyjnych

V – zbiór wartości atrybutów

$$V = \bigcup_{a \in A} V_a$$

V_a – zbiór wartości atrybutu V

$f: U \times A \rightarrow V$ – funkcja informacji

W dalszej części pracy przyjęto ponadto oznaczenia r_i jako i -tą regułę w systemie odpowiadającą przyjętej w klasycznych SWD postaci klauzuli Horna [1], gdzie każdy literał

z części przesłankowej i decyzyjnej tworzony jest w oparciu o zbiór atrybutów A oraz zbiory 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: $r_i = d_1 \wedge d_2 \wedge \dots \wedge d_m \Rightarrow DEC_i$.

W klasycznych systemach wspomaganie decyzji wnioskowanie wprzód polega na uaktywnianiu wszystkich możliwych reguł, których część przesłankowa ma w całości pokrycie w zbiorze faktów (ściślej: wszystkie deskryptory będące przesłankami reguły znajdują się również w zbiorze faktów). Dzięki takiemu warunkowi, zachowana zostaje zasada *modus ponendo ponens*. Konkluzja uaktywnionej reguły zostaje dopisana do zbioru faktów i cały proces jest powtarzany aż do momentu, gdy nie da się uzyskać nowej wiedzy.

Sytuacja ta jest sytuacją modelową, jednakże w rzeczywistych zastosowaniach niezmiernie często zdarza się, że w całym systemie nie ma reguły, której wszystkie przesłanki są spełnione. W takiej sytuacji klasyczny system wspomaganie decyzji nie spełni swojej roli i nie dostarczy żadnej nowej wiedzy.

Autorzy proponują skorzystanie z mechanizmów znanych z logiki rozmytej – uaktywniania także tych reguł, których nie wszystkie przesłanki są spełnione. Dzięki temu zostanie dostarczona nowa wiedza, oznaczona jako niepewna.

Autorzy w swych poprzednich pracach [12-17] proponują użycie mechanizmów analizy skupień w celu stworzenia grup reguł najbardziej podobnych do siebie. Dzięki temu skrócony zostanie czas wyszukiwania reguł w dużych systemach. Proponuje się również wprowadzenie wnioskowania w warunkach niekompletnej wiedzy opartego o uaktywnianie reguł, których nie wszystkie przesłanki są spełnione.

1.2. Pojęcia niepewności i niepełności wiedzy

Niepełność wiedzy rozumiana jest przez autorów jako sytuacja, gdy dla danej reguły nie wszystkie przesłanki są faktami w bazie wiedzy. Pojęcie to będzie stosowane w przypadku, gdy system wspomaganie decyzji nie będzie w stanie wykonać klasycznego wnioskowania ze względu właśnie na brak pełności wiedzy.

Pewien stopień niepewności wiedzy przypisywany jest konkluzji reguły, której nie wszystkie przesłanki były faktami w bazie wiedzy, a która została uaktywniona w procesie wnioskowania z użyciem współczynników IF.

1.3. Efektywność wnioskowania

Efektywność wnioskowania będzie przez autorów zwiększana dwutorowo. W pierwszej części, zwiększona zostanie szybkość znajdowania i uaktywniania reguł dzięki grupowaniu reguł za pomocą algorytmów mAHC oraz AHC. Grupowanie jest mechanizmem znanym z analizy skupień. Zostało ono z powodzeniem zaimplementowane w przedstawianym systemie. Grupowanie reguł w skupienia za pomocą algorytmów hierarchicznych pozwala na stworzenie struktury drzewiastej reguł (dendrogramu) znacznie zwiększającej szybkość wyszukiwania konkretnej reguły w stosunku do wyszukiwania liniowego w płaskiej bazie wiedzy. Szczegółowe rozwiązania tego problemu znajdują się w poprzednich pracach autorów.

