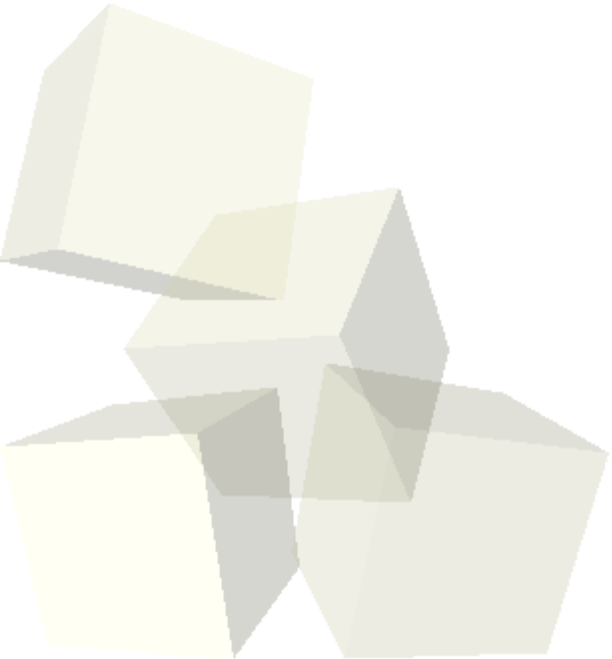
Data Mining.
Przegląd metod eksploracji danych.

Mateusz Kobos
23.11.2005



Plan prezentacji

- Ogólnie o DM
  - Czym jest DM? DM a metody AI.
  - Trochę historii
  - Gdzie się stosuje DM?
  - Do czego się stosuje DM?
  - Proces odkrywania wiedzy z baz danych
  - Główne działy DM
- Ciekawsze metody DM:
  - Odkrywanie reguł asocjacyjnych
  - Tworzenie drzew decyzyjnych
  - Grupowanie metodą WaveCluster

Czym jest DM? DM a metody AI.

- Definicja1: „Nietrywialne wydobywanie ukrytej, poprzednio nieznanej i potencjalnie użytecznej informacji z danych”
  (W.Frawley, G. Piatetsky-Shapiro, C. Matheus. **Knowledge Discovery in Databases: An Overview**. AI Magazine , Jesień 1992)
- Definicja2: „Nauka zajmująca się wydobywaniem informacji z dużych zbiorów danych lub baz danych”
  (D. Hand, H. Mannila, P. Smyt. **Principles of Data Mining**. MIT Press, Cambridge, MA, 2001)

- DM a metody AI:
  - DM operuje na dużych zbiorach danych

Trochę historii

- Ewolucja systemów Bazodanowych:
  - Lata '60 – proste metody przetwarzania plików
  - Lata '70 – wczesne '80 – systemy zarządzania danymi (Database Management Systems)
    - Relacyjne, sieciowe, hierarchiczne BD
    - SQL, transakcje, metody indeksujące, strutkury danych:
      - Tablice mieszające, B+-drzewa, ...
  - Lata '80 – teraz – zaawansowane systemy bazodanowe
    - OO, OR, dedukcyjne
    - Aplikacyjne (przestrzenne, multimedialne, naukowe, ...)
  - Lata '80 – teraz – hurtownie danych, DM
    - Hurtownie danych, technologia OLAP
    - DM i odkrywanie wiedzy (nasilenie zainteresowania w latach '90)
  - Lata '90 – teraz – internetowe BD
    - BD oparte na XML-u
    - Web mining

Gdzie się stosuje DM?

