Data Mining
Wykład 1

Wprowadzenie do Eksploracji Danych

Wprowadzenie

- Organizacja przedmiotu
  - 10 wykładów (2h)
  - Listy ćwiczeniowe (2h)
- Zasady zaliczenia przedmiotu:
  - 3 x Sprawozdanie Lab 3, Lab 6 i Lab7 w terminie do dwóch tygodni od zajęć
  - 1 x Projekt z klasyfikacji danych – termin 31/01/2020

Zagadnienia

1. Eksploracja Danych i Uczenie Maszynowe **x 1**
2. Reguły Asocjacyjne – Istota Asocjacji w Danych **x 2**
3. Wzorce sekwencji **x 2**
4. Klasyfikacja danych **x 3**
5. Analiza skupień (klasteryzacja) danych **x 2**

Plan wykładu

- Wprowadzenie do eksploracji danych
- Czym jest eksploracja danych?
- Proces odkrywania wiedzy
- Co można eksplorować
- Metody eksploracji danych
- Problemy odkrywania wiedzy
- Dziedziny zastosowań

„Toniemy” w danych…

Źródła danych

- Rozwój technologii baz danych, hurtowni danych oraz automatycznych narzędzi do gromadzenia danych;
- Upowszechnienie systemów informatycznych w szczególności mobilnych;
- banki, ubezpieczalnie, firmy, sieci handlowe, szpitale;
- Elektroniczna Dokumentacja Medyczna (EDM);
- dane eksperymentalne: fizyka, astronomia, biologia, genetyka;
- web, tekst, i e-handel, itd.

Największe bazy danych świata (2010 r.)

- World Data Centre for Climate ≈ 220 TB danych online + dodatkowe 6 PB offline
- National Energy Research Scientific Computing Center ≈ 2.8 PB danych
- AT&T (dane telekomunikacyjne) ≈ 323 TB danych + 1,9 tryliona rozmów telefonicznych 10^{12}
- Google ≈ 91 milionów zapytań dziennie + 33 tryliona rekordów
- Youtube ≈ 45 TB danych + 100 milionów filmów oglądanych dziennie + 65,000 filmów dodawanych dziennie
- Amazon ≈ 42TB danych + 59 milionów użytkowników
- …..

Czym jest eksploracja danych?

Eksploracja danych:

– proces automatycznego odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców schematów, podobieństwa lub trendów w dużych repozytoriach danych (bazach danych, hurtowniach danych, itp.)

Cel eksploracji danych:

– analiza danych i procesów w celu lepszego ich rozumienia

Typy zapytań do repozytoriów danych

- Eksploracja danych = złożone zapytania
- Zapytanie operacyjne do bazy danych:

  Np.: *Ile butelek wina sprzedano w I kwartale 2006 w sklepie w Poznaniu?*
- Zapytanie analityczne do hurtowni danych:

  Np.: *Ile sprzedano butelek wina w sieci Auchan na terenie kraju z podziałem na województwa, gatunki win oraz kwartały, w ciagu ostatnich 5 lat?*

Zapytania eksploracyjne (1)

Przykłady zapytań eksploracyjnych:

- Jakie inne jeszcze produkty, najczęściej, kupują klienci, którzy kupują wino?
- Czym różnią się koszyki klientów kupujących wino i piwo?
- W jaki sposób można scharakteryzować klientów kupujących wino?
- W jaki sposób pogrupować klientów kupujących wino?
- Czy można dokonać predykcji, że dany klient kupi wino?

Zapytania eksploracyjne (2)

Dany jest zbiór danych opisujących pacjentów szpitala.

Czy potrafimy w oparciu o ten zbiór danych:

- Poprawnie zdiagnozować pacjenta (określić chorobę)?
- Przewidzieć poprawnie wynik terapii?
- Zaproponować najlepsza terapie?

Czym jest eksploracja danych? (1)

- Alternatywne określenia technologii eksploracji danych:
  - odkrywanie wiedzy w bazach danych
    ang. Knowledge Discovery in Databases (KDD),
  - ekstrakcja wiedzy
    ang. Knowledge Extraction,
  - intelligencja binzesowa
    ang. Business Intelligence (BI),
  - pozyskiwanie wiedzy
    ang. Knowledge Retrieval

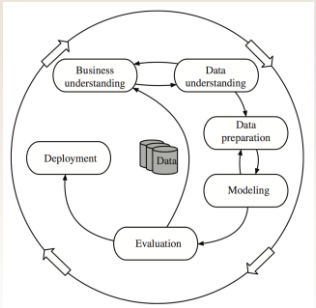
Czym jest eksploracja danych? (2)

- „Ciekawe” określenia:
  - archeologia danych,
  - kopanie w danych,
  - eksploatacja złóż danych
- Czym nie jest eksploracja danych:
  - Systemy eksperckie
  - OLAP

Czym jest eksploracja danych? (3)

Eksploracja danych (*ang. Data Mining*): zbiór technik automatycznego odkrywania nietrywialnych zależnosci, schematów, wzorców, reguł (*ang. patterns*) w dużych zbiorach danych (bazach danych, hurtowniach danych)