Z drugiej strony, przewiduje się zdolność systemu do uaktywniania reguł w warunkach wiedzy niepełnej. Grupowanie reguł pozwala na wyznaczenie skupień reguł o największym stopniu pokrycia zbioru faktów w przypadku niepełności wiedzy, a co za tym idzie – przeprowadzenie procesu wnioskowania nawet w przypadku, gdy żadna z reguł nie była w pełni pokryta zbiorem faktów. W przeciwnym wypadku jednak, gdy w grupie najbardziej podobnej do zbioru faktów znajdują się reguły, których tylko część przesłanek jest spełniona, proponowane będzie uaktywnienie tychże reguł z określeniem ich współczynnika niepełności. Konkluzje otrzymane

w ten sposób będą oznaczone jako niepewne, a ich stopień niepełności (*IF*, ang. *incompleteness factor*) będzie wyznaczany w sposób opisany przez autorów w dalszej części pracy. Formalnie, autorzy rozpatrują trzy rodzaje reguł wchodzących w skład bazy wiedzy:

- Reguły pewne
 - $IF(\text{reguła}) = 1$
- Reguły niepewne
 - $IF(\text{reguła}) \in [\text{wsp. jakości}; 1]$
- Reguły nieuaktywniane
 - $IF(\text{reguła}) \in [0; \text{wsp. jakości}]$

Każdorazowo określony zostanie współczynnik jakości, który oddzielał będzie reguły niebrane pod uwagę w wnioskowaniu.

2. Proponowane rozwiązanie

Dotychczas, autorzy próbowali wypracować optymalną metodą grupującą reguły w bazie wiedzy [12-17]. Dokonana została analiza hierarchicznych algorytmów grupujących AHC oraz mAHC [2-3]. Sprawdzono również jak różne parametry algorytmów grupujących wpływają na jakość otrzymywanych skupień oraz efektywność wnioskowania. Zbadano wpływ: kryteriów łączenia skupień w grupy, tworzenia reprezentantów skupień reguł. Zaproponowano modyfikację metody ścieżki najbardziej obiecującej wywodzącej się z prac Saltona [4] oraz przedstawiono propozycję wyznaczania poprawnej liczby skupień w danych. Przykłady reguł oraz konkretne rozwiązania użyte przez autorów zostały przedstawione w poprzednich pracach [12-17].

2.1. Współczynnik niepewności wiedzy IF

W wyniku wnioskowania, odnajdywana jest grupa będąca najbardziej podobna (ze względu na zdefiniowane wcześniej kryterium podobieństwa) w stosunku do aktualnie rozpatrywanego zbioru faktów. Po jej odnalezieniu, system przystępuje do analizy reguł wchodzących w skład tej grupy. Jeśli znajdują się tam reguły mające pełne pokrycie w zbiorze faktów, te zostają uaktywniane. W przypadku, gdy nie istnieje żadna taka reguła, której wszystkie przesłanki są prawdziwe, uaktywniane są reguły mające tylko częściowe pokrycie w zbiorze faktów. W celu odróżnienia wiedzy wyznaczonej przez reguły pewne od reguł niepewnych wprowadza się współczynnik pełności IF (ang. *incompleteness factor*). Współczynnik ten będzie wyznaczany w następujący sposób:

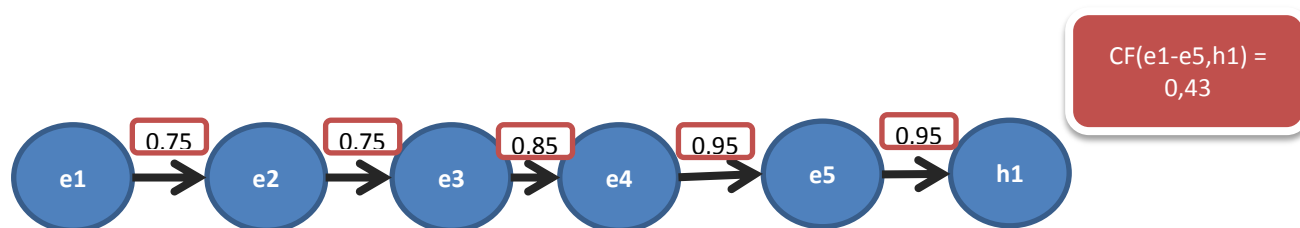
$$IF(DEC_i) = \frac{\sum_j IF(f_j)}{\text{card}(D_i)}; f_j \in (F \cap D_i)$$

Gdzie $IF(d_i)$ to współczynnik niepełności *i*-tej przesłanki, $IF(f_j)$ to współczynnik niepełności *j*-tego faktu będącego częścią wspólną zbioru deskryptorów D_i oraz zbioru faktów (F).

