STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Agnieszka NOWAK-BRZEZIŃSKA, Alicja WAKULICZ-DEJA
Uniwersytet Śląski, Instytut Informatyki

WYBÓR MIARY PODOBIEŃSTWA A EFEKTYWNOŚĆ GRUPOWANIA REGUŁ W ZŁOŻONYCH BAZACH WIEDZY

Streszczenie. Praca przedstawia wyniki eksperymentów analizujących wpływ wybranej metryki podobieństwa na efektywność grupowania reguł w złożonych bazach wiedzy. Inna metryka podobieństwa to inna struktura skupień reguł, a co za tym idzie, inny przebieg procesów wnioskowania dla takich struktur.

Słowa kluczowe: analiza skupień, regułowe bazy wiedzy, miary podobieństwa

CHOOSING A PROPOPER METRIC AND CLUSTERING RULES IN COMPOSITED KNOWLEDGE BASES EFFICIENCY

Summary. The paper presents the results of the experiments analyzing an influence choosed similarity measure on the efficiency of rules (from composited knowledge bases) clustering. In authors oppinion. Each time when we use different measure we achieving different structure of rules' clusters and the results of inference process.

Keywords: cluster analysis, rules knowledge bases, similarity measure

1. Wprowadzenie

Grupowanie reguł w złożonych bazach wiedzy pozwala realizować m.in. zadanie modularyzacji bazy wiedzy, tak ważne dla efektywności systemów wspomagania decyzji. Okazuje się, że podział całej bazy wiedzy (o dużej liczbie reguł w przypadku złożonych baz wiedzy) na mniejsze grupy, wewnątrz których reguły są do siebie bardzo podobne (bądź to ze względu na część warunkową bądź decyzyjną) istotnie skraca czas wnioskowania. Dzieje się tak dlatego, że w procesach wnioskowania poszukiwane są takie reguły do uaktywnienia, które w danym momencie posiadają potwierdzone przesłanki albo szukaną konkluzję. Jeśli

zamiast przeszukiwać wszystkie reguły, jedna po drugiej, znajdziemy grupę reguł, która powinna być przeszukana i tylko dla niej zostanie przeprowadzony proces wnioskowania, skrócenie czasu będzie oczywiste. Rzecz w tym, by właściwie reguły (faktycznie podobne) zgrupować oraz przypisać im etykiety (najlepiej opisujące takie grupy i jednocześnie najlepiej je różnicujące od innych grup w bazie wiedzy). Niezwykle ważny zatem już na początku jest wybór odpowiednich metod grupowania reguł spośród szerokiego grona znanych algorytmów analizy skupień. Wyniki analiz możliwości zastosowania wybranych algorytmów analizy skupień zarówno z zakresu metod k-optymalizacyjnych, jak i metod hierarchicznych były treścią prac [8],[9] oraz [3].

Efektywność wnioskowania na grupach reguł nie zależy jedynie od wybranego algorytmu analizy skupień, choć oczywiście wybrana metoda ma na pewno ogromne znaczenie, bowiem to od niej zależy ostateczna struktura bazy wiedzy. Należy mieć świadomość faktu, że każda metoda grupowania bazuje na kryterium podobieństwa (bądź odległości). Kryterium to jest dlatego decydującem, że skupieniem obiektów chcemy nazywać grupę obiektów do siebie podobnych ze względu na pewne cechy opisujące te obiekty [1,2]. I tak, owe podobieństwo jest mierzone poprzez zastosowanie wybranej miary podobieństwa. Oczywiście w przestrzeniach metrycznych wolimy stosować metryki odległości, gdyż one w bardzo prosty sposób pozwalają decydować o tym, czy dany obiekt powinien należeć do skupienia *A* czy też może bardziej do skupienia *B*. Oczywiście mierzona jest wówczas odległość każdego obiektu do każdego skupienia i ostatecznie obiekt przydzielany jest do tego skupienia, do którego ma mniejszą odległość w sensie geometrycznym. Pojęcie odległości można bardzo łatwo przełożyć na pojęcie podobieństwa, gdy ustalimy, że tak naprawdę duża odległość świadczy jednocześnie o małym podobieństwie dwóch obiektów, i odwrotnie, mała odległość musi oznaczać duże podobieństwo wzajemne obiektów. Ponieważ w życiu codziennym dużo częściej używamy pojęcia podobieństwa aniżeli odległości, zupełnie naturalna jest próba przełożenia tegoż pojęcia również na aspekt analizy danych. W świecie rzeczywistym nigdy nie mówimy, że dwaj niemal tak samo zachowujący się klienci banku cechują się małą odległością, lecz mówimy o nich, że są ”podobni”. Stosowanie wybranych miar czy to podobieństwa, czy odległości jest jeszcze uzależnione od typu danych, które analizujemy. Otóż okazuje się, że wybrana metryka ma ogromny wpływ na uzyskaną potem jakość grupowania. Niniejsza praca ma właśnie na celu wykazać, jak wiele zależy od wybranej ostatecznie miary.

