Laboratorium z przedmiotu Data mining

Lab1.a – wprowadzenie, wstępne przekształcanie danych

- Grzegorz Protaziuk

  email: G.Protaziuk@ii.pw.edu.pl

  konsultacje: czwartek, 13.00-14.00, pok. 301

Regulamin

- Laboratorium jest na ocenę.
  - Skala ocen 2,0 – 5,0.
  - Ocena jest wystawiana na podstawie punktów uzyskanych z dwóch równorzędnych zadań: zdania dotyczącego reguł i zadania dotyczącego klasyfikacji.
- Oceny

ocena	liczba punktów
5	> 90%
4,5	> 80%
4	> 70%
3,5	> 60%
3	> 50%

Tematyka laboratorium

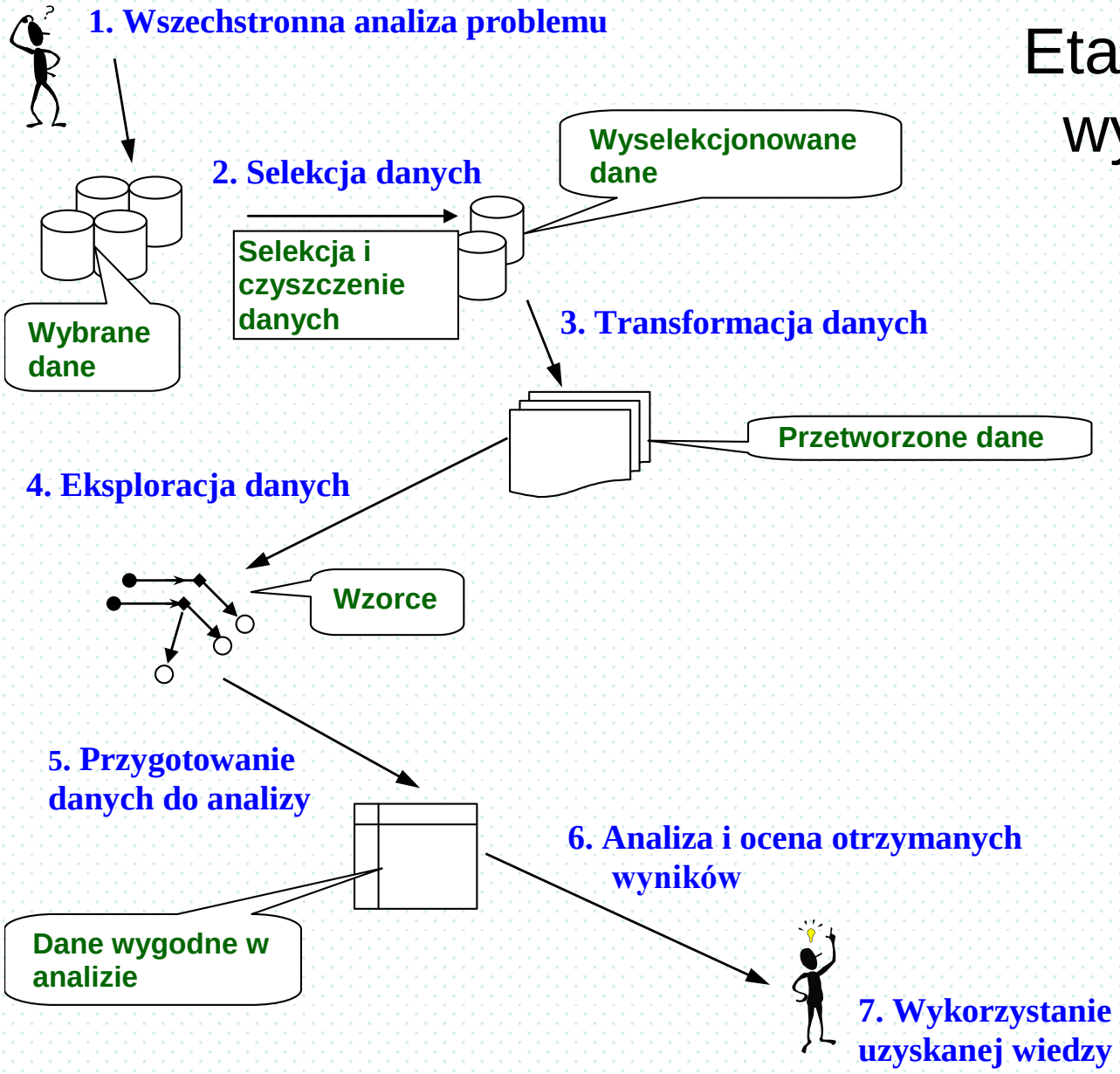
- Lab1 – wprowadzenie, reguły asocjacyjne
- Lab2 – sekwencje
- Lab3 – klasyfikacja
- Lab4 – grupowanie