Fakty pewne, znane wcześniej lub też dopisane do zbioru faktów po uaktywnieniu reguł pewnych posiadają z definicji wartość współczynnika IF równą 1. Współczynnik ten ma wartości z przedziału $[0;1]$, co pozwala na łatwą interpretację uzyskanych wyników. Przykładowo, dla zbioru faktów $\{(a, 2), (b, 3), (c, 6)\}$ oraz następujących reguł: $R1: (a, 2) \wedge (b, 3) \wedge (c, 4) \rightarrow (D, 1)$; $R2: (a, 2) \wedge (b, 4) \wedge (c, 5) \rightarrow (D, 2)$ współczynniki IF tychże reguł wynoszą odpowiednio $IF(R1) = 0,67$ oraz $IF(R2) = 0,33$. Autorzy proponują również wprowadzenie średniej wartości współczynnika IF dla całej odnalezionej grupy. W przedstawianym przykładzie wartość tego współczynnika dla grupy złożonej z reguł $R1$ oraz $R2$ wynosiłaby 0,5.

Współczynnik IF powstał dzięki inspiracji współczynnikami CF zaproponowanymi w systemie MYCIN [8]. Był on odpowiedzią na trudne do zaimplementowania metody probabilistyczne. Formalnie wartość współczynnika CF to różnica pomiędzy miarą wiarygodności (MB, ang. measure of belief) oraz miarą wątpliwości (MD ang. measure of disbelief) [9-10]. MB opisuje wiedzę systemu o zajściu danej hipotezy h znając przesłankę e , podczas gdy MD – opisuje niewiedzę.

Metoda współczynników CF wprowadziła bardzo prosty sposób obliczania i propagacji CFów. Dzięki temu implementacja tegoż systemu była stosunkowo łatwa. Niestety, ze względu na ten sam sposób propagacji współczynników CF drastycznie spada wartość współczynnika CF konkluzji w trakcie wnioskowania na długich łańcuchach przesłanek i konkluzji. Spadek ten utrudnia prawidłową interpretację końcowych wyników oraz powoduje problemy związane z niewystarczającą precyzją obliczeń zmiennoprzecinkowych.



Rys. 1. Długi łańcuch wnioskowania i spadek wartości współczynnika CF

Autorzy proponują użycie współczynnika IF, który stanowi łatwą miarę do modelowania niepewności i niepełności wiedzy zależną tylko od stopnia pełności wiedzy zapisanej w zbiorze faktów. Umożliwia korzystanie z zalet mechanizmów grupowania (m.in. wybór grupy o największym współczynniku IF). Umożliwia się również sterowanie „jakością” wnioskowania. Do użytkownika systemu należy decyzja, czy liczy się z możliwością otrzymania wiedzy niższej jakości, czy woli nie otrzymać żadnych nowych informacji w wyniku procesu wnioskowania.

