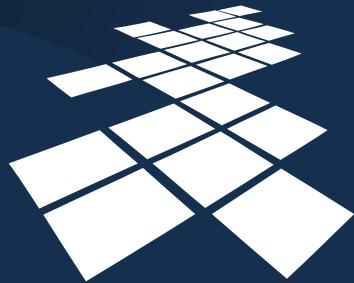


Odkrywanie asocjacji

Wprowadzenie
Sformułowanie problemu
Typy reguł asocjacyjnych



**UCZELNIA
ONLINE**

Odkrywanie asocjacji – wykład 1

Wykład jest poświęcony wprowadzeniu i zaznajomieniu się z problemem odkrywania reguł asocjacyjnych. Przybliżymy problem analizy koszyka zakupów (MBA) oraz pokażemy jak zamodelować różne przypadki świata rzeczywistego w postaci tablicy informacji. Dokonamy klasyfikacji reguł oraz objaśnimy każdy z typów reguł asocjacyjnych. Na koniec przedstawimy algorytmy: algorytm naiwny oraz ogólny algorytm odkrywania reguł asocjacyjnych.



Geneza problemu

- Geneza problemu odkrywania reguł asocjacyjnych:

problem analizy koszyka zakupów
(MBA – Market Basket Analysis)

- Dane:

baza danych zawierająca informacje o zakupach realizowanych przez klientów supermarketu

- Cel:

znalezienie grup produktów, które klienci supermarketu najczęściej kupują razem

Odkrywanie asocjacji (2)

Odkrywanie asocjacji jest jedną z najciekawszych i najbardziej popularnych technik eksploracji danych. Celem procesu odkrywania asocjacji jest znalezienie interesujących zależności lub korelacji, nazwanych ogólnie asocjacjami, pomiędzy danymi w dużych zbiorach danych. Wynikiem procesu odkrywania asocjacji jest zbiór reguł asocjacyjnych opisujących znalezione zależności lub korelacje pomiędzy danymi. Geneza problemu odkrywania reguł asocjacyjnych sięga problemu odkrywania asocjacji rozważanego w kontekście tak zwanej analizy koszyka zakupów (MBA - Market Basket Analysis). Klasyczny problem analizy koszyka zakupów polega na analizie danych zawierających informacje o zakupach zrealizowanych przez klientów supermarketu. Celem takiej analizy jest znalezienie naturalnych wzorców zachowań konsumenckich klientów poprzez analizę produktów (grup produktów), najczęściej kupowanych razem przez klientów supermarketu.



Analiza koszyka zakupów

- Cel analizy MBA:

znanie naturalnych wzorców zachowań konsumenckich klientów

- Wykorzystanie wzorców zachowań

organizacji półek w supermarketie

opracowania akcji promocyjnych

opracowania katalogu oferowanych produktów

Odkrywanie asocjacji (3)

Celem tej analizy jest znanie naturalnych wzorców zachowań konsumenckich klientów poprzez analizę produktów, które są przez klientów supermarketu kupowane najczęściej wspólnie (tj. określenie grup produktów, które klienci najczęściej umieszczają w swoich koszykach – stąd nazwa problemu). Znalezione wzorce zachowań klientów mogą być, następnie, wykorzystane do opracowania akcji promocyjnych, organizacji półek w supermarketie, opracowania koncepcji katalogu oferowanych produktów itd.



Zastosowanie MBA (1)

- Znaleziony wzorzec:

„ktoś kto kupuje pieluszki, najczęściej kupuje również piwo”

- Akcja promocyjna: (typowy trick)

Ogłoś obniżkę cen pieluszek, jednocześnie podnieś piwa

- Organizacja sklepu:

Staraj się umieszczać produkty kupowane wspólnie w przeciwnieństwach końcach sklepu, zmuszając klientów do przejścia przez cały sklep

Odkrywanie asocjacji (4)

Założymy, przykładowo, że znaleziono w bazie danych regułę asocjacyjną (wzorzec) postaci „**ktoś kto kupuje pieluszki, najczęściej kupuje również piwo**”. Wiedza, jaką niesie powyższa reguła, może być wykorzystana wielokrotnie, np.: organizując akcję promocyjną można zastosować typowy trick pod hasłem 10% obniżki ceny pieluszek. Naturalną konsekwencją akcji powinno być zwiększenie liczby klientów kupujących pieluszki i tym samym, zrekompensowanie straty wynikającej z oferowanej obniżki ceny pieluszek. Jednakże, dodatkowym źródłem dochodu może być również zwiększyły zysk z tytułu dodatkowej sprzedaży piwa kupowanej najczęściej z pieluszkami. Innym przykładem wykorzystania zdobytej wiedzy może być reorganizacja półek sklepowych w taki sposób aby umieścić produkty kupowane wspólnie obok siebie na półkach co może wpływać na zwiększenie sprzedaży obu produktów odwołując się do naturalnych preferencji konsumenckich klientów. Można również wykorzystać sytuację odwrotną umieszczając produkty kupowane wspólnie w przeciwnieństwach końcach sklepu, zmuszając klientów do przejścia przez cały sklep.



