STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Beata ZIELOSKO
Uniwersytet Śląski, Instytut Informatyki

ALGORYTM ZACHŁANNY DLA KONSTRUOWANIA CZĘŚCIOWYCH REGUŁ ASOCJACYJNYCH

Streszczenie. W artykule przedstawiono sposób konstruowania częściowych reguł asocjacyjnych z wykorzystaniem algorytmu zachłannego. Podejście to jest odmienne od znanego algorytmu A priori i jego modyfikacji, wykorzystujących zbiory częste. Przedstawione wyniki badań oraz rezultaty z przeprowadzonych eksperymentów pokazują, że algorytm zachłanny pozwala konstruować stosunkowo małą liczbę krótkich, częściowych reguł asocjacyjnych o dobrej jakości, które pokrywają wszystkie obiekty danego systemu informacyjnego.

Słowa kluczowe: reguły asocjacyjne, częściowe reguły asocjacyjne, algorytm zachłanny

GREEDY ALGORITHM FOR PARTIAL ASSOCIATION RULE CONSTRUCTION

Summary. The paper presents greedy algorithm for partial association rule construction. This approach is different from the known algorithm Apriori and its modifications based on frequent itemsets. Theoretical and experimental results show, that the greedy algorithm constructs relatively small number of short partial association rules which have good accuracy and cover all objects from given information system.

Keywords: association rules, partial association rules, greedy algorithm

1. Wprowadzenie

W artykule przedstawiono algorytm zachłanny dla konstruowania częściowych reguł asocjacyjnych. Podejście to jest odmienne od znanego, opartego na algorytmie A priori i jego licznych modyfikacji, wykorzystujących zbiory częste [1].

Częściowe reguły asocjacyjne są szczególnym przypadkiem reguł asocjacyjnych. Opierając się na wynikach badań dotyczących częściowego pokrycia zbioru oraz wynikach badań dotyczących konstruowania częściowych reguł decyzyjnych [4] i wykorzystując zalety regułowej reprezentacji wiedzy, zaproponowano podejście wykorzystujące algorytm zachłanny, zaprezentowane w pracy [5].

Problem minimalizacji długości częściowych reguł asocjacyjnych jest problemem NP-trudnym, dlatego został wykorzystany algorytm aproksymacyjny o złożoności wielomianowej (algorytm zachłanny) dla konstruowania częściowych reguł asocjacyjnych.

Reguły asocjacyjne to jedna z metod eksploracji danych. Reguły asocjacyjne stanowią model reprezentacji wiedzy o związkach i korelacjach występujących między danymi. W literaturze można znaleźć wiele przykładów potwierdzających, iż często zamiast dokładnych reguł z wieloma atrybutami, stosowane są częściowe (przybliżone) reguły, zawierające mniejszą liczbę atrybutów i pozwalające uzyskać lepsze wyniki, np. w procesie klasyfikacji [4, 9, 11]. Dokładne reguły mogą być zbyt mocno dopasowane do istniejących przykładów, poza tym opierając się na zasadzie minimalnego opisu (ang. *minimal length principle*) należy dążyć do optymalizacji opisu pojęć [8]. Kryteria optymalizacji jakości opisu wypracowane w różnych dziedzinach nie są jednoznaczne, a wybór właściwego uzależniony jest od specyfiki konkretnych zbiorów danych. Istotne jest utrzymanie właściwej równowagi pomiędzy ogólnością opisu a jego poprawnością.

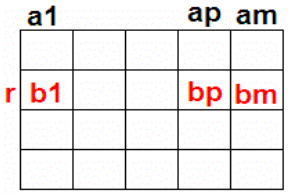
Artykuł składa się z 5 rozdziałów. W rozdziale 2 zostały przedstawione podstawowe pojęcia dotyczące częściowych reguł asocjacyjnych. Rozdział 3 prezentuje algorytm zachłanny oraz twierdzenia dotyczące oszacowania dokładności wyników uzyskiwanych za pomocą algorytmu zachłannego. Rozdział 4 zawiera wyniki eksperymentów przeprowadzonych na danych umieszczonych w UCI machine learning repository [2]. Rozdział 5 stanowi krótkie podsumowanie.

2. Podstawowe pojęcia

W rozdziale tym zostaną przedstawione podstawowe pojęcia dotyczące częściowych reguł asocjacyjnych.