2.2. Propozycja algorytmu wnioskowania

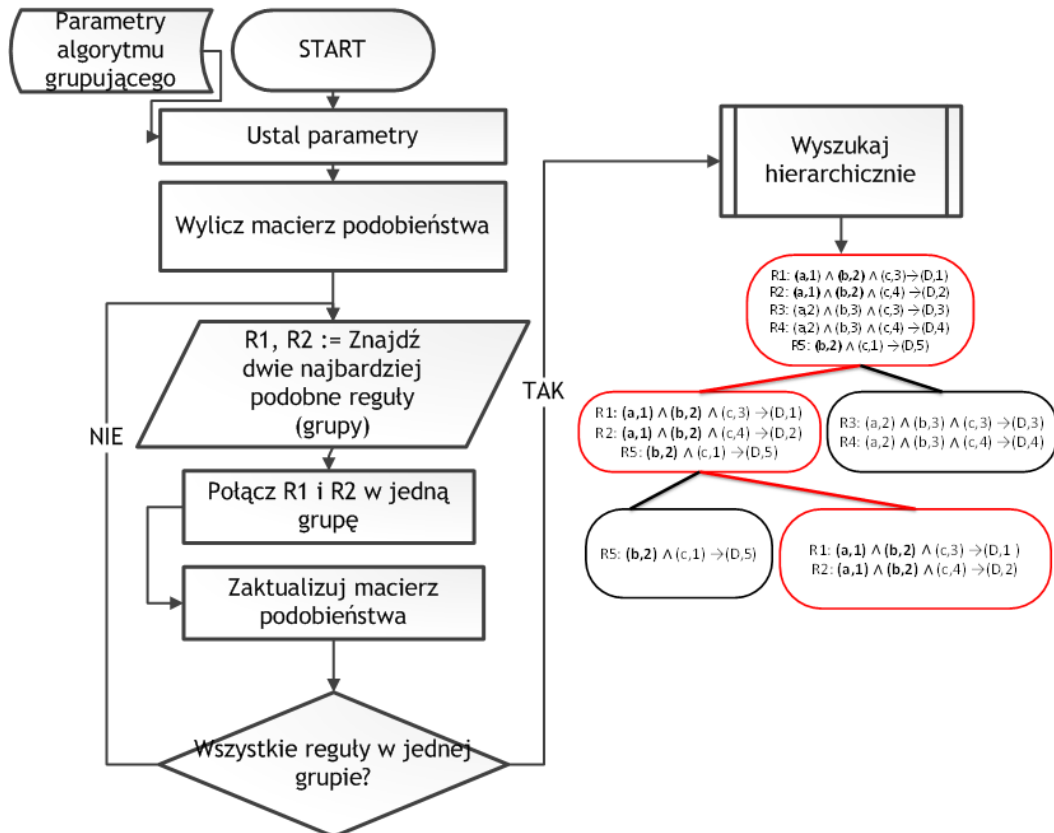
W celu optymalizacji efektywności wnioskowania, autorzy proponują użycie algorytmu AHC do grupowania reguł wchodzących w skład bazy wiedzy. Schematyczny proces działania systemu przedstawiony jest na rysunkach 2 oraz 3.

Rozpoczęcie pracy algorytmu następuje po pobraniu od użytkownika parametrów pracy (m.in. wybór metody wiązania skupień, miary odległości, wartości progowych współczynnika IF, zakładanej głębokości i innych). Po ich ustaleniu, wyliczana jest kwadratowa macierz podobieństwa. Ma ona rozmiar równy liczbie reguł znajdujących się w systemie. Na przecięciu i -tej kolumny oraz j -tego wiersza wyliczana jest wartość podobieństwa i -tej oraz j -tej reguły z bazy wiedzy. W kolejnym kroku budowana jest struktura hierarchiczna (dendrogram) reguł poprzez odnajdywanie dwóch najbardziej podobnych skupień oraz łączeniu ich w jedno. Proces ten powtarzany jest do momentu połączenia wszystkich skupień w jedną dużą grupę. Proces ten musi być powtórzony tylko raz dla każdej bazy wiedzy.

