Eksploracja danych

Piotr Lipiński

Informacje ogólne

- Informacje i materiały dotyczące wykładu będą publikowane na stronie internetowej wykładowcy, m.in.
  - prezentacje z wykładów
    - UWAGA: prezentacja to nie książka, notatki czy skrypt – to raczej streszczenie omawianego materiału, pokazanie wybranych algorytmów, przedstawienie wybranych przykładów – dlatego przejrzenie samych prezentacji czasami może nie wystarczyć do zrozumienia treści całego wykładu
  - listy zadań
  - propozycje minireferatów i minikonkursów
  - materiały dodatkowe rozszerzające treść wykładu
  - ogłoszenia bieżące

Zasady zaliczenia

- Ćwiczenia/pracownie:
  - Będzie można zdobywać punkty za:
    - listy zadań (za 80 punktów łącznie)
      - każda lista zadań będzie dotyczyć pewnego działu eksploracji danych
      - część zadań będzie polegać na zaimplementowaniu pewnych algorytmów i wykorzystaniu ich do analizy przykładowych danych, a część na omówieniu przy tablicy pewnych mechanizmów eksploracji danych
      - na realizację każdej listy zadań będzie określony czas, zwykle 1 lub 2 tygodnie
    - projekt (za 30 punktów)
    - sprawdzian (za 10 punktów)
    - punkty bonusowe za dodatkową aktywność (minireferaty, minikonkursy, itp.)
  - Łącznie będzie do zdobycia minimum 120 punktów (oprócz punktów bonusowych)
  - Na zaliczenie wymagane jest 60 punktów. Progi na poszczególne oceny to:

3.0	60 punktów
3.5	72 punktów
4.0	84 punktów
4.5	96 punktów
5.0	108 punktów

  - Na ocenę bardzo dobrą wymagane jest dodatkowo przygotowanie i wygłoszenie minireferatu.
- Wykład: egzamin

Zasady zaliczenia

- Kilka przykładów:
  - Student, który zdobędzie 40 punktów za listy zadań, 0 punktów ze sprawdzianu i 20 punktów za projekt, zaliczy zajęcia z oceną 3.0.
  - Student, który zdobędzie 60 punktów za listy zadań, 0 punktów ze sprawdzianu i 0 punktów za projekt, zaliczy zajęcia z oceną 3.0.
  - Student, który zdobędzie 55 punktów za listy zadań, 0 punktów ze sprawdzianu, 0 punktów za projekt i 5 punktów za minireferat, zaliczy zajęcia z oceną 3.0.
  - Student, który zdobędzie 75 punktów za listy zadań, 10 punktów ze sprawdzianu, 20 punktów za projekt i 5 punktów za minireferat, zaliczy zajęcia z oceną 5.0.
  - Student, który zdobędzie 75 punktów za listy zadań, 10 punktów ze sprawdzianu i 25 punktów za projekt, zaliczy zajęcia z oceną 4.5 (na ocenę 5.0 wymagany jest minireferat).
- UWAGA: Projekt może wymagać sporo pracy. Musi zawierać przemyślenie wybranego problemu, opracowanie algorytmu jego rozwiązywania, implementację tego algorytmu, przeprowadzenie eksperymentów obliczeniowych i wykonanie raportu z testowania opracowanego podejścia.

Program wykładu

- Niepewność danych
- Grupowanie danych
- Redukcja wymiarowości danych
- Klasyfikacja danych
- Reguły asocjacyjne
- Prognozowanie szeregów czasowych
- Systemy rekomendujące
- Systemy wspomagania decyzji
- Przetwarzanie dużych danych i danych multimedialnych
- Statystyka obliczeniowa

Eksploracja danych

- Eksploracja danych zajmuje się analizą dużych zbiorów danych w celu pozyskania z nich nietrywialnej i pożytecznej wiedzy.
- Różnica między informacją a wiedzą:
  - informacja = dane zgromadzone w bazie lub hurtowni danych
    - często bardzo dużych rozmiarów
    - zazwyczaj opisują zarejestrowane obserwacje pewnego zjawiska
    - zazwyczaj obarczone błędem pomiarowym lub innym zaburzeniem
    - często trudne do zrozumienia przez człowieka (człowiek nie potrafi zauważyć pewnych zależności w tych danych)
  - wiedza
    - model obserwowanego zjawiska lub jego części
    - często zawiera opis zależności między danymi
    - często wyjaśnia i pozwala zrozumieć zjawisko

Eksploracja danych

- Z informacji można w prosty sposób utrzymać wiedzę bezużyteczną:
  - zawsze można policzyć średnią (z atrybutów numerycznych) lub medianę (z atrybutów numerycznych lub kategorycznych)
  - zawsze można zrobić parę wykresów
  - zawsze można opracować przeuczony system klasyfikujący
  - można też pokusić się o nieuprawnione wnioski pseudomatematyczne