2. Grupowanie reguł w bazach wiedzy

W klasycznych regułowych bazach wiedzy efektywność w sensie szybkości działania systemu spada wówczas, gdy wzrasta rozmiar bazy wiedzy. System wspomagania decyzji jest wtedy efektywny, gdy wynikiem wnioskowania jest bądź wyprowadzona nowa wiedza (nowe fakty) bądź potwierdzenie prawdziwości hipotezy głównej. Im bardziej więc kompletna jest baza wiedzy, tym większe są szanse na osiągnięcie takich celów. A więc rozmiar bazy wiedzy jest czynnikiem niewątpliwie problematycznym, gdyż jednocześnie warunkuje efektywność wnioskowania ale i wydłuża czas przeszukiwania bazy wiedzy. Skoro nie da się zmniejszyć rozmiaru bazy wiedzy, a zakładamy, że będziemy mieć do czynienia z dużymi zbiorami reguł, powinniśmy zmienić metodę przeszukiwania bazy reguł na taką, która pozwoli skrócić czas przeszukiwania. Cel taki będzie możliwy do osiągnięcia, jeśli zmianie ulegnie struktura bazy wiedzy. Analiza skupień pozwoli nam utworzyć w bazie wiedzy grupy reguł spójnych. To dalej sprawi, że w procesach wnioskowania nie będą przeszukiwane wszystkie reguły, lecz jedynie reprezentanci grup (skupień) reguł. W ten sposób relatywnie szybko znaleziona zostanie grupa, która daje się pokryć z obserwacjami zadanymi na wejściu systemu i uaktywniane będą tylko z tej grupy reguły relewantne do podanych informacji. Nie ma wątpliwości, że proces taki powinien wpłynąć pozytywnie na efektywność systemu, a wyniki doświadczeń stały się treścią niniejszej pracy.

2.1. Hierarchiczna baza wiedzy

Hierarchiczna baza wiedzy tworzona jest przy użyciu aglomeracyjnego algorytmu analizy skupień, który pozwala budować grupy reguł podobnych ze względu na części warunkowe. W pracach [8, 9] oraz [3] przedstawiono koncepcję modelu bazy wiedzy przy użyciu dwóch algorytmów *AHC* oraz zmodyfikowanego algorytmu *mAHC*. Zakładamy, że mamy do czynienia ze złożoną bazą wiedzy, a więc takim zbiorem reguł, który zawiera dużą liczbę elementów (reguł), a dodatkowo dopuszczamy możliwość zagnieżdżania się reguł. Oznacza to mniej więcej tyle, że dany atrybut może raz wystąpić w części warunkowej a innym razem w części decyzyjnej reguły. Jest to o tyle istotne, iż stosowane w praktyce systemy z regułowymi bazami wiedzy dotąd operowały jedynie na tzw. regułach płaskich, a więc takich, gdzie własność zagnieżdżania nie była dopuszczana. Konieczność realizacji zjawiska zagnieżdżania się reguł jest naszym zdaniem o tyle istotna, że wiedza ekspercka najczęściej ma charakter łańcucha przyczynowo-skutkowego, w którym pewne decyzje są warunkami niezbędnymi do spełnienia, aby możliwe było podjęcie kolejnych decyzji. Takiej bazy wiedzy, w której konkluzja jednej reguły będzie jednocześnie przesłanką w innej regule, nie da się zapisać inaczej niż w postaci reguł zagnieżdżonych. Bazę wiedzy z regułami

zagnieżdżonymi będziemy dalej nazywać złożoną bazą wiedzy (ang. composited knowledge base) [8,9].

Formalizując definicję złożonej (hierarchicznej) bazy wiedzy powiemy, że system ze złożoną hierarchiczną bazą wiedzy będziemy definiować jako uporządkowaną szóstkę:

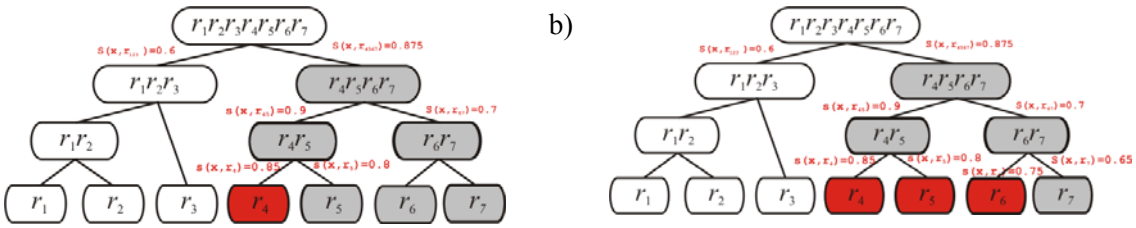
$$S_{HC} =< X, A, V, dec, F_{sim}, Tree >,$$