Zastosowanie MBA (2)

- **MBA** znajduje zastosowanie wszędzie tam, gdzie „klienci” nabierają łącznie pewien zbiór dóbr lub usług

Analiza pogody (koszykiem jest zbiór zdarzeń pogodowych, które wystąpiły w danym przedziale czasu)

Telekomunikacja (koszykiem jest zbiór rozmów telefonicznych)

Diagnostyka medyczna

Karty kredytowe

Bankowość

Odkrywanie asocjacji (5)

Market Basket Analysis znajduje zastosowanie wszędzie tam, gdzie „klienci” nabierają łącznie pewien zbiór dóbr lub usług:
może to być analiza pogody, w której koszykiem będzie zbiór zdarzeń pogodowych, występujących w danym przedziale czasu. Telekomunikacja, gdzie koszykiem będzie zbiór rozmów telefonicznych, oraz wiele innych dziedzin życia np.: diagnostyka medyczna czy też bankowość.



Model koszyka zakupów

Model koszyka zakupów jest pewną abstrakcją umożliwiającą modelowanie relacji wiele-do-wiele pomiędzy encjami „produkty” i „koszyki”



Formalnie, model koszyka zakupów można opisać za pomocą tzw. **tablicy obserwacji**

Odkrywanie asocjacji (6)

Modelując koszyk zakupów, możemy odnieść się do pewnej abstrakcji umożliwiającej modelowanie relacji wiele-do-wiele pomiędzy wspomnianymi encjami „Produkty” i „Koszyki”. Model koszyka zakupów modelujemy najczęściej w postaci tzw. tablicy obserwacji.

Tablica obserwacji (1)

- Dany jest zbiór atrybutów $A = \{A_1, A_2, \dots, A_n\}$ oraz zbiór obserwacji $T = \{T_1, T_2, \dots, T_m\}$

TR_id	A ₁	A ₂	A ₃	A ₄	A ₅
T ₁	1	0	0	0	1
T ₂	1	1	1	1	1
T ₃	1	0	1	0	0
T ₄	0	0	1	0	1
T ₅	0	1	1	1	1
T ₆	1	1	1	0	1
T ₇	1	0	1	1	1
T ₈	1	1	1	0	0

Odkrywanie asocjacji (7)

Dany jest zbiór atrybutów $A = \{A_1, A_2, \dots, A_n\}$. Przykładowa tablica obserwacji D dla zbioru atrybutów A, zawierająca 8 obserwacji $\{T_1, \dots, T_8\}$. Przedstawiona tablica obserwacji dla celów ilustracyjnych, poza zbiorem atrybutów A, zawiera dodatkową kolumnę TR_id, której wartościami są identyfikatory poszczególnych obserwacji. Za pomocą tablicy obserwacji można zamodelować różne przypadki świata rzeczywistego.



Tablica obserwacji (2)

- Elementy tablicy obserwacji:

- Atrybuty tablicy reprezentują wystąpienia encji „produkty”

- Wiersze tablicy reprezentują wystąpienia encji „koszyki”

- Dodatkowy atrybut TR_id – wartościami atrybutu są identyfikatory poszczególnych obserwacji

- Pozycja $T_i[A_j] = 1$ tablicy wskazuje, że i-ta obserwacja zawiera wystąpienie j-tego atrybutu

Odkrywanie asocjacji (8)

Tablicę obserwacji można wykorzystać również do analizy koszyka zakupów. Zbiór atrybutów tablicy obserwacji odpowiada liście produktów oferowanych przez supermarket, natomiast wiersze tablicy reprezentują klientów i ich koszyki zakupów. Dodatkowy atrybut Tr_id przedstawia identyfikatory poszczególnych obserwacji. Pozycja $T_i[A_j] = 1$ tablicy wskazuje, że i-ta obserwacja zawiera wystąpienie j-tego atrybutu.



Tablica obserwacji (3)

- „koszyki” = studenci, „produkty” = wykłady oferowane przez uczelnię

MBA – poszukiwanie wykładów, które studenci wybierają najczęściej łącznie

- „koszyki” = strony WWW, „produkty” = słowa kluczowe

MBA – poszukiwanie stron WWW opisanych tymi samymi, lub podobnymi lub podobnymi, zbiorami słów kluczowych (prawdopodobnie, znalezione strony dotyczą podobnej problematyki)

Odkrywanie asocjacji (9)

Tablica obserwacji może posłużyć na przykład, do opisu wyboru przedmiotów obieralnych przez studentów. W takim przypadku zbiór atrybutów będzie odpowiadał liście przedmiotów obieralnych oferowanych studentom przez uczelnię, natomiast wiersze tablicy obserwacji będą reprezentować studentów i ich wybory przedmiotów. Weźmy pod uwagę inny przykład, w którym koszykiem będą strony WWW natomiast produktami – słowa kluczowe, dla takiej reprezentacji możemy sformułować problem MBA jako poszukiwanie stron WWW opisanych takimi samymi, lub podobnymi, zbiorami słów kluczowych. Możemy przypuszczać iż prawdopodobnie znalezione strony będą dotyczyć podobnej tematyki.



