STUDIA INFORMATICA 2016
Volume 37 Number 1 (123)

Beata ZIELOSKO
Uniwersytet Śląski, Instytut Informatyki
Marek ROBASZKIEWICZ
EL-PLUS Sp. z o.o.

OPTYMALIZACJA CZĘŚCIOWYCH REGUŁ ASOCJACYJNYCH WZGLĘDEM LICZBY POMYŁEK

Streszczenie. Artykuł przedstawia optymalizację częściowych reguł asocjacyjnych generowanych przez algorytm zachłanny względem liczby pomyłek (błędnych zaklasyfikowań). Zaproponowana optymalizacja ma na celu: (i) uzyskanie reguł o stosunkowo dobrej jakości, które w kolejnych etapach badań zostaną wykorzystane do budowy klasyfikatorów, (ii) zmniejszenie liczby konstruowanych reguł, co ma znaczenie z punktu widzenia reprezentacji wiedzy. Praca przedstawia wyniki eksperymentalne dla zbiorów danych umieszczonych w Repozytorium Uczenia Maszynowego.

Słowa kluczowe: zbiory przybliżone, algorytm zachłanny, częściowe reguły asocjacyjne, liczba pomyłek

OPTIMIZATION OF PARTIAL ASSOCIATION RULES RELATIVE TO NUMBER OF MISCLASSIFICATIONS

Summary. In the paper, an optimization of partial association rules relative to number of misclassifications is presented. The aims of proposed optimization are: (i) construction of rules with small number of misclassifications, what is important from the point of view of construction of classifiers, (ii) decreasing the number of rules, what is important from the point of view of knowledge representation. The paper contains experimental results for data sets from UCI Machine Learning Repository.

Keywords: rough sets, greedy algorithm, partial association rules, misclassifications

1. Wprowadzenie

Reguły asocjacyjne stosowane są do wykrywania interesujących powiązań, wzorców, korelacji między danymi. Są wykorzystywane jako sposób reprezentacji wiedzy oraz do budowy klasyfikatorów [9, 10]. Jednym z najbardziej popularnych zastosowań reguł asocjacyjnych jest tzw. analiza koszykowa (ang. *market basket analysis*) [1].

Istnieje wiele podejść dla konstruowania reguł asocjacyjnych. Jednym z najbardziej popularnych jest podejście oparte na tzw. zbiorach częstych (ang. *frequent itemsets*) zaproponowane w algorytmie Apriori i jego licznych modyfikacjach [1, 3].

Reguły asocjacyjne mogą być definiowane w różny sposób. W ciągu ostatnich lat stosowany jest asocjacyjny mechanizm konstruowania reguł, gdzie poszczególne atrybuty systemu informacyjnego [7] mogą wystąpić w przesłankach lub konkluzji reguł.

W artykule reguły asocjacyjne są związane z regułami decyzyjnymi, tzn. w konkluzji występuje tylko jeden deskryptor (para „atrybut=wartość"). Podejście to było stosowane w [4, 6, 11-13].

W [5] pokazano, że dla szerokiej klasy binarnych systemów informacyjnych liczba nieredukowalnych reguł asocjacyjnych nie jest wielomianowa względem liczby atrybutów. W związku z tym poszukiwane są algorytmy, które pozwolą na konstruowanie „dobrych" reguł (o małej liczbie pomyłek) w rozsądnym czasie.

Praca ta jest kontynuacją badań [4, 12], w których pokazano, że biorąc pod uwagę założenia dotyczące klasy NP, algorytm zachłanny jest bliski najlepszym aproksymacyjnym algorytmom o złożoności wielomianowej dla minimalizacji długości częściowych reguł asocjacyjnych. Niestety, nie istnieją podobne wyniki dotyczące liczby błędnych zaklasyfikowań (liczby pomyłek). Celem badań jest pokazanie, że zaproponowana optymalizacja pozwala uzyskać częściowe reguły asocjacyjne o jak najmniejszej liczbie pomyłek, które są krótkie i mają dobre wsparcie. Ponadto zostaje zmniejszona liczba reguł, co jest istotne z punktu widzenia reprezentacji wiedzy. Uzyskane w ten sposób reguły zostaną w dalszych etapach badań wykorzystane do budowy klasyfikatorów.