Eksploracja danych

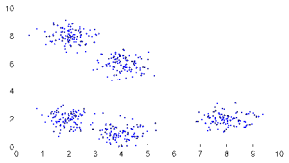
- Przykład:
  - informacja = zebrane informacje o użytkownikach kart kredytowych (klientach pewnego banku) zawierające dane osobowe i miesięczne wyciągi z kart kredytowych
    - dane są dużych rozmiarów
    - dane są nie tylko numeryczne
    - dane mogą być od siebie zależne (wydatki osób mieszkających wspólnie)
  - wiedza całkowicie bezużyteczna
    - najczęściej powtarzające się nazwisko
    - średni numer domu klientów
    - średni wiek klientów
    - średnia roczna suma wydatków klientów

Eksploracja danych

- Przykład:
  - wiedza bardziej użyteczna (m.in. uzyskana podstawowymi metodami Business Intelligence)
    - średnie roczne sumy wydatków klientów w poszczególnych przedziałach wiekowych
    - średnie roczne sumy wydatków klientów w poszczególnych rejonach geograficznych
    - średni wiek klientów w poszczególnych przedziałach rocznej sumy wydatków
    - prognozowana suma wydatków klientów w poszczególnych rejonach geograficznych w przyszłym roku
  - wiedza jeszcze bardziej użyteczna (m.in. uzyskana podstawowymi metodami eksploracji danych)
    - wyodrębnienie grup klientów zachowujących się podobnie
      - na przykład: klienci przeznaczający podobną część swoich wydatków na paliwo, odzież i podróże
    - znalezienie powiązań między wydatkami klientów
      - na przykład: duże wydatki na paliwo pociągają duże wydatki na hotele

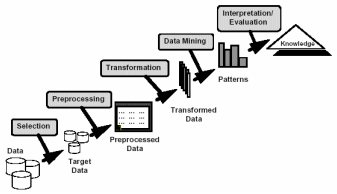
Eksploracja danych

- Przykład:
  - wiedza jeszcze bardziej użyteczna (m.in. uzyskana podstawowymi metodami eksploracji danych)



Eksploracja danych

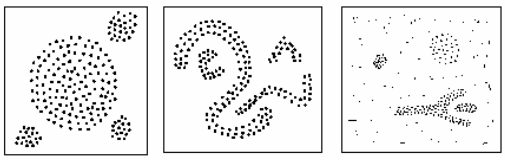
- Typowy schemat analizy danych



- Popularne narzędzia eksploracji danych:
  - Oracle Data-Mining, IBM SPSS
  - Matlab, Octave, R, Statistica
  - WEKA
  - własne algorytmy i ich implementacje

Grupowanie danych

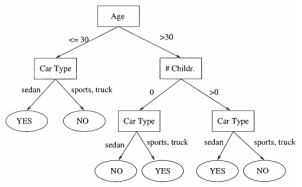
- Celem grupowania danych jest podział rekordów danych na grupy, tak aby elementy z tej samej grupy były do siebie podobne, a z różnych grup od siebie różne.
  - Zazwyczaj nie wiadomo czemu odpowiadają utworzone grupy (jak je interpretować merytorycznie).
  - Wiadomo jednak, jak je precyzyjnie zdefiniować.
  - Wiadomo też, że są statystycznie nieprzypadkowe.



Klasyfikacja danych

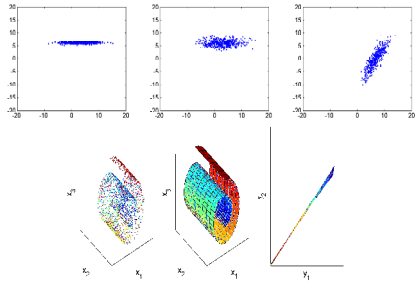
- Klasyfikator to funkcja, która przypisuje każdy rekord danych do jednej z określonych klas. Klasyfikacja polega na konstruowaniu klasyfikatora poprawnie klasyfikującego dane.
  - Potrzebne są dane uczące (poprawnie poklasyfikowane).
  - Wiadomo czemu odpowiadają utworzone klasy.

Record Id	Car	Age	Children	Subscription
1	sedan	23	0	yes
2	sports	31	1	no
3	sedan	36	1	no
4	truck	25	2	no
5	sports	30	0	no
6	sedan	36	0	no
7	sedan	25	0	yes
8	truck	36	1	no
9	sedan	30	2	yes
10	sedan	31	1	yes
11	sports	25	0	no
12	sedan	45	1	yes
13	sports	23	2	no
14	truck	45	0	yes



Eksploracja danych – redukcja wymiarowości

- Czasami dane wyglądają na bardziej skomplikowane niż są w rzeczywistości.



Eksploracja danych – systemy rekomendujące

- Systemy rekomendujące służą do rekomendowania użytkownikom produktów (najczęściej w celach komercyjnych).
  - Dla każdego użytkownika tworzony jest jego profil (charakterystyka).
  - Użytkownicy o podobnym profilu są łączeni w grupy.
  - Każdemu użytkownikowi są rekomendowane produkty wysoko oceniane przez innych użytkowników z jego grupy.
- Problemy:
  - co powinien uwzględniać profil użytkownika?
    - selekcja atrybutów, redukcja wymiarowości, analiza korelacji
    - przetwarzania danych rzadkich (ang. sparse data)
  - jak pogrupować użytkowników?
  - które z wielu wybranych produktów wyświetlić i w jakiej kolejności?