Skala problemu

- Rozwiążanie problemu MBA musi być skalowalne:

Supermarket sprzedaje ponad 150 000 produktów i przechowuje informacje o miliardach wykonanych transakcji rocznie

Web zawiera kilka miliardów stron i zawiera ponad 100 milionów słów

Odkrywanie asocjacji (10)

Analizując problem MBA, podając jego rozwiązanie musimy pamiętać o jego skalowalności. Dane, które podane są analizie często osiągają olbrzymie rozmiary. Np. supermarket sprzedaj ponad 150 000 produktów i przechowuje informacje o miliardach wykonanych transakcji rocznie, albo biorąc pod uwagę Web gdzie mamy do czynienia z kilkoma miliardami stron zawierającymi ponad 100 milionów słów.



Reguły asocjacyjne (1)

- Wynikiem analizy koszyka jest zbiór reguł asocjacyjnych postaci następującej relacji:

$$\{(A_{i1} = 1) \wedge \dots \wedge (A_{ik} = 1)\} \rightarrow \{(A_{ik+1} = 1) \wedge \dots \wedge (A_{ik+l} = 1)\} \quad (1)$$

Interpretacja reguły:

„jeżeli klient kupił produkty $A_{i1}, A_{i2}, \dots, A_{ik}$, to prawdopodobnie kupił również produkty $A_{ik+1}, A_{ik+2}, \dots, A_{ik+l}$ ”

Odkrywanie asocjacji (11)

Wynik analizy koszyka zakupów przedstawiany jest w formie zbioru reguł asocjacyjnych postaci relacji przedstawionej na powyższym slajdzie oznaczonej (1). Taką regułę można interpretować w następujący sposób: „jeżeli klient kupił produkty $A_{i1}, A_{i2}, \dots, A_{ik}$, to prawdopodobnie kupił również produkty $A_{ik+1}, A_{ik+2}, \dots, A_{ik+l}$ ”.



Reguły asocjacyjne (2)

- **Regułę asocjacyjną** (1) można przedstawić jednoznacznie w równoważnej postaci

$\theta \rightarrow \varphi$:

$$(A_{i1}, A_{i2}, \dots, A_{ik}) \rightarrow (A_{ik+1}, A_{ik+2}, \dots, A_{ik+l})$$

- Z każdą regułą asocjacyjną $\theta \rightarrow \varphi$ związane są dwie podstawowe miary określające statystyczną ważność i siłę reguły:

wsparcie - sup($\theta \rightarrow \varphi$)

ufność - conf($\theta \rightarrow \varphi$)

Odkrywanie asocjacji (12)

Regułę asocjacyjną można przedstawić jednoznacznie w równoważnej postaci $\theta \rightarrow \varphi: (A_{i1}, A_{i2}, \dots, A_{ik}) \rightarrow (A_{ik+1}, A_{ik+2}, \dots, A_{ik+l})$.

Z każdą binarną regułą asocjacyjną $\theta \rightarrow \varphi$, są związane dwie miary określające statystyczną ważność i siłę

reguły: {wsparcie} reguły (ang. support) oraz {ufność} reguły (ang. confidence).



Reguły asocjacyjne (3)

- **Statystyczna ważność i siła reguły:**

Wsparciem sup reguły asocjacyjnej $\theta \rightarrow \phi$ nazywać będziemy stosunek liczby obserwacji, które spełniają warunek $\theta \wedge \phi$, do liczby wszystkich obserwacji (wsparcie reguły = prawdopodobieństwu zajścia zdarzenia $\theta \wedge \phi$)

Ufnością conf reguły asocjacyjnej $\theta \rightarrow \phi$ nazywać będziemy stosunek liczby obserwacji, które spełniają warunek $\theta \wedge \phi$, do liczby obserwacji, które spełniają warunek θ (ufność reguły = warunkowemu prawdopodobieństwu $p(\phi | \theta)$)

Odkrywanie asocjacji (13)

{Wsparciem} (sup) reguły asocjacyjnej $\theta \rightarrow \phi$ nazywać będziemy stosunek liczby obserwacji, które spełniają warunek $\theta \wedge \phi$, do liczby wszystkich obserwacji, przy czym wsparcie reguły jest równe prawdopodobieństwu zajścia zdarzenia $\theta \wedge \phi$.