System informacyjny $I=(U,A)$ jest parą, gdzie $U=\{u_1,\ldots,u_n\}$ oznacza skończony zbiór obiektów, $A=\{a_1,\ldots a_m\}$ oznacza skończony zbiór atrybutów [7]. System informacyjny przedstawiany jest w formie tabeli zawierającej n wierszy (odpowiadających obiektom $u_1,\ldots,u_n$) oraz m kolumn (odpowiadających atrybutom $a_1,\ldots,a_m$). Tabela ta wypełniona jest przez wartości atrybutów ze zbioru A, odpowiadające obiektom ze zbioru U.

Niech $r=(b_1,...b_m)$ będzie wierszem tablicy I opisanym przez wartości atrybutów $b_1,...,b_m$, a_p jest atrybutem ze zbioru A. Przez $U(I,r,a_p)$ jest oznaczany zbiór wierszy systemu informacyjnego I, które są różne od wiersza r na przecięciu z kolumną a_p i są różne na przecięciu z przynajmniej jedną kolumną a_j taką, że $j \in \{1,...,m\} \setminus \{p\}$. Powiemy, że atrybut a_i separuje (oddziela) wiersz $r' \in U(I,r,a_p)$ od wiersza r, jeśli wiersze te posiadają różne wartości na przecięciu z kolumną a_i.



Rys.1 System informacyjny I

.

Fig.1. The information system I

Niech α będzie liczbą rzeczywistą taką, że $0 \leq \alpha < 1$. Reguła

$$(a_{i1} = b_{i1}) \wedge \ldots \wedge (a_{it} = b_{it}) \rightarrow a_p = b_p$$

jest nazywana *α-regułą asocjacyjną* (*częściową regułą asocjacyjną*) dla (I,r,a_p), jeśli atrybuty występujące w części warunkowej reguły (atrybuty $a_1,...,a_t$) oddzielają od wiersza r przynajmniej $\lceil (1-\alpha)|U(I,r,a_p)| \rceil$ wierszy ze zbioru $U(I,r,a_p)$.

Na przykład *0.1-reguła asocjacyjna* oznacza, że należy oddzielić od wiersza r przynajmniej 90% wierszy dotychczas nie oddzielonych ze zbioru $U(I,r,a_p)$.

Częściowe reguły asocjacyjne są szczególnym rodzajem reguł asocjacyjnych. Przedstawione w pracy wyniki badań dotyczą m.in. długości częściowych reguł asocjacyjnych. Ponieważ częściowe reguły asocjacyjne są konstruowane dla każdego wiersza systemu informacyjnego I, liczbę tych reguł można ograniczyć wykorzystując standardowe parametry dotyczące reguł asocjacyjnych, np. wsparcie (ang. *support*). Zatem dla częściowej reguły asocjacyjnej w postaci implikacji $X \rightarrow Y$ stosowane są następujące miary:

- długość - liczba atrybutów występujących w części warunkowej reguły,
- wsparcie (*supp*) - liczba wierszy systemu informacyjnego I, które spełniają warunek $X \wedge Y$, podzielona przez liczbę wszystkich wierszy systemu informacyjnego I. Zatem jest to prawdopodobieństwo zajścia zdarzenia $X \wedge Y$.

3. Algorytm zachłanny

W rozdziale tym zostanie przedstawiony algorytm zachłanny z parametrem α, stosowany do konstruowania częściowej reguły asocjacyjnej dla trójki (I,r,a_p) oraz twierdzenia z pracy

[5], dotyczące oszacowania dokładności wyników uzyskiwanych przy zastosowaniu tego algorytmu (podrozdział 3.1).

Poniżej został przedstawiony pseudokod algorytmu zachłannego, który konstruuje częściowe reguły asocjacyjne dla każdego wiersza r i każdego atrybutu a_p systemu informacyjnego I.

```
Dane wejściowe: system informacyjny I zawierający atrybuty a1,…,am,
wiersze r1,…,rn oraz liczba rzeczywista α taka, że 0 ≤ α < 1.
Q - zbiór atrybutów na podstawie których tworzona jest częściowa reguła
asocjacyjna

Dane wyjściowe: α-reguła asocjacyjna dla trójki (I,r,ap).

BEGIN
   FOR j=1 to n   //dla każdego wiersza
      FOR i=1 to m//dla każdego atrybutu
      r=rj;       //r będzie wierszem dla którego zostanie wygenerowana
      reguła
      ap=ai;      //ap będzie stanowić konkluzję reguły
      Q ← ∅;
      WHILE atrybuty ze zbioru Q oddzielają od wiersza r mniej niż
      ⌈(1-α)|U(I,r,ap)|⌉ wierszy ze zbioru U(I,r,ap)
      DO
         wybierz taki atrybut ai∈{a1,…,am}\{ap} o minimalnym indeksie i,
         który oddziela największą liczbę wierszy ze zbioru U(I,r,ap)
         dotychczas nie oddzielonych przez atrybuty ze zbioru Q;
               Q ← Q ∪ {ai};
         Atrybuty zawarte w zbiorze Q tworzą cześć przesłankową α-reguły
         asocjacyjnej.
      END
   END
END
```