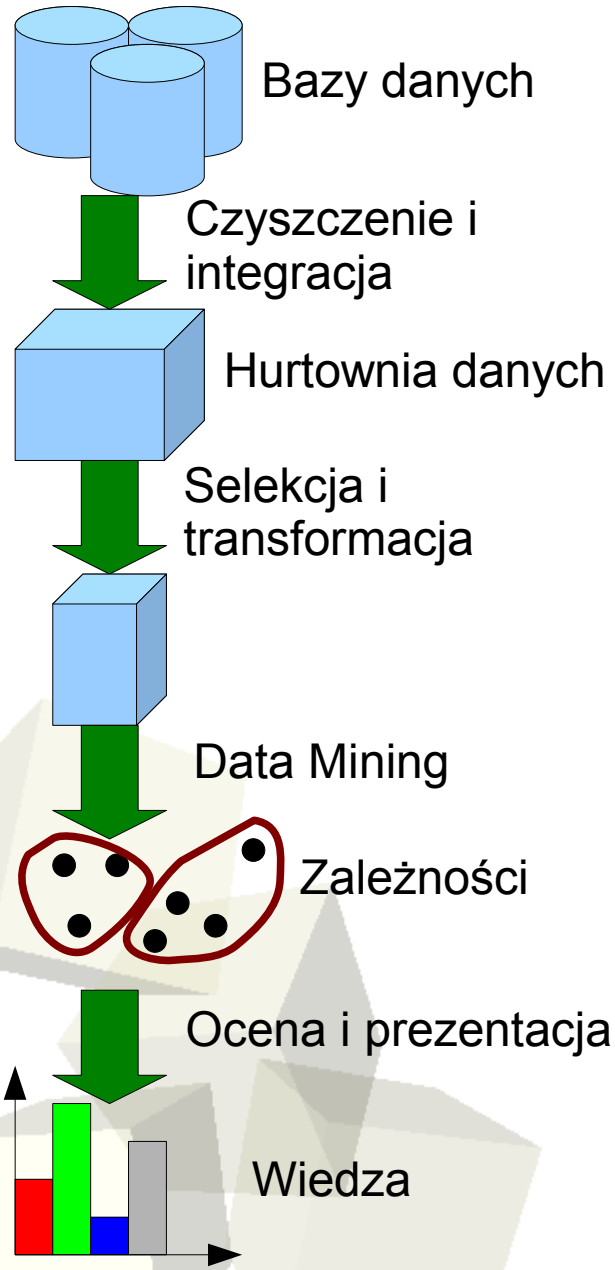
- Firmy z silnym nastawieniem na klienta (i dużymi bazami danych):
  - Sieci sklepów
  - Firmy finansowe
  - Firmy telekomunikacyjne
  - Firmy marketingowe
- Pomoc w znalezieniu relacji między czynnikami:
  - “wewnętrznymi”:
    - Cena, ustawienie produktu, zdolności pracowników
  - i “zewnętrznymi”
    - Współczynniki ekonomiczne, konkurencja, demografia klientów
- Ustalenie wpływu tych czynników na:
  - Wielkość sprzedaży,
  - Zadowolenie klienta
  - Przychody

Do czego się stosuje DM?

- Popularny przykład (ale czy prawdziwy?):
  - mężczyźni+pieluchy+piątek=piwo
- Zastosowania (USA):
  - Blockbuster – proponuje filmy na podstawie historii wypożyczeń
  - American Express – proponuje produkty posiadaczom kart na podstawie analizy miesięcznych wydatków
  - WalMart –
    - 2,600 sklepów
    - <terabajtowa hurtownia danych

Proces odkrywania wiedzy z BD

(Knowlede Discovery in Databases)



- Czyszczenie danych – usuwanie szumu i niespójnych danych
- Integracja danych – łączenie danych z różnych źródeł
- Selekcja danych – wybór ważnych (dla problemu) danych
- Transformacja danych – do postaci odpowiedniej do DM (np. sumowanie, agregacja)
- DM – zastosowanie inteligentnych metod do wydobycia zależności, wzorców
- Ocena zależności – identyfikacja interesujących zależności ze wszystkich wydobytych
- Prezentacja wiedzy

Główne działy DM

- Wyszukiwanie asocjacji (pieluchy-piwo)
- Klasyfikacja(wartości dyskretne), predykcja(wartości ciągłe)
- Grupowanie (ang. clustering)
- Eksploracja złożonych typów danych

Wyszukiwanie asocjacji

- Algorytm Apriori
- Inne, bardziej zaawansowane

Klasyfikacja, predykcja

- Klasyfikacja
  - Drzewa decyzyjne
  - Modele Bayes'a
  - Sieci neuronowe (perceptron)
  - k-Nearest Neighbours
  - Alg. Genetyczne
  - Case-based reasoning
  - Zbiory rozmyte i przybliżone
- Predykcja
  - Statystyczna regresja wielowymiarowa, inne rodzaje regresji