gdzie: $X = \{x_1,..,x_n\}$ to zbiór reguł zapisanych w postaci klauzul Horn'a, $A = \{a_1,..,a_m\}$ zbiór atrybutów warunkowych i decyzyjnych, zbiór wartości atrybutów $V_i = \bigcup_{a_i \in A} v_i, x_i \in V_i$ $V_i = \cup_{ai \in A} v_i$, $x_i \in V_i$, dla $1 \le i \le n, X = V_1 \times V_2 \times ... \times V_n, dec : x \to V_{dec}$, gdzie $V_{dec} = \{d_1,...,d_m\}$, będący zbiorem możliwych decyzji systemu, F_{sim}: $X \times X \to R \mid_{[0..1]}$ to funkcja podobieństwa tworząca drzewo, zaś dendrogram zapisany jest w $Tree = \{w_1,..,w_{2n-1}\} = \cup_{i=1}^{2n-1} w_i$. Każdy węzeł drzewa w_i jest definiowany za pomocą piątki: $w_i = \{d_i, c_i, f, i, j\}$, $i,j \in (1, 2,.., 2n-1)$, $d_i \in V_{dec}$, $c_i = X.d_i$ i c_i to odpowiednio wektory decyzji i warunków reguł, $f = F_{sim}(x_i, x_j) \to [0..1]$ to wartość podobieństwa między łączonymi skupieniami reguł, natomiast i oraz j to numery łączonych skupień [4].

2.2. Wyszukiwanie reguł w złożonej bazie wiedzy

Dzięki strukturze hierarchicznej skraca się czas poszukiwania w całej bazie wiedzy tych reguł, które mają zostać uaktywnione w zależności od wybranej metody wnioskowania. Wystarczy metodą przeszukiwania drzewa znaleźć węzeł najbardziej obiecujący (bądź kilka węzłów spełniających zadaną wartość progową) i przeszukiwać wybrany fragment drzewa [5, 3]. Znane są metody efektywnego przeszukiwania struktur drzewiastych (dodatkowo binarnych). Dokonują tego metody: pnia najbardziej obiecującego (rys. 1a) oraz minimalnej wartości progowej (rys. 1b). W metodach tych zawsze zaczynamy od korzenia i na każdym poziomie o wyszukiwaniu decyduje reguła decyzyjna, która oblicza podobieństwo (dopasowanie) i wybiera węzeł do dalszego przeszukiwania.

a) b)

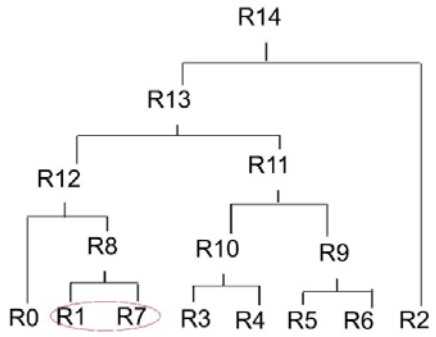


Rys. 1. Przeszukiwanie struktur drzewiastych: a) aetoda węzła najbardziej relewantnego, b) metoda minimalnej wartości progowej

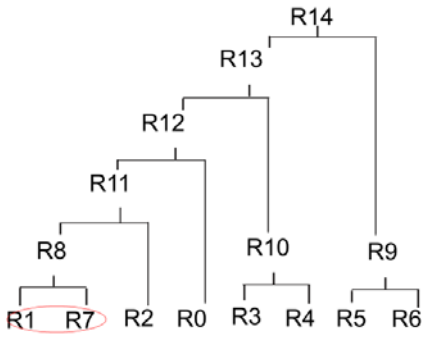
Fig. 1. Tree structures searching: a) the best node in the tree, b) the node with minimal treshold value

Zastosowanie proponowanych rozwiązań prowadzi do redukcji złożoności liniowej do złożoności logarytmicznej typowej dla tzw. przeszukiwania binarnego (połówkowego). Warto zwrócić uwagę na fakt, że w zależności od wybranej miary podobieństwa bądź

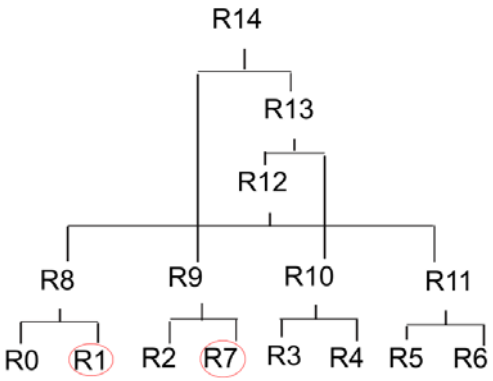
odległości przy tworzeniu skupień reguł, a także w zależności od wybranej metody łączenia skupień (metody: najbliższego sąsiada, najdalszego sąsiada i inne) inne są powstałe w efekcie drzewa skupień. Ma to niewątpliwie wpływ na efektywność przeszukiwania takich struktur drzewiastych. Rysunki 2-5 przedstawiają odpowiednie dendrogramy (inne dla każdej wybranej metryki podobieństwa bądź odległości) zbudowane algorytmem AHC dla metody „pojedynczego łączenia" skupień.



