

Integracja technik eksploracji danych z systemem zarządzania bazą danych na przykładzie Oracle9i Data Mining

Mikołaj Morzy
Marek Wojciechowski
Instytut Informatyki PP



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Eksploracja danych

- Odkrywanie wzorców w dużych wolumenach danych
- Ewolucja systemów eksploracyjnych
 - Systemy dedykowane
 - Systemy współpracujące z bazą danych (Oracle Darwin, IBM Intelligent miner)
 - Systemy ściśle zintegrowane z bazą danych (oracle9i data Mining)



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

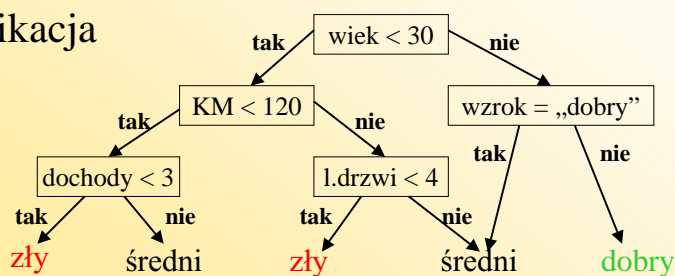
Metody eksploracji danych (1)

- Reguły asocjacyjne

- 80% klientów którzy w marcu kupili buty narciarskie i okulary słoneczne kupiło też wełniane swetry, takich zakupów dokonało 0.8% klientów kupujących w marcu

Buty narciarskie \wedge *okulary* \rightarrow *sweter wełniany* $s=0.8$ $c=80$

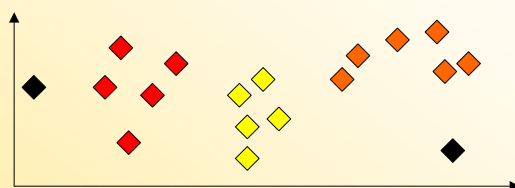
- Klasyfikacja



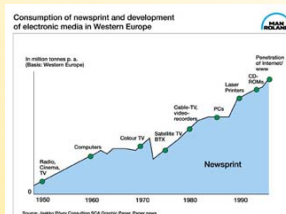
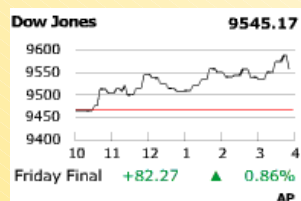
V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Metody eksploracji danych (2)

- Grupowanie obiektów (clustering)



- Przebiegi czasowe (time series)



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Klasyfikacja

- Zbiór przykładów (krotek), z których każdy należy do jednej z predefiniowanych klas
- Budowanie (trenowanie) modelu i testowanie modelu
- Wykorzystanie modelu do określania klasy do której należą nowe przykłady
- Klasyfikacja (atrybuty kategoriyczne) i predykcja (atrybuty ciągłe)



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Metody klasyfikacji

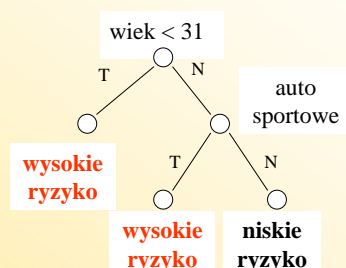
- Drzewa decyzyjne
- Klasyfikatory bayesowskie
- Sieci neuronowe
- Analiza statystyczna
- Algorytmy genetyczne
- Zbiory przybliżone



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Drzewa decyzyjne

- Każdy węzeł wewnętrzny reprezentuje test przeprowadzony na atrybucie
- Każda gałąź reprezentuje wynik testu
- Każdy liść reprezentuje klasę
- Kryteria podziału
 - Indeks GINI (CART)
 - Wzrost informacji (C4.5)
 - χ^2 (CHAID)



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Klasyfikator Bayesa

- Twierdzenie Thomasa Bayesa (1702-1761)

$$\Pr(h|d) = \Pr(d|h) * \Pr(h) * \{\sum_i \Pr(d|h_i) * \Pr(h_i)\}^{-1}$$

- Optymalny klasyfikator Bayesa

$$\arg \max \Pr(c(x)=d|t) = \sum_h \Pr(c(x)=d|h) * \Pr(h|t)$$

- Naiwny klasyfikator Bayesa

$\arg \max \Pr(c(x)=d | a_1(x)=a_1(x_0), \dots, a_n(x)=a_n(x_0))$, czyli

$$\arg \max \Pr(c(x)=d) * \Pr(a_1(x)=a_1(x_0), \dots, a_n(x)=a_n(x_0) | c(x)=d)$$