Grupowanie

- Klasyczne metody podziału:
  - K-Means, k-Medoids
- Metody hierarchiczne
  - BIRCH, CURE, Chameleon
- Metody oparte na gęstości
  - DBSCAN, OPTIC, DENCLUE
- Metody gridowe (ang. grid-based)
  - STING, WaveCluster, CLIQUE
- Metody oparte na modelu
  - Podejście statystyczne, sieci neuronowe
- Analiza odchyleń (ang. outlier analysis)

Eksploracja złożonych typów danych

- Eksploracja przestrzennych (ang. spatial) BD
- Eksploracja multimedialnych BD
- Eksploracja szeregów czasowych i danych sekwencyjnych
- Eksploracja tekstowych BD
- Web-mining

Wyszukiwanie reguł asocjacyjnych

Wyszukiwanie reguł asocjacyjnych

- Chcemy się dowiedzieć jakie produkty są kupowane razem i otrzymujemy przykładową regułę asocjacyjną:

komputer→program do zarządzania finansami
[wsparcie względne=2%, zaufanie=60%]

$$wsparcie(A) = w\ ilu\ transakcjach\ występuje\ A$$

$$wsparcie\ względne(A \to B) = P(A \cup B) = \frac{wsparcie(A \cup B)}{ilość\ wszystkich\ transakcji}$$

$$zaufanie(A \to B) = P(B|A) = \frac{P(A \cup B)}{P(A)} = \frac{wsparcie(A \cup B)}{wsparcie(A)}$$

Reguły asocjacyjne – przykład (z [1])

Dane

Dane transakcyjne:

TID	*Produkty*
1	P1, P2, P5
2	P2, P4
3	P2, P3
4	P1, P2, P4
5	P1, P3
6	P2, P3
7	P1, P3
8	P1, P2, P3, P5
9	P1, P2, P3

minimalne wsparcie=2

minimalne zaufanie=1/3

- Generujemy wszystkie zbiory ze wsparciem ≥2 (wsparcie podane w nawiasie):
  - P1(6), P2(7), P3(6), P4(2), P5(2)
  - P1P2(4), P1P3(4), P1P5(2), P2P3(4), P2P4(2), P2P5(2)
  - P1P2P3(2), P1P2P5(2)
- Z każdego z tych zbiorów generujemy reguły asocjacyjne. Np. dla P1P2P5:
  - P1P2→P5
    - Wsparcie wzgl.=2/9, zaufanie=2/4
  - P1P5→P2
    - Wsparcie wzgl.=2/9, zaufanie=2/2
  - P2P5→P1 : w=2/9, z=2/2
  - P1→P2P5 : w=2/9, z=2/6
  - ~~P2→P1P5~~ : w=2/9, z=2/7
  - P5→P1P2 : w=2/9, z=2/2

A inne algorytmy?

- Algorytm Apriori(1994) polega na sprytnym generowaniu (tak, by jak najrzadziej zaglądać do BD):
  - Wszystkich zbiorów o wsparciu ≥ minimalnego wsparcia
  - Reguł asocjacyjnych o zaufaniu ≥ od minimalnego zaufania
- Inne, ulepszone wersje

Apriori – przykład generowania

Dane

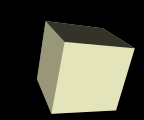
Dane transakcyjne:

TID	*Produkty*
1	P1, P2, P5
2	P2, P4
3	P2, P3
4	P1, P2, P4
5	P1, P3
6	P2, P3
7	P1, P3
8	P1, P2, P3, P5
9	P1, P2, P3