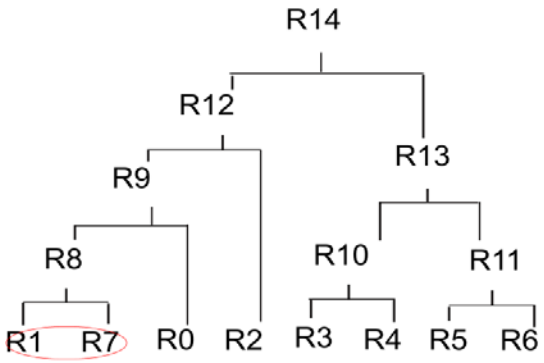
Rys. 2. Drzewo budowane miarą kosinusową
Fig. 2. Dendrogram for *cosine* measure



Rys. 3. Drzewo budowane miarą Euklidesową
Fig. 3. Dendrogram for *euclidean* measure



Rys. 4. Drzewo budowane miarą Gowera
Fig. 4. Dendrogram for *Gower* measure



Rys. 5. Drzewo budowane miarą nakładania
Fig. 5. Dendrogram for *ovlap* measure

2.3. Pomiar efektywności

Efektywność systemu z hierarchiczną bazą wiedzy będzie mierzona wartością dokładności procesu wnioskowania. Dokładność ta rozumiana będzie jako zdolność systemu do znajdowania reguł najbardziej przydatnych w procesie wnioskowania. Regułą relewantną będzie reguła bądź to pokrywająca zadany zbiór faktów, bądź taka, której konkluzja pokrywa się z celem wnioskowania. Zakładając, że w bazie wiedzy będą zadane pewne fakty (stanowiące tzw. wiedzę wejściową), proces wnioskowania polega na przeszukiwaniu struktury skupień reguł i znalezieniu pewnej grupy reguł bądź jednej reguły, która zostanie uaktywniona. Jeśli proces uaktywnienia reguły zakończy się sukcesem, wówczas mówimy o uzyskaniu wysokiej dokładności systemu. Dokładność nie będzie możliwa, gdy w procesie wyszukiwania reguł relewantnych zostanie znaleziona reguła, która ostatecznie nie zostanie

uaktywniona, a więc proces wnioskowania nie zakończy się pomyślnie. Nie będziemy dążyć do pełnej kompletności systemu, bowiem nie jest naszym celem znalezienie i uaktywnienie wszystkich reguł relewantnych. *Dokładność* (ang. *precision*) będziemy więc definiować jako zdolność systemu do niewyszukania reguł nierelewantnych lub inaczej – prawdopodobieństwo, że reguła nierelewantna nie zostanie wyszukana. Obliczyć ją można z następującego wzoru: $D_{HC} = a/a+b)$, gdzie odpowiednio: a – to liczba reguł relewantnych wyszukanych, b – to liczba reguł nierelewantnych wyszukanych, zaś c – to liczba reguł relewantnych nie wyszukanych [6].

3. Kryterium podobieństwa – wpływ miary na efektywność grupowania

Analiza chociażby pracy [7] pozwoli stwierdzić, że miar podobieństwa i odległości jest bardzo wiele, a ich skuteczność w ogromnym stopniu zależy od typu danych, które analizujemy. W niniejszej pracy zakładamy, że reguły w bazach wiedzy, które są naszymi danymi, opisane są przy użyciu zarówno cech nominalnych, jak i liczbowych. Gdy dysponujemy danymi nie tylko liczbowymi, okazuje się, że nie każda metryka będzie się dobrze zachowywać przy analizie danych. Rozważając dane wielocechowe musimy brać pod uwagę metryki, które potrafią porównać każdy typ danych. W eksperymentach wykonanych na regułowych bazach wiedzy analizowano wpływ trzech metryk na efektywność wnioskowania: miar podobieństwa kosiunusowego oraz Gowera, a także miarę odległości Euklidesowej.