Ufnością conf reguły asocjacyjnej $\theta \rightarrow \phi$ nazywać będziemy stosunek liczby obserwacji, które spełniają warunek $\theta \wedge \phi$, do liczby obserwacji, które spełniają warunek θ (ufność reguły = warunkowemu prawdopodobieństwu $p(\phi | \theta)$). Łatwo zauważyc, że ufność reguły jest równa warunkowemu prawdopodobieństwu zajścia zdarzenia ϕ pod warunkiem zajścia zdarzenia $p(\phi | \theta)$



Klasyfikacja reguł asocjacyjnych

- Klasyfikacja reguł asocjacyjnych ze względu na:
 - Typ przetwarzanych danych
 - Wymiarowość przetwarzanych danych
 - Stopień abstrakcji przetwarzanych danych
- Inne typy reguł asocjacyjnych
- Asocjacje vs. analiza korelacji

Odkrywanie asocjacji (14)

W literaturze poświęconej eksploracji danych można znaleźć wiele rodzajów reguł asocjacyjnych. Reguły te można sklasyfikować według szeregu kryteriów. Wśród tych kryteriów podstawowe znaczenie mają trzy kryteria:

typ przetwarzanych danych, wymiarowość przetwarzanych danych oraz stopień abstrakcji przetwarzanych danych.

W dalszej części wykładu, krótko scharakteryzujemy poszczególne kryteria klasyfikacji i przedstawimy rodzaje reguł asocjacyjnych wynikające z tych klasyfikacji.



Typ przetwarzanych danych (1)

- Wyróżniamy:

- **binarne reguły asocjacyjne**

- **ilościowe reguły asocjacyjne**

- Regułę asocjacyjną nazywamy **binarną**, jeżeli dane występujące w regule są danymi (zmiennymi) binarnymi
- Regułę asocjacyjną nazywamy **ilościową**, jeżeli dane występujące w regule są danymi ciągłymi i/lub kategorycznymi

Odkrywanie asocjacji (15)

Z punktu widzenia typu przetwarzanych danych wyróżniamy dwa rodzaje reguł asocjacyjnych: (1) {binarne reguły asocjacyjne} (ang. binary lub Boolean association rules) oraz (2) {ilościowe reguły asocjacyjne} (ang. quantitative association rules). Regułę asocjacyjną nazywamy {binarną regułą asocjacyjną}, jeżeli dane występujące w regule są danymi (zmiennymi) binarnymi, to znaczy, danymi, które mogą przyjmować tylko dwie wartości: '1' ({true}) lub '0' ({false}). Regułę asocjacyjną nazywamy {ilościową regułą asocjacyjną}, jeżeli dane występujące w regule są danymi ciągłymi i/lub kategorycznymi. Ilościowe reguły asocjacyjne reprezentują, najogólniej mówiąc, współwystępowanie wartości niektórych danych.



Typ przetwarzanych danych (2)

- Binarna reguła asocjacyjna:

$pieluszki = 1 \rightarrow piwo = 1$

– (reprezentuje współwystępowanie danych)

- Ilościowa reguła asocjacyjna:

-

$wiek = '30...40' \wedge wykształcenie = 'wyższe'$
 $\rightarrow opcja_polityczna = 'demokrata'$

(reprezentuje współwystępowanie wartości danych)

Odkrywanie asocjacji (16)

Binarne reguły asocjacyjne reprezentują, najogólniej mówiąc, współwystępowanie danych. Przykładem binarnej reguły asocjacyjnej może być reguła: „pieluszki=1 -> piwo=1”; Reguła ta wywiedziona w ramach analizy koszyka zakupów klientów supermarketu, stwierdza, że produkt 'pieluszki' często występuje w koszykach klientów łącznie z produktem 'piwo'. Przykładem ilościowej reguły asocjacyjnej jest reguła: „wiek ='30...40' \wedge wykształcenie = 'wyższe' -> opcja_polityczna = 'demokrata'”. Ilościowe reguły asocjacyjne reprezentują, najogólniej mówiąc, współwystępowanie wartości niektórych danych. Reguła wywiedziona z analizy danych osobowych, stwierdza, że jeżeli wiek pracownika należy do przedziału wartości '30...40' i pracownik posiada wykształcenie wyższe, to, często, jego poglądy polityczne zwrotne są w kierunku demokracji. Atrybut {wiek} jest atrybutem ciągłym, natomiast atrybuty {wykształcenie} oraz opcja_polityczna} są atrybutami kategorycznymi. W procesie odkrywania ilościowych reguł asocjacyjnych, atrybuty ciągłe podlegają dyskretyzacji. Stąd, w regule wartością atrybutu {wiek} jest pewien przedział wartości.