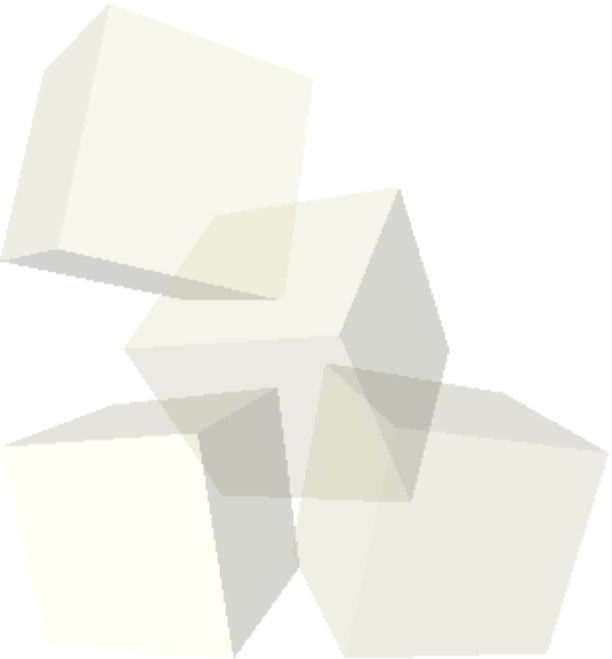
minimalne wsparcie=2

minimalne zaufanie=1/3

- Generujemy wszystkie zbiory ze wsparciem ≥2 (wsparcie podane w nawiasie), każdy punkt to 1 przeglądnięcie BD:
  - P1(6), P2(7), P3(6), P4(2), P5(2)
  - P1P2(4), P1P3(4), ~~P1P4~~(1), P1P5(2), P2P3(4), P2P4(2), P2P5(2), ~~P3P4~~(0), ~~P3P5~~(1), ~~P4P5~~(0)
  - P1P2P3(2), P1P2P5(2)
- Analogicznie generujemy reguły asocjacyjne (ale już bez patrzenia do BD)

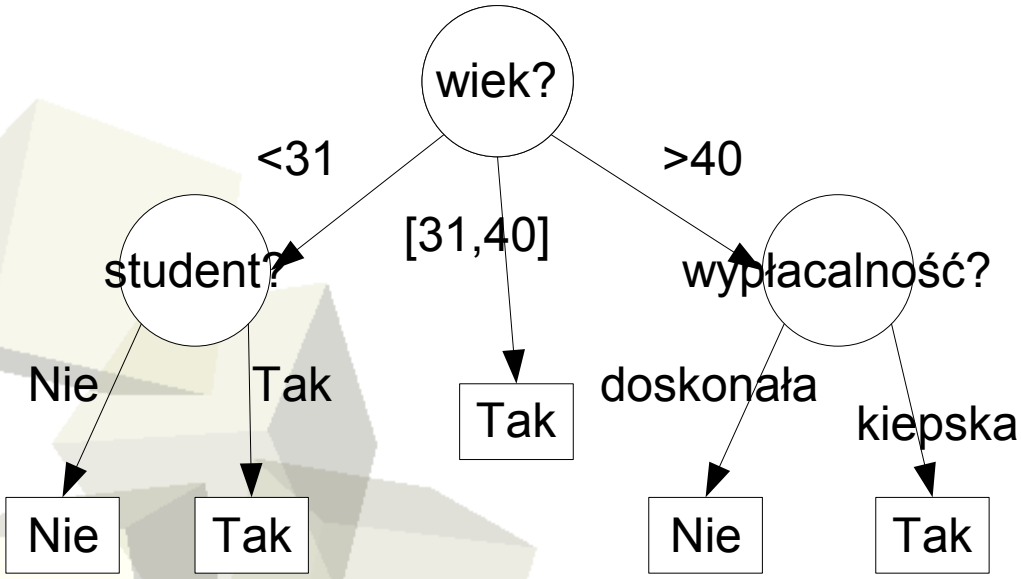


Drzewa decyzyjne



Drzewa decyzyjne

- Drzewo decyzyjne:
  - Drzewo
  - Wewnętrzne węzły – test na atrybucie
  - Gałęzie – wyniki testu
  - Liście – klasy
- Przykładowe drzewo (wskazujące, czy klient kupi komputer, czy nie):

