UNIWERSYTET PRZYRODNICZY WE WROCŁAWIU

Data Mining
Wykład 2

Odkrywanie asocjacji

Plan wykładu

- Wprowadzenie
- Sformułowanie problemu
- Typy reguł asocjacyjnych
- Proces 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

Analiza koszyka zakupów

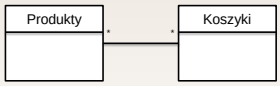
- Cel analizy MBA:
  - znalezienie naturalnych wzorców zachowań konsumenckich klientów
- Wykorzystanie wzorców zachowań:
  - organizacji półek w supermarkecie
  - opracowania akcji promocyjnych
  - opracowania katalogu oferowanych produktów

Zastosowanie MBA

- 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 przeciwległych końcach sklepu, zmuszając klientów do przejścia przez cały sklep

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

Tablica obserwacji (1)

- Dany jest zbiór atrybutów A= {A_1 , A_2 , ..., A_n } oraz zbiór obserwacji T = {T_1 , T_2 , ..., 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	0	0	1	1	0
T_4	0	1	0	0	0
T_5	1	0	0	1	0
T_6	0	0	1	0	0
T_7	1	1	1	0	0
T_8	1	1	0	0	1

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

Tablica obserwacji - przykłady

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

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)

Reguły asocjacyjne (1)

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

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

Interpretacja reguły:

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

Reguły asocjacyjne (2)

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

$$\theta \rightarrow \varphi: (A_{i1}, A_{i2}, \ldots, A_{ik}) \rightarrow (A_{ik+1}, A_{ik+2}, \ldots, A_{ik+l})$$

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

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

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

Reguły asocjacyjne (3)

- Statystyczna ważność i siła reguły:

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

Ufnością (conf) reguły asocjacyjnej $\theta \rightarrow \varphi$ nazywać będziemy s**tosunek liczby obserwacji, które spełniają warunek $\theta \wedge \varphi$, do liczby obserwacji, które spełniają warunek θ** (ufność reguły = warunkowemu prawdopodobieństwu $p(\varphi \mid \theta)$)

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

Typ przetwarzanych danych (1)

- Wyróżniamy:

binarne reguły asocjacyjne - regułę asocjacyjną nazywamy binarną, jeżeli dane występujące w regule są danymi (zmiennymi) binarnymi

ilościowe reguły asocjacyjne - regułę asocjacyjną nazywamy ilościową, jeżeli dane występujące w regule są danymi ciągłymi i/lub kategorycznymi

Typ przetwarzanych danych (2)

- Binarna reguła asocjacyjna:

pieluszki = 1→ piwo =1

– reprezentuje współwystępowanie danych

- Ilościowa reguła asocjacyjna:

wiek = '30...40' ∧ wykształcenie = 'wyższe' → opcja_polityczna ='demokrata'

- reprezentuje współwystępowanie wartości danych

Wymiarowość przetwarzanych danych (1)

- Wyróżniamy:

jednowymiarowe reguły asocjacyjne - regułę asocjacyjną nazywamy jednowymiarową, jeżeli dane występujące w regule reprezentują tę samą dziedzinę wartości.

wielowymiarowe reguły asocjacyjne - regułę asocjacyjną nazywamy wielowymiarową, jeżeli dane występujące w regule reprezentują różne dziedziny wartości.

Typ przetwarzanych danych (2)

- Jednowymiarowa reguła asocjacyjna:

pieluszki = 1→ piwo =1

- Wielowymiarowa reguła asocjacyjna:

wiek = '30...40' ∧ wykształcenie = 'wyższe' → opcja_polityczna ='demokrata'

Stopień abstrakcji przetwarzanych danych (1)

- Wyróżniamy:

jednopoziomowe reguły asocjacyjne - regułę asocjacyjną nazywamy jednopoziomową, jeżeli dane występujące w regule reprezentują ten sam poziom abstrakcji.

Wielopoziomowe reguły asocjacyjne - regułę asocjacyjną nazywamy wielopoziomową, jeżeli dane występujące w regule reprezentują różne poziomy abstrakcji.

Stopień abstrakcji przetwarzanych danych (2)

- Jednopoziomowa reguła asocjacyjna:

  pieluszki_Pampers = 1→ piwo_Zywiec =1

- Wielopoziomowa reguła asocjacyjna:

  pieluszki_Pampers = 1 ∧ piwo_Zywiec =1 → napoje = 1

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

Odkrywanie binarnych reguł asocjacyjnych

- Dane:
  - $I=\{i_1, i_2, ..., i_n\}$: 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, $X \subseteq T$

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, $0 \leq sup \leq 1$, jeżeli sup% transakcji w D wspiera zbiór $X \cup Y$

- Ufność (confidence):

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

Reguły asocjacyjne – miary (2)

- wsparcie(X → Y):

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

- ufność(X → Y):

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

Reguły asocjacyjne – miary (3)

- Ograniczenia miar (definiowane przez użytkownika):

  Minimalne wsparcie – minsup

  Minimalna ufność – minconf

- Mówimy, że reguła asocjacyjna X → Y jest silna jeżeli:

  sup(X → Y) ≥ minsup i conf(X → Y) ≥ minconf

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

Przykład

TR_{ID}	Produkty
1	A,B,C
2	A,C
3	A,D
4	B,E,F

Zakładając:

minsup = 50% oraz minconf = 50%

w przedstawionej bazie danych można znaleźć następujące reguły asocjacyjne:

A → C sup = 50%, conf = 66,6 %
C → A sup = 50%, conf = 100%

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 ≈ 200 000 elementów)

Ogólny algorytm odkrywania reguł asocjacyjnych (1)

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

1. Znajdź wszystkie zbiory elementów Li={ii1, ii2, ..., iim}, Li⊆ I, których **wsparcie(Li) ≥ minsup**.
Zbiory Li nazywać będziemy zbiorami częstymi

2. Korzystając z Algorytmu 1.2 i znalezionej kolekcji zbiorów częstych wygeneruj wszystkie reguły asocjacyjne

Ogólny algorytm odkrywania reguł asocjacyjnych (2)

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

for each zbioru częstego Li **do**
  for each podzbioru subLi zbioru Li **do**
    if wsparcie(Li)/**wsparcie**(subLi)≥minconf
    then
      output reguła subLi → (Li-subLi)
      conf(subLi → (Li-subLi)) =
      support(Li)/**support**(subLi),
      sup(subLi → (Li-subLi)) = **support**(Li)

Podsumowanie

- Typy reguł asocjacyjnych
  - Typ przetwarzanych danych
  - Wymiarowość przetwarzanych danych
  - Stopień abstrakcji przetwarzanych danych
- Proces odkrywania reguł asocjacyjnych
  - Wsparcie - $sup(\theta \rightarrow \varphi)$
  - Ufność - $conf(\theta \rightarrow \varphi)$