W pracy badane są reguły przybliżone (częściowe). W literaturze można znaleźć wiele przykładów potwierdzających, iż często, zamiast dokładnych reguł z wieloma atrybutami, stosowane są częściowe reguły, zawierające mniejszą liczbą 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 MDL (ang. *Minimal Description 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. Celem prowadzonych badań jest uzyskanie zbioru częściowych reguł asocjacyjnych o jak najmniejszej liczbie pomyłek.

Artykuł składa się z 5 rozdziałów. W rozdziale 2 zostały przedstawione podstawowe pojęcia dotyczące częściowych reguł asocjacyjnych. Rozdział 3 prezentuje algorytm zachłanny oraz optymalizację reguł względem liczby pomyłek. 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 jest tabelą zawierającą n wierszy (odpowiadających obiektom $r_1,...,r_n$) oraz m kolumn (odpowiadających atrybutom $a_1,...,a_m$). Tabela ta wypełniona jest przez wartości atrybutów ze zbioru A odpowiadające obiektom ze zbioru U. Formalnie, system informacyjny definiowany jest jak para $I=(U,A)$ [7], gdzie $U=\{r_1,\ldots,r_n\}$ jest niepustym, skończonym zbiorem obiektów (wierszy), $A=\{a_1,\ldots,a_m\}$ jest niepustym, skończonym zbiorem atrybutów takim, że atrybut $a:U\rightarrow V_a$ jest funkcją, dla dowolnego $a\in A$, V_a jest zbiorem wartości atrybutu a. System informacyjny I może zostać przekształcony do postaci tablicy decyzyjnej T [7]. Wówczas jeden z atrybutów ze zbioru A staje się wyróżnionym atrybutem określanym jako atrybut decyzyjny i występuje w konkluzji reguł, pozostałe atrybuty nazywane są atrybutami warunkowymi. Formalnie tablicę decyzyjną można zdefiniować jako $T=(U,A\cup\{d\})$, gdzie $U=\{r_1,\ldots,r_n\}$ jest niepustym, skończonym zbiorem obiektów (wierszy), $A=\{a_1,\ldots,a_{m-1}\}$ jest niepustym, skończonym zbiorem atrybutów, $a:U\rightarrow V_a$ jest funkcją dla dowolnego $a\in A$, V_a jest zbiorem wartości atrybutu a. Elementy zbioru A są nazywane atrybutami warunkowymi, $d\notin A$ jest wyróżnionym atrybutem, nazywanym atrybutem decyzyjnym.

Niech $r=(b_1,\ldots,b_m)$ będzie wierszem z I opisanym przez wartości atrybutów $b_1,\ldots,b_m$, a_p niech będzie 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 $j\in\{1,\ldots,m\}\setminus\{p\}$. Powiemy, że atrybut a_i separuje (oddziela) wiersz $r'\in U(I,r,a_p)$ od wiersza r, jeśli wiersze te mają różne wartości na przecięciu z kolumną a_i. Trójka (I,r,a_p) nazywana jest *problemem reguły asocjacyjnej* [4].

W celu definiowania reguł przybliżonych zastosowano parametr α oraz wartość $|U(I,r,a_p)|$. Niech α będzie liczbą rzeczywistą taką, że $0 \leq \alpha < 1$. Reguła

$$a_{i1} = b_{i1} \wedge \ldots \wedge a_{im} = b_{im} \rightarrow a_p = b_p \quad (1)$$

jest nazywana *α-regułą asocjacyjną* (*częściową regułą asocjacyjną*) dla (I,r,a_p), jeśli atrybuty występujące w części warunkowej reguły oddzielają od wiersza r przynajmniej $\lceil (1-\alpha)|U(I,r,a_p)| \rceil$ wierszy ze zbioru $U(I,r,a_p)$. Na przykład, 0.01-*reguła asocjacyjna* oznacza, że należy oddzielić od wiersza r przynajmniej 99% wierszy dotychczas nieoddzielonych ze zbioru $U(I,r,a_p)$.