Wymiarowość przetwarzanych danych (1)

- Wyróżniamy:

- **jednowymiarowe reguły asocjacyjne**

- **wielowymiarowe reguły asocjacyjne**

- Regułę asocjacyjną nazywamy **jednowymiarową**, jeżeli dane występujące w regule reprezentują tę samą dziedzinę wartości
- Regułę asocjacyjną nazywamy **wielowymiarową**, jeżeli dane występujące w regule reprezentują różne dziedziny wartości

Odkrywanie asocjacji (17)

Z punktu widzenia wymiarowości przetwarzanych danych wyróżniamy dwa rodzaje reguł asocjacyjnych: (1) {jednowymiarowe reguły asocjacyjne} (ang. single-dimensional association rules) oraz (2) {wielowymiarowe reguły asocjacyjne} (ang. multidimensional association rules). Regułę asocjacyjną nazywamy {jednowymiarową regułą asocjacyjną}, jeżeli dane występujące w regule reprezentują tę samą dziedzinę wartości. Regułę asocjacyjną nazywamy {wielowymiarową regułą asocjacyjną}, jeżeli dane występujące w regule reprezentują różne dziedziny wartości. Pojęcie {wymiaru} wywodzi się z terminologii magazynów danych, gdzie pojawia się w kontekście pojęcia {analiza wielowymiarowa danych}.



Wymiarowość przetwarzanych danych (2)

- Jednowymiarowa reguła asocjacyjna:

$pieluszki = 1 \rightarrow piwo = 1$

- Wielowymiarowa reguła asocjacyjna:

$wiek = '30...40' \wedge wykształcenie = 'wyższe'$
 $\rightarrow opcja_polityczna = 'demokrata'$

Odkrywanie asocjacji (18)

Przykładem reguły jednowymiarowej będzie reguła asocjacyjna „ $pieluszki = 1 \rightarrow piwo = 1$ ”, natomiast reguła „ $wiek = '30...40' \wedge wykształcenie = 'wyższe' \rightarrow opcja_polityczna = 'demokrata'$ ” - jest wielowymiarową regułą asocjacyjną, gdyż występują w niej trzy wymiary: { $wiek$ }, { $wykształcenie$ }, oraz { $opcja_polityczna$ }. Każdy z wymiarów jest reprezentowany jako osobny predykat reguły.



Stopień abstrakcji przetwarzanych danych (1)

- Wyróżniamy:

- **jednopoziomowe reguły asocjacyjne**

- **wielopoziomowe reguły asocjacyjne**

- Regułę asocjacyjną nazywamy **jednopoziomową**, jeżeli dane występujące w regule reprezentują ten sam poziom abstrakcji
- Regułę asocjacyjną nazywamy **wielopoziomową**, jeżeli dane występujące w regule reprezentują różne poziomy abstrakcji

Odkrywanie asocjacji (19)

Z punktu widzenia stopnia abstrakcji przetwarzanych danych wyróżniamy dwa rodzaje reguł asocjacyjnych: (1) {jednopoziomowe reguły asocjacyjne} (ang. single-level association rules) oraz (2) {wielopoziomowe} lub {uogólnione reguły asocjacyjne} (ang. multilevel lub generalized association rules). Regułę asocjacyjną nazywamy {jednopoziomową regułą asocjacyjną}, jeżeli dane występujące w regule reprezentują ten sam poziom abstrakcji. Przykładem takich danych są konkretne produkty zakupione w supermarkecie, wykłady, na które zarejestrowali się studenci na studiach, słowa kluczowe występujące w dokumentach tekstowych, itd. Czasami dane występujące w bazie danych tworzą pewną hierarchię poziomów abstrakcji. Przykładowo, produkty w supermarkecie można poklasifykować według kategorii produktu: produkt 'pieluszki_Pampers' należy do kategorii 'pieluszki', która, z kolei, należy do kategorii 'środki_czystości'; produkt 'piwo_żywiec' należy do kategorii 'piwo', która, z kolei, należy do kategorii 'napoje_alkoholowe' i, dalej, do kategorii 'napoje'; wreszcie, produkt 'czekolada_milka' należy do kategorii 'słodycze'. Reguły, które opisują asocjacje występujące pomiędzy danymi reprezentującymi różne poziomy abstrakcji, nazywamy {wielopoziomowymi regułami asocjacyjnymi}.