- Założenie o warunkowej niezależności atrybutów:

$$\Pr(a_1(x)=v_1, \dots, a_n(x)=v_n | c(x)=d) = \prod_i \Pr(a_i(x)=v_i | c(x)=d)$$



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Przykład

RID	DOCHOD	TYP	RYZYKO
1	> 2000	SPORT	NISKIE
2	<= 2000	SPORT	WYSOKIE
3	<= 2000	COMBI	NISKIE
4	> 2000	COUPE	WYSOKIE
5	> 2000	COMBI	NISKIE
6	<= 2000	SPORT	WYSOKIE
7	<= 2000	COUPE	???

$Pr(\text{niskie})=3/6$ $Pr(\text{wysokie})=3/6$
 $Pr(\text{dochod} > 2000 | \text{niskie}) = 2/3$
 $Pr(\text{dochod} \leq 2000 | \text{niskie}) = 1/3$
 $Pr(\text{dochod} > 2000 | \text{wysokie}) = 1/3$
 $Pr(\text{dochod} \leq 2000 | \text{wysokie}) = 2/3$
 $Pr(\text{typ} = \text{sport} | \text{niskie}) = 1/3$
 $Pr(\text{typ} = \text{combi} | \text{niskie}) = 2/3$
 $Pr(\text{typ} = \text{coupe} | \text{niskie}) = 0/3$
 $Pr(\text{typ} = \text{sport} | \text{wysokie}) = 2/3$
 $Pr(\text{typ} = \text{combi} | \text{wysokie}) = 0/3$
 $Pr(\text{typ} = \text{coupe} | \text{wysokie}) = 1/3$

$Pr(\text{niskie}) * Pr(>2000, \text{coupe} | \text{niskie}) = 3/6 * 2/3 * 0.1/3 = 1/90$

$Pr(\text{wysokie}) * Pr(>2000, \text{coupe} | \text{wysokie}) = 3/6 * 1/3 * 1/3 = 1/18$



V Seminarium PLOUG

Projektowanie i implementowanie magazynów danych

Reguły asocjacyjne

- Zbiór transakcji klientów gdzie każda transakcja to zbiór elementów (produktów)
- Odnalezienie zbiorów elementów często występujących razem w transakcjach klientów
- Wygenerowanie reguł i obliczenie współczynników statystycznych opisujących współwystępowanie elementów



V Seminarium PLOUG

Projektowanie i implementowanie magazynów danych

Zastosowanie reguł asocjacyjnych

- Analiza koszyka zakupów
- Rozkład półek i towarów na półkach
- Konstruowanie wiązanych ofert sprzedaży
- Marketing bezpośredni
- Diagnozy lekarskie
- Telekomunikacja
- Analiza dostępów do serwisów WWW
- Automatyczna personalizacja serwisów WWW



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Sformułowanie problemu

- Zbiór elementów $I = \{ i_1, \dots, i_n \}$
- Transakcja T (zbiór elementów) $T \subseteq I$
- Baza danych D (zbiór transakcji)
- Transakcja T wspiera zbiór elementów X jeśli $X \subseteq T$
- Reguła asocjacyjna:
 $X \rightarrow Y$, gdzie $X, Y \subseteq I$ i $X \cap Y = \emptyset$
„piwo” \wedge „czipsy” \wedge „karkówka” \rightarrow „plast. Talerze”



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Miary stosowane do reguł

- **Wsparcie** reguły $X \rightarrow Y$ to liczba transakcji w D wspierających $(X \cup Y)$
- **Ufność** reguły $X \rightarrow Y$ to liczba transakcji wspierających X które również wspierają Y
- **Lift** reguły $X \rightarrow Y$ porównuje stosunek prawdopodobieństwa wystąpienia Y razem z X do prawdopodobieństwa wystąpienia Y z dowolnym innym zbiorem