3.1. Dokładność algorytmu zachłannego

W rozdziale tym zostaną przedstawione twierdzenia z pracy [5] dotyczące oszacowania dokładności wyników uzyskiwanych za pomocą algorytmu zachłannego, na podstawie wyników badań P. Slavika [10], dotyczących częściowego pokrycia zbioru oraz wyników badań zawartych w [4]. Wyniki te pozwalają lepiej zrozumieć wybór algorytmu zachłannego, spośród aproksymacyjnych algorytmów o złożoności wielomianowej, dla minimalizacji długości częściowych reguł asocjacyjnych. Niech I będzie systemem informacyjnym zawierającym m kolumn oznaczonych przez atrybuty $a_1,\ldots, a_m$, $r = (b_1,\ldots b_m)$ będzie wierszem systemu I i atrybut $a_p \in \{a_1,\ldots,a_m\}$.

Przez $L_{min}(\alpha) = L_{min}(\alpha,I,r,a_p)$ jest oznaczana minimalna długość częściowej reguły asocjacyjnej dla trójki (I,r,a_p).

Przez $L_{greedy}(\alpha) = L_{greedy}(\alpha,I,r,a_p)$ jest oznaczana długość częściowej reguły asocjacyjnej konstruowanej przez algorytm zachłanny dla trójki (I,r,a_p).

Twierdzenie 1. [5]

Niech $0 \leq \alpha < 1$ i $\lceil (1-\alpha)|U(I,r,a_p)| \rceil \geq 2$. Wówczas

$L_{greedy}(\alpha) < L_{min}(\alpha)(\ln \lceil (1-\alpha)|U(I,r,a_p)| \rceil - \ln \ln \lceil (1-\alpha)|U(I,r,a_p)| \rceil + 0.78)$.

Twierdzenie 2. [5]

Niech $0 \leq \alpha < 1$. Wówczas dla każdej liczby naturalnej $t \geq 2$ istnieje problem częściowej reguły asocjacyjnej (I, r, a_p) taki, że $\lceil (1-\alpha)|U(I,r,a_p)| \rceil = t$ i

$L_{greedy}(\alpha) > L_{min}(\alpha)(\ln \lceil (1-\alpha)|U(I,r,a_p)| \rceil - \ln \ln \lceil (1-\alpha)|U(I,r,a_p)| \rceil - 0.31)$.

Twierdzenie 3. [5]

Niech $0 \leq \alpha < 1$ i $U(I,r,a_p) \neq \emptyset$. Wówczas

$L_{greedy}(\alpha) \leq L_{min}(\alpha)(1+\ln (\max_{j \in \{1,\ldots,m\} \setminus \{p\}} |U(I,r,a_p,a_j)|))$.

Poniżej zostaną przedstawione twierdzenia dotyczące aproksymacyjnych algorytmów o złożoności wielomianowej. Twierdzenia te bazują na wynikach badań D. Ślęzaka [11], U.Feigego [3] oraz wynikach badań przedstawionych w [6].

Twierdzenie 4. [5]

Niech $0 \leq \alpha < 1$. Wówczas problem konstruowania α-reguły asocjacyjnej o minimalnej długości jest NP-trudny.

Twierdzenie 5. [5]

Niech $\alpha \in R$ i $0 \leq \alpha < 1$. Jeśli NP $\not\subseteq$ DTIME $(n^{O(\log \log n)})$, wówczas dla każdej liczby ε, $0 < \varepsilon < 1$ nie istnieje algorytm o złożoności wielomianowej, który dla danego problemu reguły asocjacyjnej (I,r,a_p) z $U(I,r,a_p) \neq \emptyset$, konstruuje α-regułę asocjacyjną dla (I,r,a_p), której długość wynosi najwyżej $(1-\varepsilon) L_{min}(\alpha,I,r,a_p) \ln| U(I,r,a_p)|$.