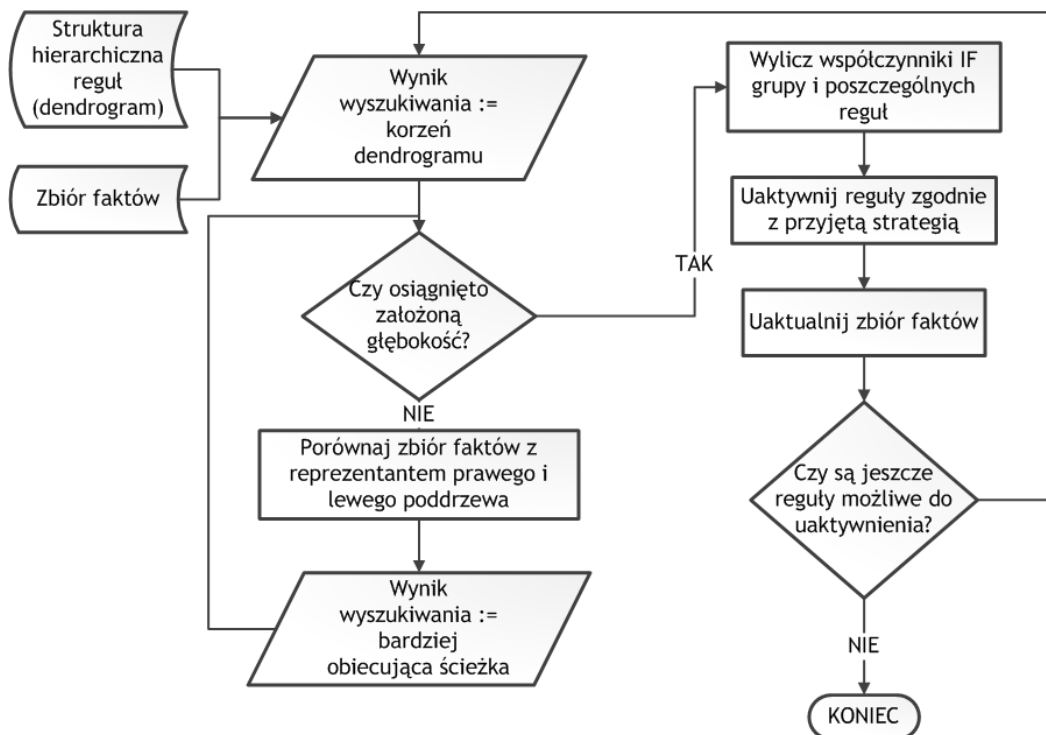
Otrzymaawszy dendrogram, algorytm przystępuje do właściwego procesu wnioskowania korzystając z utworzonej struktury hierarchicznej. Rozpoczynając od korzenia drzewa, aktualny zbiór faktów jest porównywany z reprezentantem lewego i prawego poddrzewa. Do dalszej analizy wybierana jest ścieżka bardziej obiecująca. Krok ten powtarzany jest do momentu uzyskania zakładanej głębokości. W wyniku działania, zwracana jest grupa obiektów najbardziej podobnych w stosunku do początkowego zbioru faktów. Uaktywniane zostają zarówno reguły pewna jak i

niepewne, a ich konkluzje zostają dodane do zbioru faktów. Wyliczane są wartości współczynników IF.

System działa do momentu, w którym uaktywnianie dalszych reguł spowodowałoby dopisanie do bazy faktów wiedzy o zbyt niskiej jakości w stosunku do przyjętego minimum.



Rys. 2. Ogólny schemat algorytmu wnioskowania

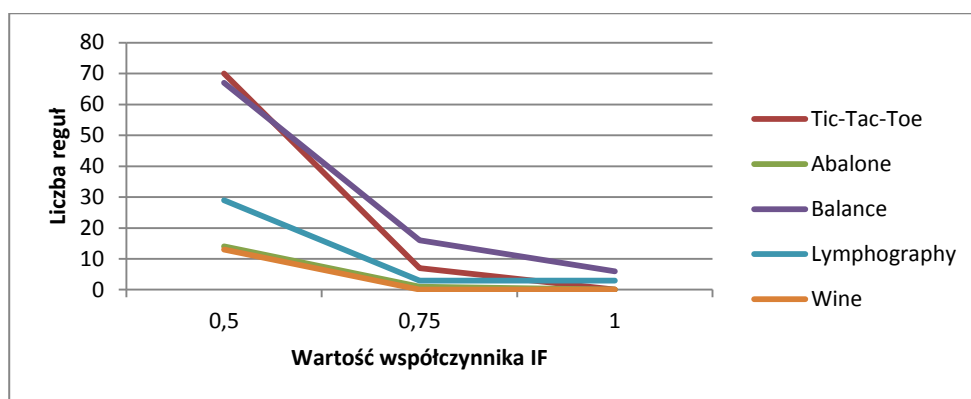


Rys. 3. Schemat wyszukiwania hierarchicznego

3. Eksperymenty obliczeniowe

Podobnie jak w poprzednich pracach, autorzy wykonali eksperymenty obliczeniowe korzystając z ogólnodostępnych baz zawartych w Machine Learning Repository. Skorzystano z baz o różnym stopniu złożoności: Wine, Lymphography, Balance, Abalone, Tic-Tac-Toe. W celu przeprowadzenia eksperymentów, za pomocą pakietu RSES wygenerowano reguły minimalne w oparciu o ww. bazy. Następnie reguły te poddano mechanizmowi grupowania i wnioskowania opisanego w tym opracowaniu.