Proces odkrywania wiedzy (1)



Proces odkrywania wiedzy (2)

- Odkrywanie wiedzy a eksploracja danych
  - Eksploracja danych stanowi jeden z etapów procesu odkrywania wiedzy
- Etapy procesu odkrywania wiedzy (*ang. KDD process*):
  1. Zapoznanie się z wiedza dziedzinowa aplikacji - aktualna wiedza i cele aplikacji
  2. Integracja danych
  3. Selekcja danych

Proces odkrywania wiedzy (2)

1. Czyszczenie danych: (około 60% czasu)
2. Konsolidacja i transformacja danych
3. Wybór metody (metod) eksploracji danych
4. Wybór algorytmów eksploracji danych
5. Eksploracja danych
6. Interpretacja, analiza i ocena wyników wizualizacja, transformacja, usuwanie redundantnych wzorców, itd.
7. Wykorzystanie pozyskanej wiedzy

Mieszanka wielu dyscyplin

- Systemy baz danych, hurtownie danych, OLAP
- Statystyka
- Uczenie maszynowe i odkrywanie wiedzy
- Techniki wizualizacji danych
- Teoria informacji
- Wyszukiwanie informacji
- Inne dyscypliny:
  - Sieci neuronowe, modelowanie matematyczne, rozpoznawanie obrazów, technologie internetowe, itd..

Co można eksplorować?

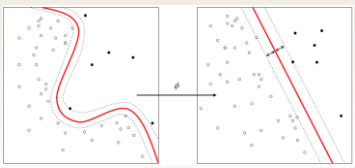
- Relacyjne bazy danych
- Hurtownie danych
- Repozytoria danych
- Zaawansowane systemy informatyczne
  - Obiektowe i obiektowo-relacyjne bazy danych
  - Przestrzenne bazy danych
  - Przebiegi czasowe i temporalne bazy danych
  - Tekstowe i multimedialne bazy danych
  - WWW
  - itd.

Metody eksploracji danych

- klasyfikacja/regresja
- grupowanie
- odkrywanie sekwencji
- odkrywanie charakterystyk
- analiza przebiegów czasowych
- odkrywanie asocjacji
- wykrywanie zmian i odchyleń
- eksploracja WWW
- eksploracja tekstów

Metody eksploracji: klasyfikacja

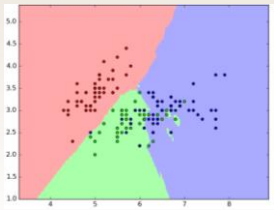
Metoda analizy danych, której celem jest predykcja wartości określonego atrybutu w oparciu o pewien zbiór danych treningowych



Wiele technik:
statystyka,
drzewa decyzyjne,
sieci neuronowe,
...

Metody eksploracji: grupowanie

Znajdź „naturalne” pogrupowanie obiektów w oparciu o ich wartości



zastosowania grupowania:
- grupowanie dokumentów
- grupowanie klientów
- segmentacja rynku

Metody eksploracji: odkrywanie asocjacji

znajdowanie związków pomiędzy występowaniem grup elementów w zbiorach danych

przykłady asocjacji:

- klienci, którzy kupują pieluszki, kupują również piwo
- klienci, którzy kupują chleb, masło i ser, kupują również wodę mineralna i ketchup

Metody eksploracji: odkrywanie wzorców sekwencji

znajdowanie najczęściej występujących sekwencji elementów

przykłady odkrywania wzorców sekwencji:

- kurs akcji BPH, który podczas ostatnich trzech sesji wzrósł o 0.5%, 0.9%, 0.1%, na następnej sesji spadnie o 0.5%
- klienci, którzy kupili farbę emulsyjna, kupią w najbliższym czasie pędzel płaski

Metody eksploracji: odkrywanie charakterystyk

znajdowanie zwięzłych opisów (charakterystyk) podanego zbioru danych

przykłady odkrywania charakterystyk:

opis pacjentów chorujących na anginę

pacjenci chorujący na anginę cechują się temperatura ciała większą ni 37.5 C, bólem gardła, osłabieniem organizmu

Problemy odkrywania wiedzy

- w dużych bazach danych mogą zostać odkryte tysiące reguł
- człowiek nie potrafi rozumieć i przeanalizować bardzo dużych zbiorów informacji
- różni użytkownicy systemu bazy danych są zainteresowani różnymi typami reguł z różnych relacji
- odkrywanie reguł jest procesem bardzo złożonym obliczeniowo

Dziedziny zastosowań

- Nauka
- Biznes
- Web
- Administracja
- Handel i Marketing
- Finanse i Bankowość
- Telekomunikacja
- Medycyna
- Inne …

Podsumowanie

- **Systemy baz danych**
  – narzędzie do przechowywania danych
- **Hurtownie danych**
  – narzędzie do wspomagania podejmowania decyzji
- **Eksploracja danych**
  – narzędzie do analizy zgromadzonych danych