Z twierdzenia 3 wynika, że $L_{greedy}(\alpha) \leq L_{min}(\alpha)(1 + \ln |U(I,r,a_p)|)$. Z tej nierówności i z twierdzenia 5 wynika, że przyjmując założenie NP $\not\subseteq$ DTIME $(n^{O(loglogn)})$, algorytm zachłanny pozwala uzyskać wyniki, bliskie wynikom uzyskiwanym przez najlepsze algorytmy aproksymacyjne o złożoności wielomianowej dla minimalizacji długości częściowej reguły asocjacyjnej.

4. Wyniki eksperymentów

Rozdział ten przedstawia wyniki eksperymentów, które miały na celu zbadanie zależności pomiędzy:

- długością częściowych reguł asocjacyjnych a wartością parametru α;

W tym celu dla analizowanych zbiorów danych została wyznaczona minimalna, średnia i maksymalna długość częściowych reguł asocjacyjnych;

- liczbą reguł asocjacyjnych a wartością parametru minimalnego wsparcia;
- liczbą reguł wyznaczonych dla określonej wartości parametru α, z uwzględnieniem zadanego współczynnika minimalnego wsparcia.

Eksperymenty zostały przeprowadzone na zbiorach danych umieszczonych w repozytorium UCI repository of machine learning [2]:

- car (7 atrybutów, 1728 wierszy);
- flags (27 atrybutów, 194 wiersze); ze zbioru tego zostały usunięte atrybuty: „area", „population" i „name of the country", ich pozostawienie wymagałoby przeprowadzenia dyskretyzacji;
- kr-vs-kp (37 atrybutów, 3196 wierszy);
- lenses (5 atrybutów, 24 wiersze);
- lymphography (19 atrybutów, 148 wierszy);
- spect (23 atrybuty, 267 wierszy);
- zoo (17 atrybutów, 101 wierszy). Ze zbioru został usunięty atrybut „animal name".

Tabela 1 przedstawia minimalne (min), średnie (avg) i maksymalne (max) długości częściowych reguł asocjacyjnych, w zależności od wartości parametru α, α={0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5}

Tabela 1

Długości reguł asocjacyjnych w zależności od wartości α

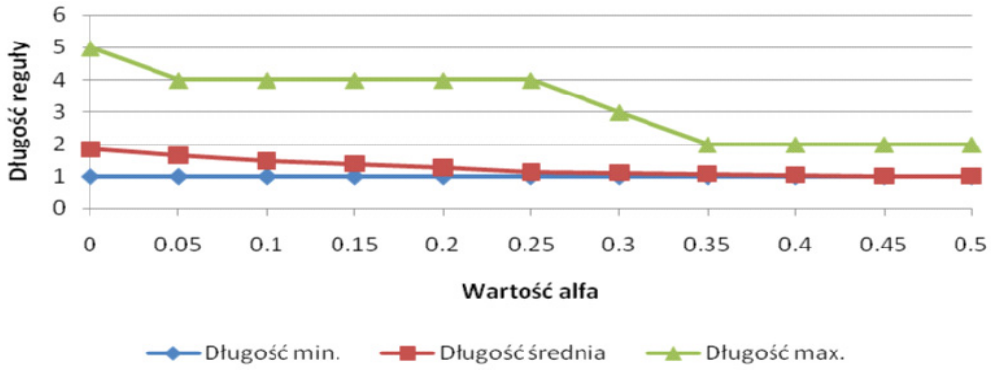
Nazwa zbioru	**Długość**	**Parametr α $\lceil \log_2(1/\alpha) \rceil$**										
		0.0	**0.05 5**	**0.1 4**	**0.15 3**	**0.2 3**	**0.25 2**	**0.3 2**	**0.35 2**	**0.4 2**	**0.45 2**	**0.5 1**
Car	min	1	1	1	1	1	1	1	1	1	1	1
	avg	2.13	2.06	1.95	1.94	1.93	1.02	1	1	1	1	1
	max	3	3	2	2	2	2	1	1	1	1	1
Flags	min	1	1	1	1	1	1	1	1	1	1	1
	avg	1.39	1.16	1.05	1.01	1	1	1	1	1	1	1
	max	3	2	2	2	2	1	1	1	1	1	1
Lenses	min	1	1	1	1	1	1	1	1	1	1	1
Lenses	avg	2.83	2.77	2.65	2.52	2.04	1.73	1.71	1.26	1.24	1.24	1
	max	4	4	4	3	3	2	2	2	2	2	1
Lympho-graphy	min	1	1	1	1	1	1	1	1	1	1	1
	avg	2.5	1.85	1.63	1.47	1.22	1.08	1.01	1	1	1	1
	max	6	3	3	2	2	2	2	1	1	1	1
Zoo	min	1	1	1	1	1	1	1	1	1	1	1
	avg	1.84	1.65	1.48	1.38	1.27	1.13	1.1	1.01	1	1	1
	max	5	4	4	4	4	4	3	2	2	2	2