Po wygenerowaniu reguł minimalnych w pierwszym zaproponowanym eksperymencie wylosowano około 10% deskryptorów występujących w całym systemie. Deskryptory te zostały dopisane do początkowo pustego zbioru faktów. Następnie sprawdzono pokrycie reguł w bazie wiedzy takim zbiorem faktów. Sprawdzono tym samym jaki procent reguł mógłby zostać uaktywniony w przypadku wnioskowania tylko na regułach pewnych oraz jak obniżenie progu pokrycia reguły wpływa na liczbę reguł możliwych do uaktywnienia. Sprawdzono wartości 0,5 oraz 0,75 współczynnika IF (czyli przyjęto, iż np. dla $IF=0,75$ tylko 75% przesłanek musiało być spełnionych, aby uznać regułę za możliwą do uaktywnienia). Wyniki przedstawia rys. 4.



	0,5	0,75	1
Tic-Tac-Toe	70	7	0
Abalone	14	1	0
Balance	67	16	6
Lymphography	29	3	3
Wine	13	0	0

Rys. 4. Liczba możliwych do uaktywnienia reguł a minimalny współczynnik IF

Jak widać, jeśli dopuszczone zostanie wnioskowanie tylko na regułach, których wszystkie przesłanki są spełnione (a więc wartość współczynnika IF wynosi 1), to aż w trzech przypadkach nie otrzymamy żadnej nowej wiedzy. Jeśli tylko obniżymy minimalny próg współczynnika IF do wartości 0,75 system będzie w stanie zwrócić informacje, które mogą posłużyć do uszczegółowienia zapytania, a co za tym idzie – do skutecznego wspomaganie decyzji podejmowanej przez człowieka-eksperta.

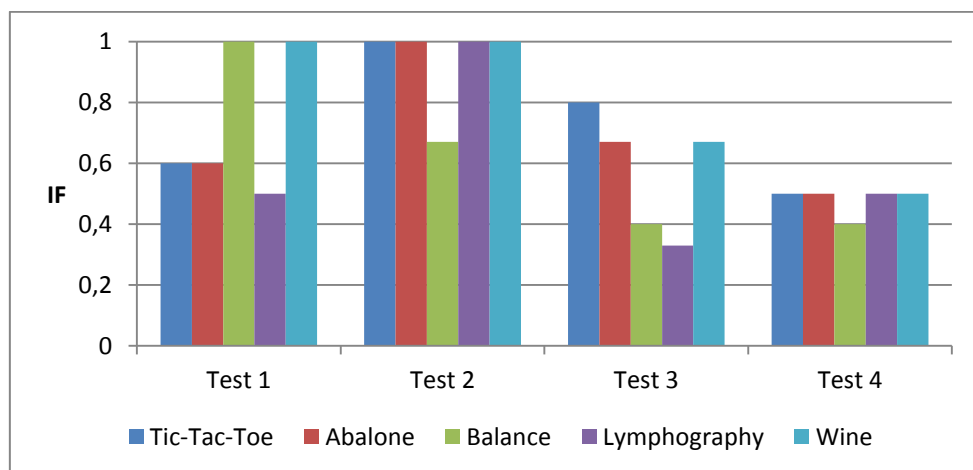
3.1. Maksymalny współczynnik IF wewnątrz odnalezionnej grupy

Kolejnym eksperymentem było sprawdzenie maksymalnego współczynnika IF wewnątrz odnalezionnej grupy. W celu jego zbadania, do systemu zawierającego już hierarchiczną strukturę reguł, zadano pytanie składające się z różnej liczby par atrybut-wartość. Pytanie takie tworzyło w

całości bazę faktów. W wyniku skorzystania a algorytmu ścieżki najbardziej obiecującej, system zwracał grupę reguł uznanych za najbardziej relewantne. W trakcie eksperymentu liczone wartości współczynników IF dla każdej reguły wchodzącej w skład skupienia, a następnie wybierano maksymalną wartość.