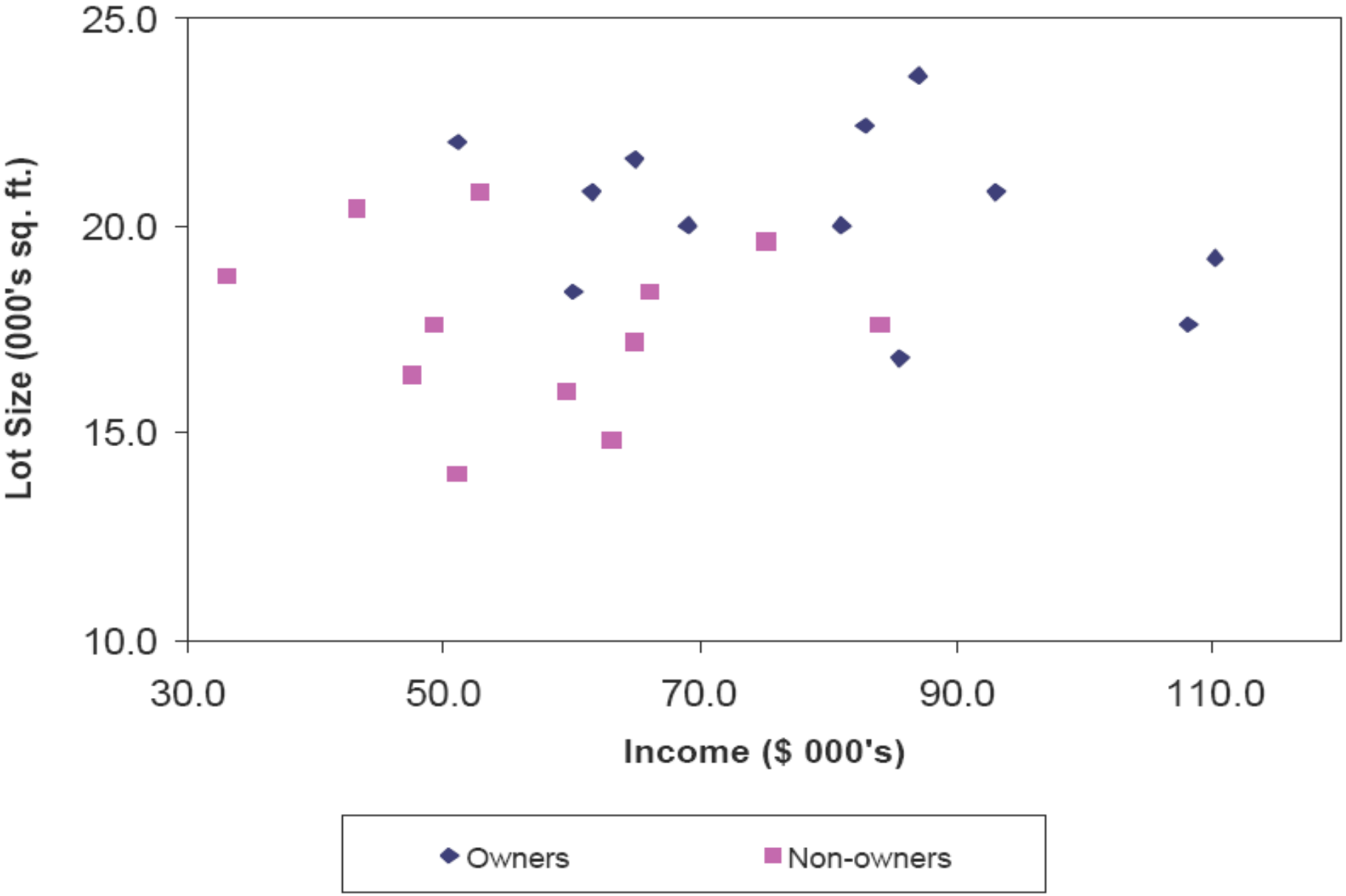


Na podstawie rysunku w [1]