Liczba pomyłek (błędnych zaklasyfikowań) reguły (1) to liczba wierszy w tablicy decyzyjnej, dla których prawdziwa jest część warunkowa reguły (wartości atrybutów warunkowych są takie same jak w regule), a wartość atrybutu decyzyjnego jest inna niż wartość atrybutu a_p występującego w konkluzji reguły. Wsparcie reguły (1) to liczba wierszy w tablicy decyzyjnej, dla których prawdziwa jest część warunkowa reguły i wartość atrybutu decyzyjnego jest taka sama jak wartość atrybutu a_p występującego w konkluzji reguły. Długość reguły (1) to liczba deskryptorów (par „*atrybut=wartość*") występujących w części warunkowej reguły.

3. Algorytm zachłanny

Poniżej został przedstawiony pseudokod algorytmu zachłannego [12], który konstruuje α-regułę asocjacyjną dla trójki (I,r,a_p). Algorytm ten jest stosowany sekwencyjnie dla każdego wiersza 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 asocja-
cyjna.

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

BEGIN
FOR i=1 to m
   ap=ai;        //ap będzie atrybutem występującym w konkluzji 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 atrybut ai∈{a1,…,am}\{ap} o minimalnym indeksie i, który oddziela
       największą liczbę wierszy ze zbioru U(I,r,ap) dotychczas nieoddzielonych
       przez atrybuty ze zbioru Q;
       Q ← Q ∪ {ai};
       Atrybuty zawarte w zbiorze Q tworzą warunki α-reguły asocjacyjnej.
       END
   END
END
```

Liczba pomyłek α-reguły asocjacyjnej skonstruowanej dla trójki (I,r,a_p) jest oznaczana jako $M_{greedy}(\alpha,I,r,a_p)$.

Poniżej zostanie przedstawiona optymalizacja α-reguł asocjacyjnych konstruowanych przez algorytm zachłanny względem liczby pomyłek. Minimalna liczba pomyłek α-reguły asocjacyjnej dla wiersza r systemu informacyjnego I i α jest oznaczana jako $Min^{\mu}(\alpha,I,r)$. Wartość ta jest wyznaczana na podstawie liczby pomyłek α-reguł asocjacyjnych skonstruowanych dla trójki (I,r,a_p), p=1,…,m.

$$Min^{\mu}(\alpha,I,r)=min\{ M_{greedy}(\alpha,I,r,a_p): p=1,\ldots,m \}.$$

W wyniku optymalizacji, spośród wszystkich α-reguł asocjacyjnych dla (I,r,a_p), p=1,…,m, wybierane są tylko te reguły, dla których spełnione jest równanie:

$$M_{greedy}(\alpha,I,r,a_p)= Min^{\mu}(\alpha,I,r).$$

Optymalizacja przeprowadzana jest dla każdego wiersza systemu informacyjnego I.

4. Wyniki eksperymentów

Rozdział ten przedstawia wyniki eksperymentów, które miały na celu:

- porównanie liczby pomyłek α-reguł asocjacyjnych konstruowanych przez algorytm zachłanny przed optymalizacją i po optymalizacji,
- porównanie liczby α-reguł asocjacyjnych konstruowanych przez algorytm zachłanny przed optymalizacją i po optymalizacji,
- porównanie długości α-reguł asocjacyjnych konstruowanych przez algorytm zachłanny po optymalizacji względem liczby pomyłek oraz po optymalizacji względem długości [13],
- wyznaczenie średniego wsparcia α-reguł asocjacyjnych po optymalizacji względem liczby pomyłek.