W pracy [5] została sformułowana nieformalnie tzw. 0.5-hipoteza: dla większej części systemów informacyjnych I, dla każdego wiersza r i każdego atrybutu a_p, podczas konstruowania częściowej reguły asocjacyjnej dla trójki (I,r,a_p) algorytm zachłanny w każdym kroku wybiera atrybut, który oddziela od wiersza r przynajmniej połowę wierszy dotychczas nie oddzielonych ze zbioru $U(I, r, a_p)$.

Można pokazać, że jeśli 0.5-hipoteza jest prawdziwa dla danego systemu informacyjnego, wówczas długość reguły konstruowanej przez algorytm zachłanny wynosi najwyżej $\lceil \log_2(1/\alpha) \rceil$. Wyniki eksperymentów przedstawione w tabeli 1 pokazują, że rozważana właściwość nie zachodzi tylko dla zbioru zoo. Wyniki te stanowią pośrednie potwierdzenie 0.5-hipotezy.

Na podstawie wyników w tabeli 1 można zauważyć, że wraz ze wzrostem wartości parametru α zmniejsza się średnia i maksymalna długość częściowych reguł asocjacyjnych. Na przykład dla α=0.5 wartości te wynoszą 1 dla przedstawionych pięciu zbiorów danych. Minimalne długości częściowych reguł asocjacyjnych także wynoszą 1. Średnia długość częściowych reguł asocjacyjnych od wartości α=0.25, z wyjątkiem zbioru lenses, również wynosi 1. Wyniki eksperymentów z tabeli 1 zostały także zaprezentowane w formie graficznej (rys. 2 i rys. 3).

Rysunek 2 przedstawia minimalną, średnią i maksymalną długość częściowych reguł asocjacyjnych dla pliku zoo.



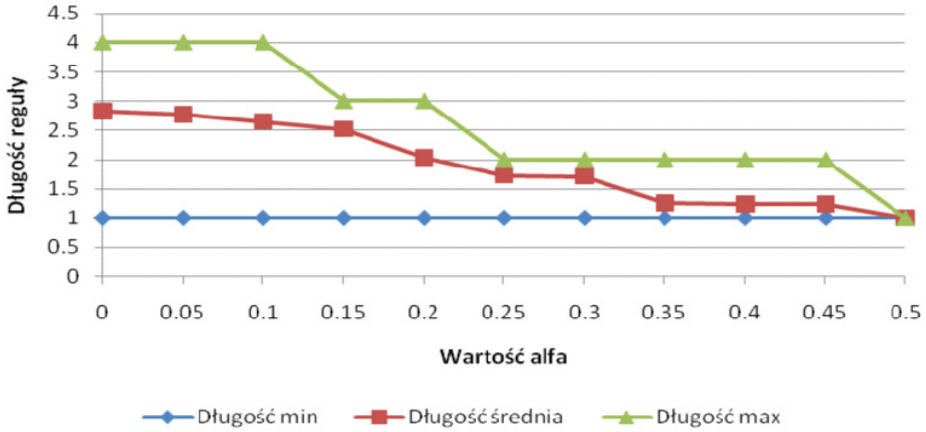
Rys. 2. Minimalna, średnia i maksymalna długość częściowych reguł asocjacyjnych dla zbioru danych zoo

Fig. 2. The minimal, average and maximal length of partial association rules for the dataset zoo

Rysunek 3 przedstawia minimalną, średnią i maksymalną długość częściowych reguł asocjacyjnych dla pliku lenses.