3.1. Miary podobieństwa Gowera

Miara podobieństwa zaproponowana przez J.C. Gowera (1971 r.) dla obiektów o cechach ilościowych i jakościowych liczona jest wg wzoru:

$$p(o_i, o_j) = \frac{\sum_{k=1}^{n} s_{ijk} w_{ijk}}{\sum_{k=1}^{n} w_{ijk}},$$

gdzie waga w_{ijk} jest równa 0, gdy wartość *k*-tej zmiennej nie jest znana dla jednego lub dla obu obiektów o_i oraz o_j, natomiast 1 w przeciwnym przypadku. Oczywiście, jeśli $w_{ijk} = 0$, to także $s_{ijk} = 0$. Gdy jednak wagi są równe 0 dla wszystkich zmiennych, wartość $p(o_i,o_j)$ pozostaje nieokreślona. Z kolei wartość ocen podobieństwa obiektów o_j oraz o_i ze względu na *k*-tą zmienną (s_{ijk}) zależy od typu danych. Dla danych jakościowych s_{ijk} = 1, jeśli

porównywane obiekty nie różnią się ze względu na *k*-tą zmienną, w przeciwnym przypadku s_{ijk} przyjmuje wartość zero. Dla danych ilościowych:

$$s_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{R_k},$$

gdzie x_{ik} oraz x_{jk} to wartości *k*-tej zmiennej dla obu obiektów, R_k – różnica między maksymalną a minimalną wartością danej zmiennej. Niezależnie od wybranej metody wnioskowania okazuje się, iż bardzo ważna jest metryka przyjęta do grupowania reguł [2].

4. Eksperymenty

Eksperymenty wykonane na potrzeby niniejszej pracy mają na celu analizę wpływu wybranej metryki podobieństwa (bądź odległości) w procesie grupowania reguł, a potem w procesie wyszukiwania reguł w strukturze hierarchicznej, na efektywność pracy systemu. Okazuje się, że jeśli do grupowania użyjemy innej metryki niż do przeszukiwania struktury skupień reguł, nie możemy liczyć na wysokie wartości parametrów efektywności. Dodatkowo powiemy, że nie jest bez znaczenia, jaką metrykę wybraliśmy. Wykonane eksperymenty dotyczyły szerszego grona metryk, niż to wynika z przeglądnięcia tabel 1-6. Wśród analizowanych metryk były: odległość Euklidesowa, podobieństwo Gowera oraz miary: kosinusowa i nakładania. Z uwagi na ograniczenia objętościowe niniejszej pracy w tabelach 1-6 umieszczono jedynie część wyników. I tak, stosując miarę nakładania praktycznie we wszystkich przypadkach nie udało się efektywnie przeprowadzić procesu wnioskowania (dlatego pomijamy wyniki dla tej miary w tabelach 1-6). Nie wolno nie wspomnieć o tym, że na wartość parametru efektywności, którym w tym przypadku był parametr dokładności (o którym mowa w punkcie 2.3 niniejszej pracy), miały wpływ nie tylko wybrana metoda grupowania, metoda łączenia skupień oraz metryka podobieństwa bądź odległości. Istotną rolę w tego typu systemach odgrywa zawsze również to, jaki procent całej bazy wiedzy stanowią reguły relewantne (a więc te, dające się uaktywnić w procesie wnioskowania).

4.1. Źródło danych

Źródłem danych (a więc zbiorem reguł) podlegających grupowaniu algorytmem aglomeracyjnych był zbiór reguł generowany przy użyciu algorytmu indukcji częściowych reguł decyzyjnych [10,11]. W wyniku zastosowania algorytmu dla danej tablicy decyzyjnej (w formacie przyjętym w repozytorium UCI Machine Learning[12]) generowany jest zbiór reguł decyzyjnych o uogólnionej części warunkowej. Oczywiście nic nie stoi na przeszkodzie, by

reguły dla danej bazy wiedzy zostały pozyskane nie w sposób automatyczny poprzez algorytm indukcji reguł decyzyjnych, ale np. wprost od eksperta (bądź zespołu ekspertów). Wówczas możliwa jest sytuacja, w której wiedza eksperta będzie stanowić swego rodzaju łańcuch przyczynowo-skutkowy, a powstałe w ten sposób reguły będą mogły być zagnieżdżone.

4.2. Wyniki eksperymentów

Tabela 1

Wyniki dla miary Gowera- zbiór „spect_all”

set	rlavg	NR	Nn	Ndata	Nrr	NRrs	Precision	%KB
spect_0	3.18	267	16	1	38	133	0	3%
spect_0	3.18	267	12	9	6	218	1	2.25%
spect_0	3.18	267	6	9	4	116	0	1.12%
spect_001	3.18	267	24	1	38	195	0	4.5%
spect_001	3.18	267	6	2	11	260	0	1.12%
spect_001	3.18	267	6	3	6	260	0	1.12%
spect_001	3.18	267	24	9	19	264	0	4.5%
spect_001	3.18	267	12	9	6	218	1	2.25%
spect_01	3.18	267	8	9	1	218	1	1.5%
spect_01	3.18	267	8	9	19	218	1	1.5%
spect_1	1.55	267	26	1	38	76	1	4.87%
spect_1	1.55	267	10	2	4	116	0	1.87%
spect_2	1.29	267	16	5	6	218	1	3%
spect_2	1.29	267	16	2	19	218	1	3%
spect_2	1.29	267	16	1	6	188	1	3%
spect_5	1.022	267	8	2	6	218	1	1.5%
spect_5	1.022	267	14	1	34	31	1	2.62%
spect_5	1.022	267	20	1	11	188	1	3.75%