Celem eksperymentów jest pokazanie, że zaproponowana optymalizacja pozwala uzyskać częściowe reguły asocjacyjne o mniejszej liczbie pomyłek, które są krótkie i mają dobre wsparcie. Ponadto zostaje zmniejszona liczba reguł, co jest istotne z punktu widzenia reprezentacji wiedzy.

Eksperymenty zostały przeprowadzone na zbiorach danych umieszczonych w Repozytorium Uczenia Maszynowego [2]. Każdy zbiór danych traktowany był jako system informacyjny i dla każdego wiersza została skonstruowana przez algorytm zachłanny α-reguła asocjacyjna dla (I,r,a_p), p=1,…,m. Następnie została przeprowadzona optymalizacja α-reguł asocjacyjnych względem liczby pomyłek.

Tabela 1 przedstawia liczbę pomyłek α-reguł asocjacyjnych, α={0,05, 0,15, 0,25, 0,35} po optymalizacji. Dla wierszy z I została wyznaczona minimalna liczba pomyłek α-reguły asocjacyjnej (kolumna Min), średnia (kolumna Avg) i maksymalna liczba pomyłek α-reguły

asocjacyjnej (kolumna Max) po optymalizacji. Kolumna Wier oznacza liczbę wierszy w systemie informacyjnym *I*, kolumna Atr – liczbę atrybutów.

Tabela 1

Liczba pomyłek α-reguł asocjacyjnych po optymalizacji

			α=0,05			α=0,15			α=0,25			α=0,35		
Zbiór danych	**Wier**	**Atr**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**
Adult-stretch	16	5	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0	0
Balance-scale	625	5	1	3.0	8	7	19.6	22	27	33.8	54	27	33.8	54
Breast-cancer	266	10	0	0.3	3	0	1.4	8	0	2.8	16	0	3.1	16
Cars	1728	7	0	6.0	22	0	31.2	72	0	52.3	198	0	53.4	198
Flags	193	26	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0	0
Hayes-roth	69	5	0	0.2	1	0	0.5	3	0	2.9	10	0	5.4	10
House-votes	279	17	0	0.4	2	0	1.8	3	0	2.1	3	2	2.1	3
Lenses	24	5	0	0.0	0	0	0.2	1	0	0.4	2	0	0.2	3
Lymphography	148	18	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0	0
Monks-1-test	432	7	0	0.0	0	0	6.0	12	0	24.0	36	0	36.0	54
Monks-3-test	432	7	0	0.0	0	0	5.1	18	0	20.9	36	0	19.3	36
Shuttle-land.	15	7	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0	0
Soybean-small	47	36	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0	0
Teeth	23	9	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0	0
Tic-tac-toe	958	10	0	7.2	16	8	15.5	33	15	52.3	113	48	84.4	143
Zoo-data	59	17	0	0.0	0	0	0.0	0	0	0.0	0	0	0.0	0

Tabela 2

Porównanie liczby pomyłek α-reguł asocjacyjnych

Zbiór danych	**Wier**	**Atr**	**α=0,05**	**α=0,15**	**α=0,25**	**α=0,35**
Adult-stretch	16	5	[0,6]	[1,3]	[1,4]	[1,6]
Balance-scale	625	5	5,2	1,2	2,5	2,5
Breast-cancer	266	10	**10,0**	6,7	5,1	4,9
Cars	1728	7	3,9	2,0	4,8	4,7
Flags	193	26	[1,1]	[3]	[3,2]	[3,2]
Hayes-roth	69	5	6,5	5,2	2,6	1,7
House-votes	279	17	**8,0**	5,1	**6,9**	**9,4**
Lenses	24	5	[0,7]	**7,5**	4,8	**14,0**
Lymphography	148	18	[1,1]	[3,7]	[6,5]	[7]
Monks-1-test	432	7	[6]	3,6	2,3	1,9
Monks-3-test	432	7	[6,3]	3,9	2,6	3,3
Shuttle-landing	15	7	[0,3]	[0,4]	[0,6]	[0,7]
Soybean-small	47	36	[0,1]	[0,8]	[1,1]	[1,1]
Teeth	23	9	[0,1]	[0,1]	[0,3]	[0,3]
Tic-tac-toe	958	10	2,3	2,0	2,0	1,6
Zoo-data	59	17	[0,4]	[0,9]	[1,2]	[1,4]