Stopień abstrakcji przetwarzanych danych (2)

- Jednopoziomowa reguła asocjacyjna:

$$\text{pieluszki_Pampers} = 1 \rightarrow \text{piwo_Żywiec} = 1$$

- Wielopoziomowa reguła asocjacyjna:

$$\text{pieluszki_Pampers} = 1 \wedge \text{piwo_Żywiec} = 1 \rightarrow \text{napoje} = 1$$

(produkt napoje reprezentuje pewna abstrakcję, będącą generalizacją określonych produktów)

Odkrywanie asocjacji (20)

Przykładem jednopoziomowej reguły asocjacyjnej jest reguła: „pieluszki_Pampers = 1 → piwo_Żywiec = 1”.
Przykładem wielopoziomowej reguły asocjacyjnej jest reguła: „pieluszki_Pampers = 1 ∧ piwo_Żywiec = 1 → napoje = 1”.
Zauważmy, że dane występujące w regule reprezentują różne poziomy abstrakcji: dana ‘napoje’ reprezentuje wyższy poziom abstrakcji aniżeli dana opisująca konkretny produkt ‘piwo_Żywiec’. Produkt napoje reprezentuje pewną abstrakcję, będącą generalizacją określonych produktów.



Odkrywanie binarnych reguł asocjacyjnych

- Dane:
 - $I = \{i_1, i_2, \dots, i_n\}$: zbiór literałów, nazywanych dalej **elementami**
 - Transakcja T : zbiór elementów, takich że $T \subseteq I$ i $T \neq \emptyset$
 - Baza danych D : zbiór transakcji
 - Transakcja T wspiera element $x \in I$, jeżeli $x \in T$
 - Transakcja T wspiera zbiór $X \subseteq I$, jeżeli T wspiera każdy element ze zbioru X , $X \subseteq T$

Odkrywanie asocjacji (21)

Sformułowanie problemu odkrywania silnych jednopoziomowych jednowymiarowych binarnych reguł asocjacyjnych, przedstawione wcześniej abstrahuje od rzeczywistych metod przechowywania danych. Założenie, że obiektem eksploracji danych jest zerojedynkowa tablica obserwacji jest mało realistyczne w praktyce. Stąd, w literaturze, znacznie częściej spotykamy alternatywne sformułowanie problemu odkrywania binarnych reguł asocjacyjnych.



Reguły asocjacyjne – miary (1)

- Binarna reguła asocjacyjna

Binarną regułą asocjacyjną (krótko, regułą asocjacyjną) nazywamy relację postaci $X \rightarrow Y$, gdzie $X \subset I$, $Y \subset I$, i $X \cap Y = \emptyset$

- Wsparcie (*support*)

Reguła $X \rightarrow Y$ posiada wsparcie sup w bazie danych D, $0 \leq \text{sup} \leq 1$, jeżeli sup% transakcji w D wspiera zbiór $X \cup Y$

- Ufność (*confidence*)

Reguła $X \rightarrow Y$ posiada ufność conf w bazie danych D, $0 \leq \text{conf} \leq 1$, jeżeli conf% transakcji w D, które wspierają zbiór X, wspierają również Y

Odkrywanie asocjacji (22)

Binarną regułą asocjacyjną nazywamy implikację postaci $X \rightarrow Y$, gdzie $X \subset I$, $Y \subset I$, i $X \cap Y = \emptyset$. Zbiór X nazywamy poprzednikiem reguły (ang. body, antecedent), natomiast zbiór Y następnikiem reguły (ang. head, consequent).



Reguły asocjacyjne – miary (2)

- ufność($X \rightarrow Y$)

oznacza stosunek liczby transakcji zawierających $X \cup Y$ do liczby transakcji zawierających Y – miara ta jest asymetryczna względem zbiorów stanowiących poprzednik i następnik reguły

- wsparcie($X \rightarrow Y$)

oznacza liczbę transakcji w bazie danych, które potwierdzają daną regułę – miara wsparcia jest symetryczna względem zbiorów stanowiących poprzednik i następnik reguły

Odkrywanie asocjacji (23)

Wsparcie jest istotną miarą wartościującą daną regułę asocjacyjną, gdyż określa liczbę transakcji w analizowanym zbiorze D , które potwierdzają daną regułę. Odwołując się do przykładu supermarketu, wsparcie reguły określa liczbę klientów, którzy zachowują się zgodnie z daną regułą. Łatwo zauważyć, że miara wsparcia jest symetryczna względem zbiorów stanowiących poprzednik i następnik reguły, to znaczy, jeżeli reguła asocjacyjna posiada w zbiorze D wsparcie s, to takie samo wsparcie w zbiorze D posiada reguła asocjacyjna. Reguły o niewielkim wsparciu są mało reprezentatywne, gdyż opisują zachowanie niewielkiej grupy klientów. Z drugiej strony, reguły o wysokim wsparciu są, najczęściej, mało interesujące dla użytkowników, gdyż ze względu na swoją powszechność są użytkownikom dobrze znane i nie wnoszą niczego nowego do ich wiedzy o świecie. Ufność reguły określa na ile odkryta reguła asocjacyjna jest "pewna". Reguły o niskiej ufności są mało wiarygodne, natomiast reguły charakteryzujące się wysoką ufnością są "prawie pewne". Miara ufności, w przeciwnieństwie do wsparcia, jest asymetryczna względem zbiorów stanowiących poprzednik i następnik reguły w tym sensie, że $\text{ufność}(X \rightarrow Y) \neq \text{ufność}(Y \rightarrow X)$.