Konieczny jest opis informacji ujętych w tabelach 1-6. Kolumna „set” odpowiada oczywiście nazwie analizowanego zbioru danych, dalej kolumna „rlavg” oznacza średnią długość reguły (jej części warunkowej), „NR” – liczbę reguł w bazie wiedzy, „Nn” – liczbę węzłów w utworzonej strukturze, „Ndata” – liczbę danych wprowadzonych jako fakty do systemu, „Nrr” – liczbę reguł relewantnych, „NRrs” – liczbę reguł faktycznie analizowaną, „Precision” – wartości miary dokładności wyszukiwania reguł, a „%KB” – procent bazy wiedzy faktycznie przeglądany.

Tabela 2

Wyniki dla miary odległości Euklidesowej – zbiór „spect_all”

set	rlavg	NR	Nn	Ndata	Nrr	NRrs	Precision	%KB
spect_0	3.18	267	42	1	38	9	1	7.87%
spect_0	3.18	267	14	9	6	151	0	2.62%
spect_0	3.18	267	20	9	4	46	0	3.75%
spect_001	3.18	267	38	1	38	9	0	7.12%
spect_001	3.18	267	40	2	11	27	0	7.5%
spect_001	3.18	267	10	3	6	184	0	1.87%
spect_001	3.18	267	56	9	19	172	0	10.5%
spect_001	3.18	267	12	9	6	151	0	2.25%
spect_01	3.18	267	10	9	1	151	0	1.87%
spect_01	3.18	267	14	9	19	188	0	2.62%
spect_1	1.55	267	42	1	38	9	1	7.87%
spect_1	1.55	267	10	2	4	116	0	1.87%
spect_2	1.29	267	38	5	6	109	1	7.12%
spect_2	1.29	267	40	2	19	172	1	7.5%
spect_2	1.29	267	22	1	6	1	1	4.12%
spect_5	1.022	267	40	2	6	109	1	7.5%
spect_5	1.022	267	34	1	34	2	1	6.37%
spect_5	1.022	267	36	1	11	1	1	6.75%

Tabela 3

Wyniki dla miary kosinusowej- zbiór „spect_all”

set	rlavg	NR	Nn	Ndata	Nrr	NRrs	Precision	%KB
spect_0	3.18	267	14	1	38	76	1	2.62%
spect_0	3.18	267	16	9	6	169	1	3%
spect_0	3.18	267	18	9	4	151	0	3.37%
spect_001	3.18	267	14	1	38	76	1	2.62%
spect_001	3.18	267	12	2	11	162	0	2.25%
spect_001	3.18	267	24	3	6	209	0	4.5%
spect_001	3.18	267	14	9	19	76	0	2.62%
spect_001	3.18	267	28	9	6	218	1	5.25%
spect_01	3.18	267	14	9	1	53	0	2.62%
spect_01	3.18	267	14	9	19	76	0	2.62%
spect_1	1.55	267	16	1	38	76	1	3%
spect_1	1.55	267	28	2	4	232	1	5.25%
spect_2	1.29	267	22	5	6	218	1	4.12%
spect_2	1.29	267	30	2	19	235	1	5.62%
spect_2	1.29	267	16	1	6	188	1	3%
spect_5	1.022	267	14	2	6	218	1	2.62%
spect_5	1.022	267	18	1	34	31	1	3.37%
spect_5	1.022	267	14	1	11	188	1	2.62%

W tabelach 1-3 przedstawiono wyniki analizy 3 różnych metryk podobieństwa bądź odległości dla zbioru 267 reguł („spect_all.kb”), zaś w tabelach 4-6 wyniki dla zbioru 101 reguł („zoo.kb”). Analiza zbioru pierwszego pokazuje wyraźnie, iż miary odległości Euklidesowej oraz podobieństwa kosinusowego nie gwarantują poprawności procesu wnioskowania. Co więcej, możemy zauważyć, że najmniej błędów występuje przy stosowaniu miary Gowera; przeważnie udawało się uaktywnić regułę relewantną. Drugi zbiór był bardziej

stabilny. W zasadzie we wszystkich przypadkach uaktywniono reguły zapewniające poprawność wnioskowania.