Algorytm Apriori

```
L1 = {frequent 1-itemset};  
for (k=2; Lk-1 ≠ ∅; k++) do  
begin  
  Ck = apriori_gen(Lk-1);  
  forall transactions t ∈ T do  
    begin  
      forall candidates c ⊆ t do  
        c.count++;  
      end;  
  Lk = {c ∈ Ck | c.count ≥ minsup}  
end;  
Answer = ∪k Lk;
```



Inne rodzaje reguł asocjacyjnych

- Uogólnione reguły asocjacyjne (reguły wielopoziomowe)
- Ilościowe reguły asocjacyjne
- Wzorce sekwencji
- Reguły inter-transakcyjne



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Architektura Oracle9i Data Mining

- ODM Application Programming Interface
 - ODM API to zbiór klas i metod wykorzystywanych przez programistę
- ODM Data Mining Server
 - ODM DMS to komponent po stronie serwera, zbiór skompilowanych klas i procedur PL/SQL oraz repozytorium



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

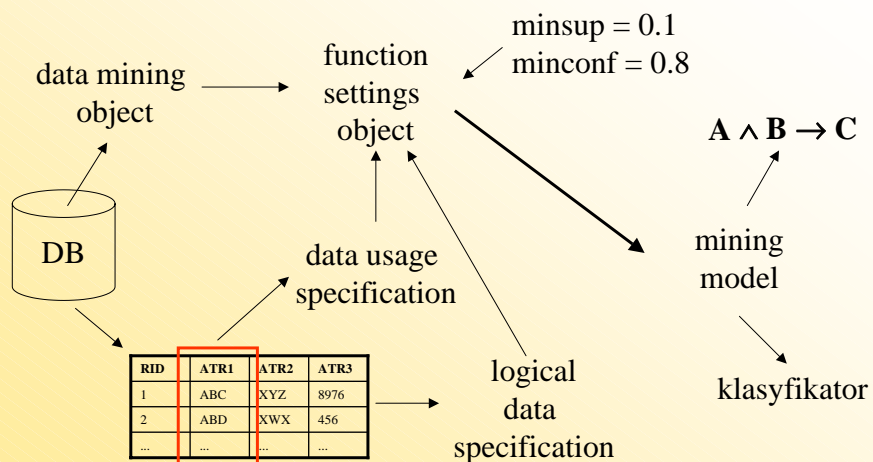
Oracle9i Data Mining – dostępne metody

- **Uczenie nadzorowane**
 - Klasyfikacja za pomocą naiwnego klasyfikatora Bayesa, budowanie modelu, testowanie modelu, stosowanie modelu do nowych danych
- **Uczenie bez nadzoru**
 - Odkrywanie reguł asocjacyjnych za pomocą algorytmu Apriori
- **Przechowywanie wyników eksploracji w repozytorium**



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Oracle9i Data Mining - proces



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Format danych

- Fizyczna specyfikacja danych
klasa: *PhysicalDataSpecification*

– format transakcyjny

SEQ_ID	ATRYBUT	WARTOŚĆ
1	KOLOR	BIAŁY
1	MARKA	FIAT
1	ROCZNIK	1998
2	KOLOR	GRANAT

– format kategoriyczny

SEQ_ID	KOLOR	MARKA	ROCZNIK
1	BIAŁY	FIAT	1998
2	GRANAT	RENAULT	2001
3	CZARNY	LANCIA	1999
4	ZIELONY	AUDI	1996



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Dyskretyzacja

- klasy:
CategoricalDiscretization,
NumericalDiscretization
- Dyskretyzacja jawna
 - Reguły mapowania, dolne i górne granice kat.
- N najczęstszych
 - Liczba interesujących kategorii
- Podział na kwantyle
 - Liczba interesujących kwantyli



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Inne klasy

- Specyfikacja funkcji eksploracji
- Model eksploracji
- Wynik eksploracji
- Algorytm eksploracji
- Reguła asocjacyjna
- Klasyfikator



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych

Repozytorium

- Zbiór relacji przechowujących funkcje, modele i wyniki eksploracji
- ODM_CONFIGURATION
- ODM_MINING_FUNCTION_SETTINGS
- ODM_MINING_MODEL
- ODM_MESSAGE_LOG
- ...



V Seminarium PLOUG
Projektowanie i implementowanie magazynów danych