Jednym z problemów związanych z algorytmem A priori jest duża liczba konstruowanych reguł. Podejście zaproponowane w artykule nie posiada tej wady, tzn. liczba konstruowanych reguł wynosi najwyżej mn, gdzie m jest liczbą atrybutów, a n liczbą obiektów dla danego systemu informacyjnego. Poniżej zostanie pokazane, że liczba różnych reguł konstruowanych

przez algorytm zachłanny może być znacznie mniejsza niż mn. Liczbę tych reguł można ograniczyć, jeśli zostaną uwzględnione dodatkowe parametry, np. wsparcie (support).



Rys. 3. Minimalna, średnia i maksymalna długość częściowych reguł asocjacyjnych dla zbioru danych lenses
Fig. 3. The minimal, average and maximal length of partial association rules for the dataset lenses

Tabela 2 przedstawia liczbę wygenerowanych częściowych reguł asocjacyjnych, w zależności od minimalnej wartości parametru wsparcia. Dla zbiorów danych kr-vs-kp, lymphography, zoo i spect, stosując algorytm zachłanny zostały skonstruowane częściowe reguły asocjacyjne dla każdego wiersza i każdego atrybutu ze zbioru $\{a_1,\ldots,a_m\}$ oznaczonego jako a_p oraz wartości $\alpha=\{0.0\}$ i minimalnej wartości wsparcia supp={0, 10%, 20%, 30%, 40%, 50%}. W przypadku powtarzających się reguł uwzględniana była tylko jedna.

Tabela 2
Liczba reguł w zależności od wartości minimalnego wsparcia

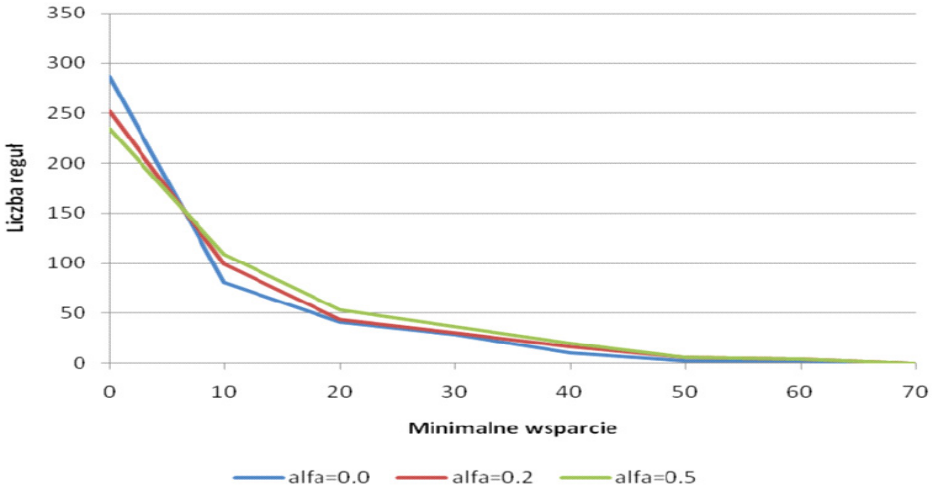
Nazwa zbioru	**Minimalne wsparcie (supp)**					
	0	**10%**	**20%**	**30%**	**40%**	**50%**
Kr-vs-kp	15383	460	165	94	51	33
Lymphography	1688	75	23	12	8	3
Spect	2718	188	52	25	5	2
Zoo	287	81	41	29	11	3

Można dokonać porównania teoretycznej górnej granicy mn liczby reguł konstruowanych przez algorytm zachłanny i rzeczywistej liczby różnych reguł konstruowanych przez przedstawiony algorytm.

- Dla zbioru kr-vs-kp, mn = 118252, a rzeczywista liczba skonstruowanych reguł wynosi 15383.
- Dla zbioru lymphography, mn = 2812, a rzeczywista liczba skonstruowanych reguł wynosi 1688.
- Dla zbioru spect, mn = 6141, a rzeczywista liczba skonstruowanych reguł wynosi 2718.
- Dla zbioru zoo, mn = 1717, a rzeczywista liczba skonstruowanych reguł wynosi 287.

Zatem można zauważyć, że rzeczywista liczba różnych reguł skonstruowanych przez algorytm zachłanny często jest znacznie mniejsza niż teoretyczna górna granica mn.

Rysunek 4 przedstawia liczbę skonstruowanych przez algorytm zachłanny częściowych reguł asocjacyjnych, w zależności od wartości minimalnego wsparcia, dla pliku zoo.