Test został podzielony na cztery przypadki testowe, z których każdy był powtórzony czterokrotnie, a uzyskane wyniki zostały uśrednione. Przypadki testowe można zaklasyfikować następująco:

- T1 oraz T2 - baza faktów składała się ze wszystkich przesłanek losowo wybranej reguły z bazy.
- T3 – baza faktów to około 80% przesłanek losowo wybranej reguły.
- T4 – baza faktów to około 50% przesłanek losowo wybranej reguły.



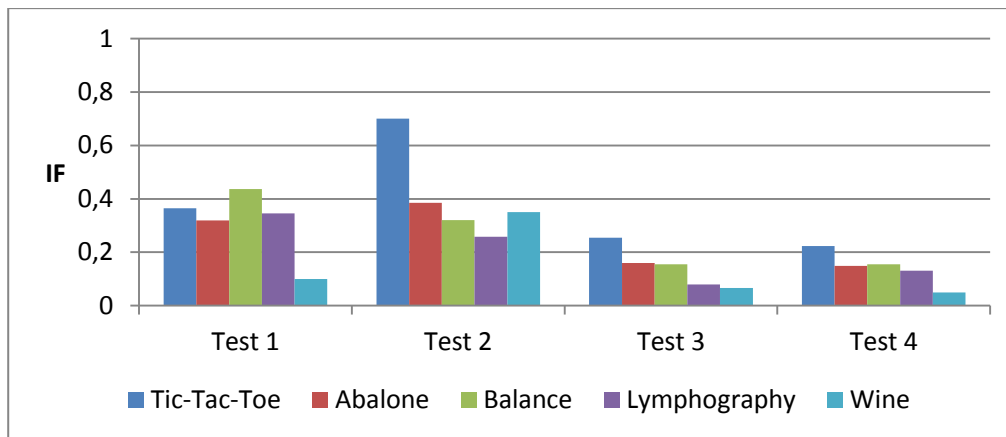
Rys. 5. Wyniki pierwszego eksperymentu

Wyniki eksperymentu pozwalają sądzić, iż proponowany system dobrze radzi sobie z odnajdowaniem grup reguł relewantnych w stosunku do zadawanego zbioru faktów. Należy tutaj podkreślić, że w teście nr 3 maksymalna możliwa wartość współczynnika IF wynosi około 0,8, a w teście 4 – około 0,5.

Jak widać, algorytm daje rezultaty zbliżone do optimum w większości przypadków. Dla większości baz udało się znaleźć reguły najbardziej odpowiednie do uaktywnienia.

3.2. Średni współczynnik IF wewnątrz odnalezionnej grupy

Kolejny z eksperymentów miał na celu zbadanie średniej wartości współczynnika IF wewnątrz odnalezionnej grupy. Sposób przeprowadzenia eksperymentu był analogiczny do poprzedniego, z tą różnicą, iż badano tutaj średnią wartość współczynnika IF wewnątrz odnalezionnej grupy. Dzięki temu możliwym jest zbadanie, czy system generuje skupienia o wysokiej jakości, w których reguły są spójne i mają dużą część wspólnych przesłanek. Wyniki przedstawione są na rysunku 5.



Rys. 6. Wyniki drugiego eksperymentu

Otrzymane wyniki pokazują, iż system zwraca stosunkowo liczne skupienia, dzięki czemu średnia wartość współczynnika IF wewnątrz skupienia jest niska. Autorzy sugerują tutaj dalsze badania pozwalające na większe odróżnienie skupień od siebie, a co za tym idzie – na wzrost średniej wartości współczynnika IF wewnątrz odnalezionnej grupy.

4. Wnioski oraz kierunki dalszych badań

Autorzy zaproponowali skorelowanie mechanizmów analizy skupień, elementów logiki rozmytej oraz systemów wspomagania decyzji w celu rozwiązania problemu niepełności wiedzy w trakcie wnioskowania w SWD. Proponowane rozwiązanie pozwala na uaktywnianie reguł pewnych (przy wiedzy pełnej) oraz reguł niepewnych (przy wiedzy niepełnej) dzięki określeniu współczynnika niepełności wiedzy (IF), a tym samym otrzymanie nowej wiedzy z systemu nawet w czasie, gdy klasyczne algorytmy wnioskujące nie dadzą żadnych rezultatów.