Tabela 4

Wyniki dla miary Gowera – zbiór „zoo”

#set	rlavg	NR	Nn	Ndata	Nrr	NRrs	Precision	%KB
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%
Zoo_01	1.48	101	18	1	41	46	1	9%
Zoo_01	1.48	101	12	2	13	87	1	6%
Zoo_01	1.48	101	12	4	1	26	1	6%
Zoo_1	1.059	101	12	1	8	30	1	6%
Zoo_1	1.059	101	8	2	1	52	1	4%
Zoo_2	1.059	101	12	1	8	30	1	6%
Zoo_2	1.059	101	10	2	2	89	1	5%
Zoo_5	1	101	14	1	41	45	1	7%

Tabela 5

Wyniki dla miary odległości Euklidesowej – zbiór „zoo”

#set	rlavg	NR	Nn	Ndata	Nrr	NRrs	Precision	%KB
Zoo_0	1.48	101	12	1	41	45	1	6%
Zoo_0	1.48	101	14	2	2	98	1	7%
Zoo_0	1.48	101	14	3	1	90	1	7%
Zoo_001	1.48	101	12	1	41	45	1	6%
Zoo_001	1.48	101	26	2	4	40	1	13%
Zoo_001	1.48	101	14	2	4	39	1	7%
Zoo_001	1.48	101	16	4	1	52	1	8%
Zoo_01	1.48	101	12	1	41	1	1	6%
Zoo_01	1.48	101	8	2	13	3	1	4%
Zoo_01	1.48	101	12	4	1	26	1	6%
Zoo_1	1.059	101	8	1	8	24	1	4%
Zoo_1	1.059	101	8	2	1	52	1	4%
Zoo_2	1.059	101	10	1	8	24	1	5%
Zoo_2	1.059	101	12	2	2	89	1	6%
Zoo_5	1	101	16	1	41	0	1	8%

Skuteczność poszczególnych metryk: odległości Euklidesowej oraz podobieństwa Gowera i kosinusowego wyrażona jest w tabelach 1-3 w postaci parametru Precision. Wartość „1” oznacza, że szukanie reguły do uaktywnienia zakończyło się sukcesem, zaś wartość „0” odpowiada nieznalezieniu reguły do uaktywnienia. Można zauważyć, że miarą najskuteczniejszą w szukaniu reguł (relewantnych) do uaktywnienia jest miara Gowera. Jednak nie jest to reguła. Okazuje się, że wiele zależy od typu danych podlegających analizie. Zbiór „spect_all” był dość specyficzny. Dużo lepsze wyniki dostarczył zbiór „zoo”. Praktycznie w każdym przypadku udawało się znaleźć reguły relewantne w procesach wnioskowania – wyniki przedstawiają tabele 4-6.

Tabela 6

Wyniki dla miary kosinusowej – zbiór „zoo”

#set	rlavg	NR	Nn	Ndata	Nrr	NRrs	Precision	%KB
Zoo_0	1.48	101	14	2	2	98	1	7%
Zoo_0	1.48	101	12	1	41	45	1	6%
Zoo_0	1.48	101	26	3	1	90	1	13%
Zoo_001	1.48	101	14	1	41	45	1	7%
Zoo_001	1.48	101	16	4	1	52	1	8%
Zoo_01	1.48	101	14	1	41	46	1	7%
Zoo_01	1.48	101	12	2	13	87	1	6%
Zoo_01	1.48	101	18	4	1	26	1	9%
Zoo_1	1.059	101	22	1	8	51	1	11%
Zoo_1	1.059	101	12	2	1	52	1	6%
Zoo_2	1.059	101	20	1	8	51	1	10%
Zoo_2	1.059	101	12	2	2	89	1	6%
Zoo_5	1	101	16	1	41	45	1	8%

5. Podsumowanie

W pracy pokazano wpływ wybranych metryk podobieństwa bądź odległości na efektywność grupowania reguł w złożonych bazach wiedzy. Przez efektywność w tym aspekcie będziemy rozumieć efektywność procesów wnioskowania w bazach wiedzy, w których reguły podlegały grupowaniu. Kryterium podobieństwa części warunkowych reguł stanowić miała wybrana metoda podobieństwa bądź odległości. Wnioskowanie wtedy kończy się sukcesem, gdy udaje się uaktywnić regułę relewantną względem podanych faktów bądź regułę, której konkluzja pokrywa się z celem wnioskowania. Analiza skupień, a konkretnie algorytmy hierarchiczne użyte tutaj do budowy skupień reguł w efekcie tworzą drzewa binarne nazywane często dendrogramami. Drzewa te można efektywnie przeszukiwać metodami o złożoności logarytmicznej, co przy klasycznym przeszukiwaniu liniowym daje oczywistą korzyść. Jednak przeszukiwanie połówkowe tutaj stosowane wymaga dobrej jakościowo struktury utworzonego drzewa, a ta zależy przede wszystkich od dobrze dobranych parametrów algorytmów grupowania. Okazało się, że metryką najbardziej odporną na problemy ze znalezieniem reguł do uaktywnienia jest metryka podobieństwa Gowera, m.in. dlatego, że jest to miara dobrze radząca sobie z danymi różnego typu: ilościowymi i jakościowymi. Miara kosinusowa jest zbyt zależna od długości wektora opisującego analizowany obiekt i tym lepiej sobie radzi, im długość wektora jest większa. Jako że reguły w bazach analizowanych były dość krótkie, toteż wyniki osiągane przez metodę podobieństwa kosinusowego użytego do budowy skupień reguł nie były zadowalające. Najwięcej błędów w poszukiwaniu reguł do uaktywnienia w procesach wnioskowania dostarczyła miara nakładania (ang. *overlap*). Zapewne powodem tego był fakt, że w oryginale metryka ta stosowana była dla danych binarnych nie zaś dla danych o wielu