Zostały także przeprowadzone eksperymenty, które pozwoliły uzyskać liczbę pomyłek α-reguł asocjacyjnych, α={0.05, 0.15, 0.25, 0.35}, przed optymalizacją. Tabela 2 przedstawia porównanie średniej liczby pomyłek α-reguł asocjacyjnych przed optymalizacją i po optymalizacji. Każda komórka tej tabeli zawiera wartość, która jest równa średniej liczbie pomyłek α-reguł asocjacyjnych przed optymalizacją podzieloną przez średnią liczbę pomyłek α-reguł asocjacyjnych po optymalizacji. Wartość w nawiasach kwadratowych oznacza średnią liczbę pomyłek przed optymalizacją, gdy średnia liczba pomyłek po optymalizacji wynosiła 0,0. Wykonane eksperymenty pokazują, że liczba pomyłek α-reguł asocjacyjnych jest niemalejąca wraz ze wzrostem wartości α. Ponadto liczba pomyłek α-reguł asocjacyjnych po optymalizacji, dla wszystkich zbiorów danych i wartości α={0,05, 0,15, 0,25, 0,35}, jest mniejsza niż przed optymalizacją. W szczególności, dla zbioru Brest-cancer i α=0,05 – 10 razy, dla zbioru House-votes i α=0,05 – 8 razy, dla zbioru Lenses i α=0,35 – 14 razy. Największe wartości różnicy zostały zapisane czcionką pogrubioną.

Tabela 3 przedstawia liczbę α-reguł asocjacyjnych przed optymalizacją (kolumna greedy) i po optymalizacji (kolumny oznaczone przez wartości α), α={0,05, 0,15, 0,25, 0,35}.

Tabela 3

Liczba α-reguł asocjacyjnych

Zbiór danych	**Wier**	**Atr**	**greedy**	**α=0,05**	**α=0,15**	**α=0,25**	**α=0,35**
Adult-stretch	16	5	80	32	24	24	20
Balance-scale	625	5	3125	700	2089	700	700
Breast-cancer	266	10	2660	607	394	413	419
Cars	1728	7	12096	1872	2371	1751	1751
Flags	193	26	5018	2964	2236	2220	2220
Hayes-roth	69	5	345	103	73	74	94
House-votes	279	17	4743	563	465	464	464
Lenses	24	5	120	60	44	39	33
Lymphography	148	18	2664	1388	760	688	687
Monks-1-test	432	7	3024	720	432	648	540
Monks-3-test	432	7	3024	612	588	648	432
Shuttle-landing	15	7	105	83	80	74	73
Soybean-small	47	36	1692	910	659	623	623
Teeth	23	9	207	195	189	185	185
Tic-tac-toe	958	10	9580	1220	1159	1107	1102
Zoo-data	59	17	1003	718	612	592	592

Tabela 4 przedstawia porównanie liczby α-reguł asocjacyjnych przed i po optymalizacji względem liczby pomyłek, α={0,05, 0,15, 0,25, 0,35}. Każda komórka tej tabeli zawiera wartość, która jest równa liczbie α-reguł asocjacyjnych przed optymalizacją podzielona przez liczbę α-reguł asocjacyjnych po optymalizacji.

Jednym z problemów związanych z algorytmem Apriori jest duża liczba konstruowanych reguł. Podejście zaproponowane w pracy nie ma tej wady, tzn. liczba konstruowanych reguł wynosi najwyżej *mn*, gdzie *m* jest liczbą atrybutów, a *n* liczbą obiektów dla danego systemu informacyjnego. Liczba reguł przed optymalizacją jest taka sama dla każdej wartości α. Wyniki przedstawione w tabelach 4 i 5 pokazują, że liczba reguł po optymalizacji znacznie się zmniejsza, w szczególności dla zbioru House-votes – około 10 razy, dla zbioru Tic-tac-toe – około 8 razy. Największe wartości różnicy zostały zapisane czcionką pogrubioną.

Tabela 4

Porównanie liczby α-reguł asocjacyjnych