Reguły asocjacyjne – miary (3)

- Ograniczenia miar (definiowane przez użytkownika):

Minimalne wsparcie – **minsup**

Minimalna ufność – **minconf**

- Mówimy, że reguła asocjacyjna $X \rightarrow Y$ jest **silna** jeżeli

$$\text{sup}(X \rightarrow Y) \geq \text{minsup} \text{ i } \text{conf}(X \rightarrow Y) \geq \text{minconf}$$

- Dana jest baza danych transakcji
Należy znaleźć wszystkie silne binarne reguły asocjacyjne

Odkrywanie asocjacji (24)

W przypadku analizy dużych wolumenów danych liczba znajdowanych (odkrywanych) reguł asocjacyjnych jest być bardzo dużą i, oczywiście, nie wszystkie znalezione reguły są równie interesujące i ważne z punktu widzenia użytkownika systemu eksploracji danych. Najczęściej, reguła asocjacyjna jest uważana za interesującą i ważną, jeżeli wartości wsparcia i ufności reguły przekraczają pewne zadane wartości progowe: minimalny próg wsparcia (ang. minimum support threshold) i minimalny próg ufności (ang. minimum confidence threshold). Wartości tych progów są definiowane przez użytkownika, lub eksperta dziedzinowego, i stanowią parametry wejściowe procesu odkrywania reguł asocjacyjnych dla danego zbioru analizowanych danych. Mówimy, że reguła jest silna, jeżeli wartość współczynnika wsparcia ($\text{sup}(X \rightarrow Y)$) jest większa lub równa minimalnemu progowi wsparcia (minsup) oraz wartość ufności reguły ($\text{conf}(X \rightarrow Y)$) jest większa lub równa minimalnemu progowi ufności. Celem odkrywania reguł asocjacyjnych jest znalezienie wszystkich silnych binarnych reguł asocjacyjnych.



Przykład

Trans_Id	Produkty
100	A, B, C
200	A, C
300	A, D
400	B, E, F

Zakładając
minsup = 50% oraz minconf = 50%
w przedstawionej bazie danych można znaleźć
następujące reguły asocjacyjne:

$A \rightarrow C$ sup = 50%, conf = 66,6 %
 $C \rightarrow A$ sup = 50%, conf = 100%

Odkrywanie asocjacji (25)

Rozważmy powyższy przykład. Mamy dana tabelę informacji zawierającą identyfikator transakcji oraz produkty jakie w ramach tej transakcji zostały zakupione. Przyjmujemy, że interesują nas tylko reguły z 50% wsparciem minsup = 50% i jednocześnie z co najmniej 50% ufnością minconf=50%. W przedstawionej bazie danych możemy znaleźć następujące reguły asocjacyjne:

$A \rightarrow C$ dla której możemy obliczyć sup = 50% (liczba transakcji, które potwierdzają daną regułę czyli 2/4), conf=66,6% (stosunek liczby transakcji zawierających $A \rightarrow C$ do liczby transakcji zawierających A czyli 2/3), analogicznie obliczamy dla reguły $C \rightarrow A$ sup = 50% oraz conf=100%.



Inne miary oceny reguł asocjacyjnych

Conviction:

$$\text{conviction } (A \rightarrow C) = \frac{(|D| - \text{sup}(C))}{(|D| (1 - \text{conf}(A \rightarrow C)))}$$

Lift:

$$\text{lift } (A \rightarrow C) = \frac{(|D| \text{conf} (A \rightarrow C))}{(\text{sup} (A \rightarrow C))}$$

Interest:

$$\text{interest } (A \rightarrow C) = \frac{\text{sup} (A, C)}{(\text{sup} (A) * \text{sup} (C))}$$

Odkrywanie asocjacji (26)

Istnieją jeszcze inne miary określające ocenę otrzymanych reguł asocjacyjnych, np.: Coviction, Lift oraz Interest.



Algorytm naiwny

1. Dany jest zbiór elementów I i baza danych D
2. Wygeneruj wszystkie możliwe podzbiory zbioru I i następnie, dla każdego podzbioru oblicz wsparcie tego zbioru w bazie danych D
3. Dla każdego zbioru, którego wsparcie jest większe/równe minsup , wygeneruj regułę asocjacyjną – dla każdej otrzymanej reguły oblicz ufność reguły

Liczba wszystkich możliwych podzbiorów zbioru I wynosi $2^{|I|} - 1$ (rozmiar $I \approx 200\ 000$ elementów)

Odkrywanie asocjacji (27)

Naiwny algorytm odkrywania zbiorów częstych składa się z trzech głównych kroków.