Rys. 4. Liczba reguł w zależności od minimalnego wsparcia dla zbioru danych zoo
Fig. 4. The number of rules dependent on minimal support for dataset zoo

Na podstawie wyników z tabeli 2 oraz rys. 4 można zauważyć, że wraz ze wzrostem minimalnej wartości wsparcia liczba skonstruowanych częściowych reguł asocjacyjnych zmniejsza się. Zatem parametr ten ma znaczenie w sytuacji, kiedy użytkownik chce ograniczyć liczbę konstruowanych częściowych reguł asocjacyjnych. Na przykład, dla zbioru zoo liczba skonstruowanych częściowych reguł asocjacyjnych (z pominięciem reguł powtarzających się), przy wartości minimalnego wsparcia supp = 10%, wynosi 81, natomiast przy wartości supp = 30% wynosi 29.

Wyniki w tabeli 2 dotyczyły sytuacji, kiedy $\alpha = 0$. Poniżej zostanie rozważony przypadek, kiedy parametr α wzrasta.

Tabela 3 przedstawia liczbę wygenerowanych częściowych reguł asocjacyjnych, w zależności od wartości parametru α i minimalnej wartości parametru wsparcia.

Dla zbiorów danych lymphography, spect i zoo zostały skonstruowane częściowe reguły asocjacyjne dla każdego wiersza i każdego atrybutu ze zbioru $\{a_1,\ldots,a_m\}$ oznaczonego jako a_p oraz wartości $\alpha=\{0.0, 0.05, 0.1\}$ i minimalnej wartości wsparcia supp={10%, 20%, 30%, 40%, 50%}. W przypadku powtarzających się reguł uwzględniana była tylko jedna.

Na podstawie wyników z tabeli 3 można zauważyć zależność pomiędzy parametrem minimalnego wsparcia i długością reguł, która związana jest z wartością α. Wraz ze wzrostem wartości α, częściowa reguła asocjacyjna staje się krótsza, co należy rozumieć, że jest bardziej „ogólna”. W związku z tym wraz ze wzrostem wartości α, przy stałym zadanym współczynniku minimalnego wsparcia, liczba reguł (postaci $X \rightarrow Y$) spełniających

prawdopodobieństwo zajścia zdarzenia $X \wedge Y$ zwiększa się. Na przykład, dla zbioru lymphography, przy wartości minimalnego wsparcia supp = 10%, liczba skonstruowanych częściowych reguł asocjacyjnych dla $\alpha = 0.0$ wynosi 75, dla $\alpha = 0.05$ wynosi 106, natomiast dla $\alpha = 0.1$ wynosi 114.

Tabela 3

Liczba reguł w zależności od parametru α i minimalnej wartości wsparcia

Nazwa zbioru	**alfa**	**Minimalne wsparcie (supp)**				
		10%	**20%**	**30%**	**40%**	**50%**
Lympho-graphy	0.0	75	23	12	8	3
	0.05	106	27	13	8	3
	0.1	114	30	14	10	6
Spect	0.0	188	52	25	5	2
	0.05	329	78	31	15	4
	0.1	367	97	50	26	10
Zoo	0.0	81	41	29	11	3
	0.05	94	42	29	15	4
	0.1	97	43	28	15	5

Zatem dla większych wartości α należałoby rozważyć silniejsze ograniczenia związane z parametrem minimalnego wsparcia w celu uzyskania stosunkowo małej liczby reguł.

5. Podsumowanie

W pracy przedstawiono wyniki badań teoretycznych, na podstawie [5] i eksperymenttalnych dotyczących algorytmu zachłannego dla konstruowania częściowych reguł asocjacyjnych. Pokazują one, że przyjmując pewne założenia dotyczące klasy NP, algorytm zachłanny pozwala uzyskać wyniki bliskie wynikom uzyskiwanym przez najlepsze algorytmy aproksymacyjne o złożoności wielomianowej, dla minimalizacji długości częściowej reguły asocjacyjnej. Algorytm ten pozwala konstruować stosunkowo małą liczbę krótkich częściowych reguł asocjacyjnych o dobrej jakości, które pokrywają wszystkie obiekty danego systemu informacyjnego. Stosując parametr minimalnego wsparcia można zmniejszyć liczbę generowanych reguł.