Zbiór danych	**Wier**	**Atr**	**α=0,05**	**α=0,15**	**α=0,25**	**α=0,35**
Adult-stretch	16	5	2,5	3,3	3,3	4,0
Balance-scale	625	5	4,5	1,5	4,5	4,5
Breast-cancer	266	10	4,4	**6,8**	6,4	6,3
Cars	1728	7	**6,5**	5,1	**6,9**	**6,9**
Flags	193	26	1,7	2,2	2,3	2,3
Hayes-roth	69	5	3,3	4,7	4,7	3,7
House-votes	279	17	**8,4**	**10,2**	**10,2**	**10,2**
Lenses	24	5	2,0	2,7	3,1	3,6
Lymphography	148	18	1,9	3,5	3,9	3,9
Monks-1-test	432	7	4,2	**7,0**	4,7	5,6
Monks-3-test	432	7	4,9	5,1	4,7	**7,0**
Shuttle-landing	15	7	1,3	1,3	1,4	1,4
Soybean-small	47	36	1,9	2,6	2,7	2,7
Teeth	23	9	1,1	1,1	1,1	1,1
Tic-tac-toe	958	10	**7,9**	**8,3**	**8,7**	**8,7**
Zoo-data	59	17	1,4	1,6	1,7	1,7

Tabela 5

Długość α-reguł asocjacyjnych po optymalizacji względem liczby pomyłek

			α=0,05			**α=0,15**			**α=0,25**			**α=0,35**		
Zbiór danych	**Wier**	**Atr**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**
Adult-stretch	16	5	1	1,9	4	1	1,2	2	1	1,2	2	1	1,0	1
Balance-scale	625	5	2	2,0	2	2	2,0	2	1	1,0	1	1	1,0	1
Breast-cancer	266	10	1	1,8	3	1	1,2	2	1	1,0	2	1	1,0	1
Cars	1728	7	1	1,8	3	1	1,5	2	1	1,0	2	1	1,0	1
Flags	193	26	1	1,3	3	1	1,0	2	1	1,0	1	1	1,0	1
Hayes-roth	69	5	1	2,0	3	1	1,6	2	1	1,3	2	1	1,0	1
House-votes	279	17	1	2,2	5	1	1,3	3	1	1,1	2	1	1,0	2
Lenses	24	5	1	2,0	4	1	1,5	3	1	1,2	2	1	1,0	1
Lymphography	148	18	1	1,5	3	1	1,1	2	1	1,0	2	1	1,0	1
Monks-1-test	432	7	1	2,6	3	1	1,8	2	1	1,5	2	1	1,0	1
Monks-3-test	432	7	1	1,8	3	1	1,8	2	1	1,3	2	1	1,0	1
Shuttle-land.	15	7	1	1,2	5	1	1,1	4	1	1,0	2	1	1,0	1
Soybean-small	47	36	1	1,3	3	1	1,1	2	1	1,0	2	1	1,0	2
Teeth	23	9	1	1,1	3	1	1,0	2	1	1,0	1	1	1,0	1
Tic-tac-toe	958	10	2	2,8	3	2	2,0	2	1	1,5	2	1	1,0	2
Zoo-data	59	17	1	1,2	3	1	1,0	2	1	1,0	2	1	1,0	2

Tabela 5 przedstawia długość α-reguł asocjacyjnych, α={0,05, 0,15, 0,25, 0,35}, po optymalizacji względem liczby pomyłek. Dla każdego wiersza z *I* została uzyskana minimalna liczba pomyłek α-reguły asocjacyjnej dla *I*, *r* i α. Następnie, dla wierszy z *I* została wyznaczona minimalna długość α-reguły asocjacyjnej (kolumna Min), średnia (kolumna Avg) i maksymalna długość α-reguły asocjacyjnej (kolumna Max).

Tabela 6 przedstawia porównanie średniej długości α-reguł asocjacyjnych, α={0,05, 0,15, 0,25, 0,35}, po optymalizacji względem liczby pomyłek i po optymalizacji względem długości [13]. Każda komórka tej tabeli zawiera wartość, która jest równa średniej długości α-reguł asocjacyjnych po optymalizacji względem liczby pomyłek podzielonej przez średnią długość α-reguł asocjacyjnych po optymalizacji względem długości.

Tabela 6

Porównanie średniej długości α-reguł asocjacyjnych

Zbiór danych	**α=0,05**	**α=0,15**	**α=0,25**	**α=0,35**
Adult-stretch	1,9	1,2	1,2	1,0
Balance-scale	1,0	1,9	1,0	1,0
Breast-cancer	1,2	1,2	1,0	1,0
Cars	1,2	0,9	1,0	1,0
Flags	1,3	1,0	1,0	1,0
Hayes-roth	1,1	1,1	1,3	1,0
House-votes	**2,2**	1,3	1,1	1,0
Lenses	1,5	1,3	1,1	1,0
Lymphography	1,5	1,1	1,0	1,0
Monks-1-test	0,9	0,9	1,5	1,0
Monks-3-test	1,0	1,0	1,3	1,0
Shuttle-landing	1,2	1,1	1,0	1,0
Soybean-small	1,3	1,1	1,0	1,0