Środowisko

R-Studio

- R jest oprogramowaniem na licencji typu open-source, oferującym różnorodne metody analizy danych (statystyczne, algorytmy eksploracji danych i danych tekstowych) oraz metody do tworzenia prezentacji graficznych wyników.
- strona www: https://www.r-project.org/

Etapy procesu wykrywania wiedzy



Metody eksploracji danych

- Reguły asocjacyjne
  - Apriori, ECLAT – pakiet arules
- Reguły sekwencyjne
  - cSpade – pakiet arulesSequence
- Grupowanie
  - Metody partycjonujące - k-means, pam
  - Metody hierarchniczne - agnes
  - Medoty gęstościowe – dbscan
  - Pakiety: cluster, fpc
- Klasyfikacja
  - Drzewa decyzyjne - pakiety: party, rpart , c50
  - Lasy losowe – pakiet randomForest
  - Svm, Naive Bayes – pakiet e1071

Transformacja danych

- Postać relacyjna i transakcyjna w obszarze eksploracji danych
- Dyskretyzacja danych
- Problem niekompletności danych

Postać relacyjne i transakcyjna

- **Postać relacyjna** – tabela, znany, dobrze zdefiniowany zbiór atrybutów.
- **Postać transakcyjna** – dwa atrybuty: identyfikator obiektu, cecha obiektu.

Jest możliwość prostej zamiany danych z jednej postaci w drugą (dla danych kompletnych).

Transformacje danych (2)

Dyskretyzacja atrybutów ciągłych

- przez eksperta
- na dwie wartości na podstawie średniej
- na 4 wartości na podstawie średniej i odchylenia standardowego
- **Na n przedziałów o równym rozmiarze** – najczęściej spotykana
- Na n równolicznych przedziałów
- Inne (użycie klasyfikatorów, dyskretyzacja boolowska, …).

Transformacje danych (3)

Zastąpienie atrybutów o wartościach nominalnych atrybutami o wartościach liczbowych

dla każdej wartości tworzony jest atrybut o wartościach binarnych: przyjmuje wartość 1, gdy w oryginalnych danych występuje dana wartość, w przeciwnym przypadku przyjmuje wartość 0.

Niekompletność

- **Wartości brakujące**
  - Chwilowa niedostępność danych — brakujące dane mogą być łatwo uzupełnione z innych źródeł (np. kod pocztowy ze stron poczty).
  - Brak danych spowodowany ogólną niedoskonałością metod i urządzeń, służących do ich zbierania i zapisywania — takich danych zazwyczaj nie da się w łatwy sposób uzupełnić.
- **Wartości niedostępne** – w danym typie obiektów pojawiają się instancje, do których nie mają zastosowania pewne fragmenty opisu tego typu.

Metody uzupełniania danych

- Pominięcie obiektów
- Proste uzupełnianie danych
  - wartością specjalną,
  - dominantą,
  - średnią
- Użycie klasyfikatora:
  - kNN,
  - drzewa decyzyjne,
  - teorii zbiorów przybliżonych
- Statystyczne