Kolejnym etapem badań będzie przeprowadzenie większej liczby eksperymentów oraz wykorzystanie innych miar dotyczących reguł asocjacyjnych, np. parametru ufności oraz wsparcia i ufności łącznie i zbadanie zależności liczby reguł od wartości tych parametrów, oraz porównanie wyników np. z wynikami dotyczącymi reguł generowanych za pomocą algorytmu A priori. W następnym etapie zostaną przeprowadzone eksperymenty pozwalające

określić jakość klasyfikatorów konstruowanych z wykorzystaniem częściowych reguł asocjacyjnych.

BIBLIOGRAFIA

1. Agrawal R., Srikant, R.: Fast algorithms for mining association rules in large databases, Proc. 20th International Conference on Very Large Data Bases (Bocca J.B., Jarke M., Zaniolo C., Eds.), Morgan Kaufmann, 1994.
2. Asuncion A., Newman D.: UCI machine learning repository, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, School of Information and Computer Sciences, 2007.
3. Feige U.: A threshold of lnn for approximating set cover (preliminary version). Proc. 28th Annual ACM Symposium on the Theory of Computing, ACM Press, New York, 1996.
4. Moshkov M.Ju., Piliszczuk M., Zielosko B.: Partial Covers, Reducts and Decision Rules in Rough Sets: Theory and Applications. Studies in Computational Intelligence, Vol. 145, Springer, Heidelberg, 2009.
5. Moshkov M.Ju., Piliszczuk M., Zielosko B.: Greedy algorithm for construction of partial association rules. Fundamenta Informaticae, 92, Vol. 3, 2009, s. 259÷277.
6. Nguyen H.S., Ślęzak D.: Approximate reducts and association rules – correspondence and complexity results, Proc. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (Zhong N., Skowron A., Ohsuga S.), LNCS (LNAI) 1711, Springer, Heidelberg, 1999.
7. Pawlak Z.: Rough Sets – Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, 1991.
8. Rissanen J.: Modeling by shortest data description. Automatica, 14, 1978, s. 465÷471.
9. Skowron A.: Rough sets in KDD. Proc. 16th IFIP World Computer Congress (Shi Z., Faltings B., Musen M., Eds.), Publishing House of Electronic Industry, 2000.
10. Slavik P.: Approximation algorithms for set cover and related problems, Ph.D. Thesis, University of New York at Buffalo, 1998.
11. Ślęzak D.: Approximate entropy reducts. Fundamenta Informaticae, 53, 2002, s. 365÷390.

Recenzenci: Dr inż. Henryk Josiński,
Dr inż. Marek Sikora

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

In the paper, greedy algorithm for partial association rules construction is considered. This algorithm is different from the Apriori algorithm and its modifications based on frequent itemsets. Partial association rules are a special kind of association rules. Association rules are used in data mining and can be considered as a way of knowledge representation. Exact rules can be overfitted, i.e., dependent essentially on the noise or adjusted too much to the existing examples. Then instead of exact rules with many attributes, it is more appropriate to work with partial (approximate) rules with smaller number of attributes. The problem of minimization of association rule length is NP-hard, so we use approximate algorithm (greedy algorithm), which allows us to construct shorter rules in reasonable time. Theoretical results show, that under some natural assumptions on the class NP, the greedy algorithm is close to the best polynomial approximate algorithms for the minimization of the length of partial association rules.

Experimental results (Table 1) show that greedy algorithm allows us to construct relatively small number of short partial association rules with good accuracy which cover all objects from given information system. These results can be considered also as a confirmation of 0.5-hypothesis from [5]: for the most part of information systems I, for each row r and each attribute a_p, under the construction of partial association rule for (I,r,a_p), during each step the greedy algorithm chooses an attribute which separates from the row r at least one-half of unseparated rows from $U(I,r,a_p)$.

One of the problems connected with the use of Apriori algorithm is the large number of constructed rules. The approach considered in this paper has no this deficiency: the number of constructed rules is at most mn, where m is the number of attributes and n is the number of objects in considered information system. Experimental results show that the number of different rules constructed by greedy algorithm can be essentially less than the theoretical bound mn. Also we can decrease the number of rules essentially if we consider additional parameters, for example, support (see Table 2 and 3).

In future work author would like to do more experiments and use another parameter confidence, or confidence and support together. The next step will be to create classifiers based on partial association rules and evaluate accuracy of classification.

Adres

Beata ZIELOSKO: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39,
41-200 Sosnowiec, Polska, beata.zielosko@us.edu.pl .