Teeth	1,1	1,0	1,0	1,0
Tic-tac-toe	1,3	1,0	1,4	1,0
Zoo-data	1,2	1,0	1,0	1,0

Wyniki w tabelach 5 i 6, pokazują że długość reguł maleje wraz ze wzrostem wartości α. Największa różnica występuje dla zbioru House-votes i α=0,05 – 2 razy. Dla większości zbiorów danych średnie długości reguł są zbliżone, dla α=0,35 – takie same. Zaproponowana optymalizacja względem liczby pomyłek pozwala uzyskać krótkie α-reguły asocjacyjne.

Tabela 7 przedstawia średnie wsparcie α-reguł asocjacyjnych po optymalizacji względem liczby pomyłek. Dla każdego wiersza z *I* została uzyskana wartość wsparcia α-reguły asocjacyjnej dla *I*, *r* i α. Następnie, dla wierszy z *I* została wyznaczona średnia wartość wsparcia α-reguły asocjacyjnej. Dla większości zbiorów danych średnie wartości wsparcia, dla α={0,05, 0,15, 0,25, 0,35}, są większe niż 10%. Tylko dla zbioru Balance-scale i α=0,15

wartość wsparcia jest bliska 1%. Natomiast dla zbiorów Adult-stretch i House-votes oraz dla α={0,15, 0,25} wartości wsparcia wynoszą powyżej 30%, dla α=0,35 – 40%.

Tabela 7

Średnie wsparcie α-reguł asocjacyjnych

Zbiór danych	**α=0,05**	**α=0,15**	**α=0,25**	**α=0,35**
Adult-stretch	29,7%	**37,5%**	**37,5%**	**40,0%**
Balance-scale	3,1%	0,9%	12,4%	12,4%
Breast-cancer	5,1%	8,7%	11,1%	13,1%
Cars	18,2%	16,6%	23,9%	24,0%
Flags	7,0%	8,4%	8,4%	8,4%
Hayes-roth	7,8%	9,7%	11,5%	14,1%
House-votes	20,5%	**32,8%**	**39,3%**	**41,5%**
Lenses	19,7%	24,4%	27,0%	**30,7%**
Lymphography	17,7%	26,0%	28,0%	28,0%
Monks-1-test	7,4%	11,8%	12,5%	16,7%
Monks-3-test	11,3%	10,1%	18,3%	25,2%
Shuttle-landing	13,9%	14,2%	14,5%	14,5%
Soybean-small	16,0%	20,3%	21,2%	21,2%
Teeth	14,6%	15,0%	15,2%	15,2%
Tic-tac-toe	1,8%	4,5%	8,3%	13,1%
Zoo-data	21,1%	23,1%	23,4%	23,4%

5. Podsumowanie

W pracy została zaproponowana optymalizacja α-reguł asocjacyjnych względem liczby pomyłek. Podejście to różni się od podejścia opartego na zbiorach częstych, ale pozwala uzyskać reguły o stosunkowo dobrej jakości w rozsądnym czasie. Przeprowadzone eksperymenty pokazały, że stosując optymalizację, można uzyskać częściowe reguły asocjacyjne o mniejszej liczbie pomyłek, które są krótkie i mają dobre wsparcie. Ponadto znacznie zmniejszyła się liczba reguł, co jest istotne z punktu widzenia reprezentacji wiedzy. Uzyskane w wyniku optymalizacji reguły zostaną w dalszych etapach badań wykorzystane do budowy klasyfikatorów.

BIBLIOGRAFIA

1. Agrawal R., Srikant R.: Fast algorithms for mining association rules in large databases. [in:] Bocca J.B., Jarke M., Zaniolo C. (eds.): Proceedings of 20th International Conference on Very Large Data Bases, Morgan Kaufmann, 1994.

2. Asuncion A., Newman D.: UCI machine learning repository, http://www.ics.uci.edu/~mlearn/MLRepository.html, School of Information and Computer Sciences, University of California, Irvine (dostęp luty 2016).
3. Borgelt C.: Simple algorithms for frequent item set mining. [in:] Koronacki J., Raś Z. W., Wierzchoń S.T., Kacprzyk J. (eds.): Advances in Machine Learning II, Studies in Computational Intelligence, Vol. 263, Springer, 2010, s. 351÷369.