- (1) Dany jest zbiór elementów I i baza danych D .
- (2) Wygeneruj wszystkie możliwe podzbiory zbioru I i następnie, dla każdego podzbioru oblicz wsparcie tego zbioru w bazie danych D .
- (3) Dla każdego zbioru, którego wsparcie jest większe/równe minsup , wygeneruj regułę asocjacyjną – dla każdej otrzymanej reguły oblicz ufność reguły.

Trywialna metoda znajdowania zbiorów częstych mogłaby polegać na wygenerowaniu wszystkich podzbiorów zbioru I i, następnie, na obliczeniu, dla każdego podzbioru L wartości wsparcia tego podzbioru. W praktyce takie rozwiązanie jest nieakceptowne ze względu na ilość potencjalnych zbiorów częstych. Liczba wszystkich możliwych podzbiorów zbioru I wynosi $2^{|I|} - 1$, gdzie $|I|$, w zastosowaniach praktycznych (np. analizie koszyka), może sięgać nawet 300 000 elementów.

Ogólny algorytm odkrywania reguł asocjacyjnych (1)

- **Algorytm 1.1:** Ogólny algorytm odkrywania reguł asocjacyjnych

- Znajdź wszystkie zbiory elementów $L_i = \{i_{i1}, i_{i2}, \dots, i_{im}\}$, $L_i \subseteq I$, których **wsparcie(L_i) $\geq \text{minsup}$**
Zbiory L_i nazywać będziemy **zbiorami częstymi**
- Korzystając z Algorytmu 1.2 i znalezionej kolekcji zbiorów częstych wygeneruj wszystkie reguły asocjacyjne

Odkrywanie asocjacji (28)

Pierwszy algorytm odkrywania silnych binarnych reguł asocjacyjnych przedstawiono w roku 1993. W tym samym roku, przedstawiono też algorytm SETM, który w procesie odkrywania silnych binarnych reguł asocjacyjnych wykorzystuje operatory relacyjne. W roku 1994 pojawiła się fundamentalna praca Agrawala i Srikanta, w której przedstawiono dwa nowe algorytmy odkrywania silnych binarnych reguł asocjacyjnych: Apriori i AprioriTID. Algorytmy te stały się, w późniejszym czasie, podstawą wielu nowych algorytmów odkrywania binarnych reguł asocjacyjnych. Cechą wspólną wszystkich algorytmów odkrywania silnych binarnych reguł asocjacyjnych jest identyczny ogólny schemat działania algorytmu. Schemat ten, został umieszczony na slajdzie. Schemat został nazwany "ogólnym algorytmem odkrywania silnych binarnych reguł asocjacyjnych".

Algorytm 1.1 składa się z dwóch kroków. W pierwszym kroku znajdują się wszystkie zbiory częste, które reprezentują zbiory elementów występujących wspólnie w transakcjach. W kroku drugim, na podstawie znalezionych zbiorów częstych, generowane są wszystkie silne

binarne reguły asocjacyjne, których ufność jest nie mniejsza niż zadany próg minimalnej ufności minconf.



Ogólny algorytm odkrywania reguł asocjacyjnych (2)

- **Algorytm 1.1:** Ogólny algorytm odkrywania reguł asocjacyjnych

```
for each zbioru częstego  $L_i$  do
    for each podzbioru  $subL_i$  zbioru  $L_i$  do
        if wsparcie( $L_i$ ) / wsparcie( $subL_i$ ) ≥ minconf
        then
            output reguła  $subL_i \rightarrow (L_i - subL_i)$ 
            conf( $subL_i \rightarrow (L_i - subL_i)$ ) =
            support( $L_i$ ) / support( $subL_i$ ),
            sup( $subL_i \rightarrow (L_i - subL_i)$ ) = support( $L_i$ )
```

Odkrywanie asocjacji (29)

Kluczowe znaczenie, z punktu widzenia efektywności algorytmu odkrywania silnych binarnych reguł asocjacyjnych, ma pierwszy krok algorytmu - znajdowanie zbiorów częstych, to jest, podzbiorów zbioru I , których wsparcie jest większe lub równe minimalnej wartości wsparcia $minsup$. W ostatnim kroku algorytmu wygenerowane reguły są poddawane analizie. W zbiorze wynikowym pozostaną tylko te reguły, których współczynnik ufności będzie co najmniej tak dobry jak minimalny próg wsparcia. W ten sposób otrzymujemy tylko silne reguły asocjacyjne.

Jak już wspomnieliśmy wcześniej, problem odkrywania binarnych reguł asocjacyjnych był bardzo intensywnie analizowany od roku 1993 i zaproponowano szereg algorytmów odkrywania binarnych reguł asocjacyjnych różniących się, głównie, dwoma elementami: metodą odkrywania zbiorów częstych oraz metodą obliczania wsparcia zbiorów elementów.