wartościach. Skuteczność miary Gowera wynika niewątpliwie z tego, że miara ta dostosowuje się do typu danych i pozwala dobrze oszacować wartość podobieństwa między analizowanymi obiektami zarówno w sytuacji, gdy analizujemy cechy ilościowe, jak i jakościowe.

Na potrzeby niniejszej pracy wykonano szereg eksperymentów, w których analizowano efektywność wnioskowania. Przetestowano kilka zbiorów z repozytorium *UCI Machine Learning*[12], różniących się liczbą atrybutów, ich typem, liczbą obiektów. Dodajmy przy tym, iż oryginalne obiekty podlegały algorytmom indukcji reguł i w efekcie w eksperymentach obiektami były reguły dla obiektów analizowanych zbiorów. Z uwagi na ograniczenia dotyczące rozmiarów niniejszej pracy przedstawiono wyniki uzyskane tylko dla niewielkiej części zanalizowanych zbiorów. Jako że reguły w bazach wiedzy są specyficznym typem danych do analizy, miary podobieństwa bądź odległości powinny być dobierane ściśle do typu danych (najlepiej by były to miary wielocechowe). Z tego względu kolejne wykonywane eksperymenty będą dotyczyły bardziej złożonych baz wiedzy (np. o większej liczbie nie tylko wartości atrybutów ale i samych atrybutów budujących części warunkowe reguł), a także innych metryk stosowanych jako kryterium w procesie grupowania oraz takich, które pozwalają oceniać efektywność grupowania.

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., Cwik 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. Bren M., Batagejl V.: The Metrix Index. University of Ljubljana, Preprint series, Vol. 35, 561, Ljubljana, Slovenia 1997.
8. Nowak A., Wakulicz-Deja A., Simiński, R.: Knowledge representation for composited knowledge bases. Control and Cybernetics, Intelligent Information Systems 2008, Zakopane 2008, ISBN 978-83-60434-44-4, s. 405÷414.

9. Nowak A., Wakulicz-Deja A.: The inference processes on composited knowledge bases. Control and Cybernetics, Intelligent Information Systems 2009, 2008, ISBN 978-83-60434-44-4, s. 415÷422.
10. 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, ISBN 978-83-60434-59-8, s. 579÷588.
11. Nowak A., Zielosko B.: Clustering of partial decision rules. Advanced In Intelligent and Soft Computing, Man – Machine Interactions, Springer-Verlag 2009, s. 183÷190.
12. UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/

Recenzenci: Dr hab. inż. Adam Pelikant, prof. Pol. Łódzkiej,
Dr inż. Marek Sikora

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

The article presents the results of experiments analyzing the influence of the measure that was used in clustering algorithm on the efficiency of created structure. Each choosed measure gives different structure of rules' clusters. Then the inference process is also different. Authors analyzed two kind of clustering methods: nonhierarchical and hierarchical. Finally agglomerative hierarchical clustering algorithm was choosed. As a result It builts binary tree structure, called dendrogram. In such tree, leaves are rules from knowledge base, and at the higher levels tere are rules' clusters. Authors propose effective methods of searching such structure, called „the best node in the tree" and „the node with minimal treshold value". Such methods In the logarithmic time efficiency (instead of linear) find the relevant node and do the inference process on such rule. The experiment are based on two different set of rules. The first one, called „spect_all" consists of 267 rules, and the second one, called „zoo" consists of 101 rules. Checking the results (when we changing the measures used during the clustering process) we can see that the best results (the highest values of precision parameter) we achieve when we use Gower's similarity measure, which is an universal measure for multidimensional and different type data (like rules in knowledge base).

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 .

Alicja WAKULICZ - DEJA Uniwersytet Śląski, Instytut Informatyki, Wydział Informatyki i Nauki o Materiałach, ul. Będzińska 39, 41-200 Sosnowiec, Polska, alicja.wakulicz-deja@us.edu.pl .