Przedstawiony współczynnik IF stanowi także alternatywę przeszukiwania struktury skupień reguł, które wcześniej realizowano metodą węzła najbardziej obiecującego. Dzięki wyliczaniu średniej wartości IF grup reguł, możliwym jest wybranie tej, która będzie najbardziej relewantna w stosunku do aktualnie rozpatrywanego zbioru faktów.

Najważniejszym jednak elementem jest możliwość uzyskania dodatkowej wiedzy z systemu dzięki uaktywnianiu reguł niepewnych.

Autorzy w swych dalszych badaniach pragną skupić się na optymalnym doborze parametrów algorytmu grupującego do zastosowania wraz z metodą współczynników IF.

Adnotacja

Autor otrzymał stypendium w ramach projektu DoktoRIS – Program stypendialny na rzecz innowacyjnego Śląska współfinansowanego przez Unię Europejską w ramach Europejskiego Funduszu Społecznego.

Bibliografia

- [1] Chandru, V., Hooker, J.: Optimization methods for logical inference. John Wiley & Sons, New York (1999).
- [2] Kaufman, L., Rousseeuw, P. J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990).
- [3] Myatt, G.: Making Sense of Data a Practical Guide to Exploratory Data Analysis and Data Mining. John Wiley and Sons, Inc., New Jersey (2007).

- [4]Salton, G.: Automatic Information Organization and Retrieval. McGraw-Hill, (1975).
- [5]Bazan, J., Szczuka, M., Wróblewski, J.: A new version of rough set exploration system. In : Third International Conference on RSCTC. Springer-Verlag, Malvern, PA (2002).
- [6]Pawlak, Z.: Rough set approach to knowledge-based decision suport. European Journal of Operational Research, (1997).
- [7]Frank, A., Asuncion, A.: UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>], Irvine, CA: University of California, School of Information and Computer Science, (2010).
- [8]Buchanan, B.G.; Shortliffe, E.H.: Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison-Wesley, (1984).
- [9]Dempster, A.P.: A generalization of Bayesian inference. Journal of the Royal Statistical Society, Series B 30 (1968).
- [10]Shafer, Glenn.: A Mathematical Theory of Evidence. Princeton University Press, (1976).
- [11]Reichgelt H.: Knowledge Representation: An AI Perspective. Ablex Publishing Corporation, New Jersey, USA, (1991).
- [12]Wakulicz-Deja A., Nowak – Brzezińska A., Jach T.: Inference processes using incomplete knowledge in Decision Support Systems – chosen aspects. Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science, (2012).
- [13]Jach T., Nowak-Brzezińska A.: Wybrane aspekty wnioskowania w systemach z wiedzą niepełną. Studia Informatica, Zeszyty Naukowe Politechniki Śląskiej, Vol. 33, Number 2A (2012).
- [14]Jach T., Nowak-Brzezińska A.: Wnioskowanie w systemach z wiedzą niepełną. Studia Informatica, Zeszyty Naukowe Politechniki Śląskiej, Vol. 32, Number 2A, (2011).
- [15]Wakulicz-Deja A., Nowak – Brzezińska A., Jach T.: Inference processes in decision support systems with incomplete knowledge. Rough Sets and Knowledge Technology, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, (2011).
- [16]Nowak – Brzezińska A., Jach T., Xięski T.: Wybór algorytmu grupowania a efektywność wyszukiwania dokumentów, Studia Informatica, Zeszyty Naukowe Politechniki Śląskiej, Vol. 31, Number 2A, (2010).
- [17]Nowak-Brzezińska A., Jach T., Xięski T., Analiza hierarchicznych i niehierarchicznych algorytmów grupowania dla dokumentów tekstowych, Studia Informatica, Zeszyty Naukowe Politechniki Śląskiej, Vol. 30, No. 2A, (2009).