4. Moshkov M.J., Piliszczuk M., Zielosko B.: Greedy algorithm for construction of partial association rules. Fundamenta Informaticae, Vol. 92(3), 2009, s. 259÷277.
5. Moshkov M.J., Skowron A., Suraj Z.: On minimal rule sets for almost all binary information systems. Fundamenta Informaticae, Vol. 80(1-3), 2007, s. 247÷258.
6. Nguyen H.S., Ślęzak D.: Approximate reducts and association rules – correspondence and complexity results. [in:] Zhong N., Skowron A., Ohsuga S. (eds.): Proc. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, LNCS (LNAI), Vol. 1711, Springer, Heidelberg 1999.
7. Pawlak Z.: Systemy Informacyjne – Podstawy teoretyczne. Wydawnictwa Naukowo-Techniczne, Warszawa 1983.
8. Rissanen J.: Modeling by shortest data description. Automatica, Vol. 14, 1978, s. 465÷471.
9. Skowron A.: Rough sets in KDD. [in:] Shi Z., Faltings B., Musen M. (eds.): Proc. 16th IFIP World Computer Congress, Publishing House of Electronic Industry, 2000.
10. Stefanowski J., Vanderpooten D.: Induction of decision rules in classification and discovery-oriented perspectives. International Journal of Intelligent Systems, Vol. 16(1), 2001, s. 13÷27.
11. Wieczorek A., Słowiński R.: Generating a set of association and decision rules with statistically representative support and anti-support. Inf. Sci., Vol. 277, 2014, s. 56÷70.
12. Zielosko B.: Algorytm zachłanny dla konstruowania częściowych reguł asocjacyjnych. Studia Informatica, Vol. 31, No. 2A(89), Gliwice 2010, s. 225÷236.
13. Zielosko B., Robaszkiewicz M.: Greedy Algorithm for Optimization of Association Rules Relative to Length. KES Smart Innovation Systems and Technologies, 2016 (artykuł przyjęty do publikacji).

Abstract

In the paper, an optimization of partial association rules relative to number of misclassifications is presented. The aims of proposed optimization are: (i) construction of rules with small number of misclassifications, what is important from the point of view of construction

of classifiers, (ii) decreasing the number of rules, what is important from the point of view of knowledge representation.

The paper can be considered as a continuation of reserach presented in [13]. Greedy algorithm is different from the Apriori algorithm and its modifications based on frequent itemsets. However, it allows one to obtain "important", in some sense, rules in a reasonable time.

Experimental results show that proposed optimization relative to number of misclassifications allows us to obtain α-association rules with small number of misclassifications, enough good coverage and small length. Obtained rules will be used for construction of classifiers.

Adresy

Beata ZIELOSKO: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39,
41-200 Sosnowiec, Polska, beata.zielosko@us.edu.pl.
Marek ROBASZKIEWICZ: EL-PLUS Sp. z o.o., ul. Inwalidzka 11, 41-506 Chorzów, Polska,
marek.robaszkiewcz@gmail.com.