Drzewa decyzyjne - przykład

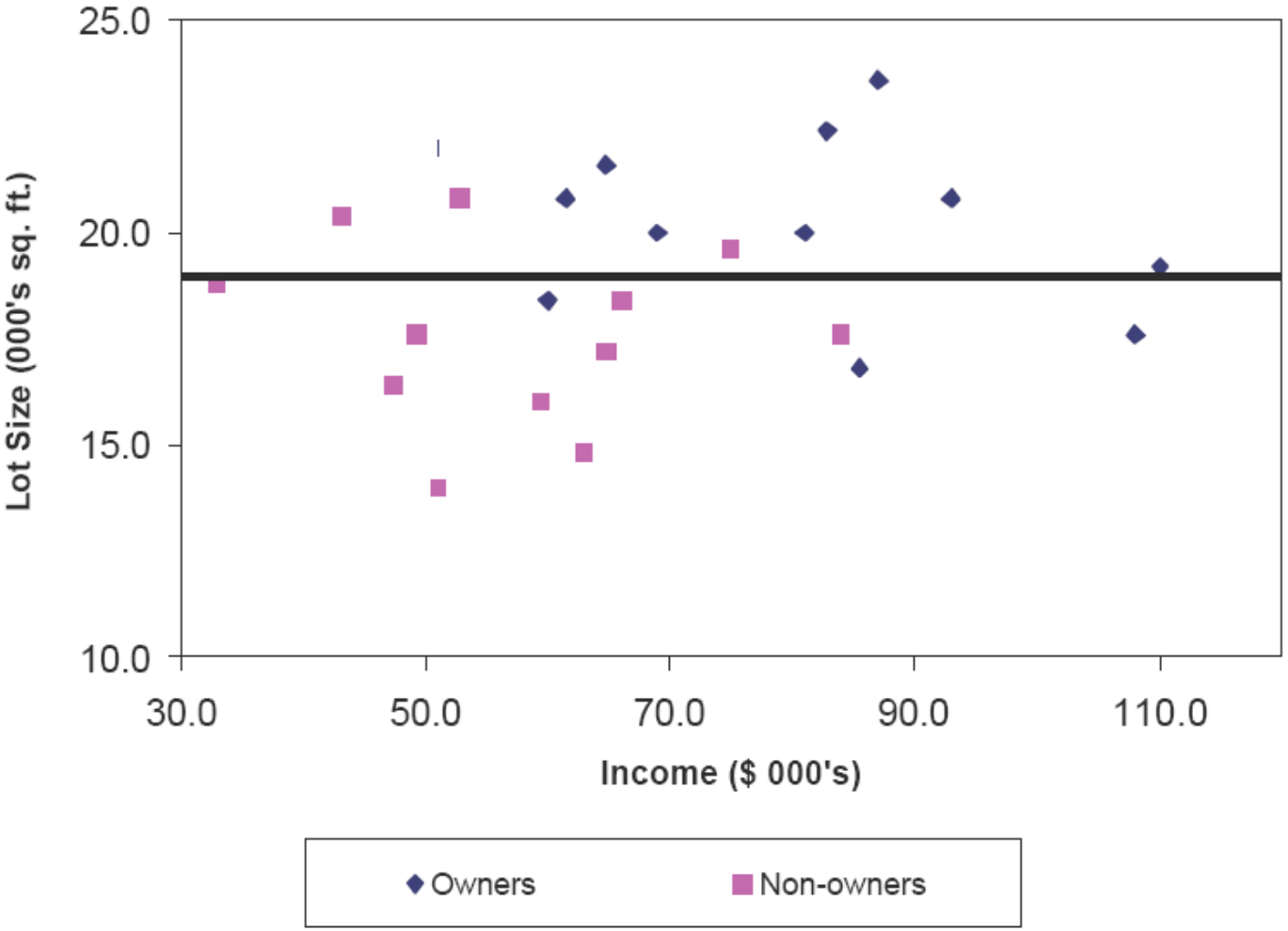
- Fabryka kosiarek samojezdnych chce podzielić rodziny na te, które:
  - chciałyby kupić kosiarkę i
  - nie chciałyby kupić kosiarki
- Zebrano dane o 12 rodzinach posiadających kosiarkę i 12 nieposiadających kosiarki:
  - Rozmiar działki (lot size)
  - Przychód (income)
- Przedstawimy analizę tych danych na przykładzie algorytmu 2-etapowego algorytmu CART(1984):
  - Rekurencyjny podział przestrzeni
  - Obcinanie gałęzi przy użyciu danych walidacyjnych

Drzewa decyzyjne - przykład



Rysunek z [3]

Drzewa decyzyjne - przykład



- Dzielimy przestrzeń tak, by nowopowstałe części były jak najbardziej jednorodne

Rysunek z [3]

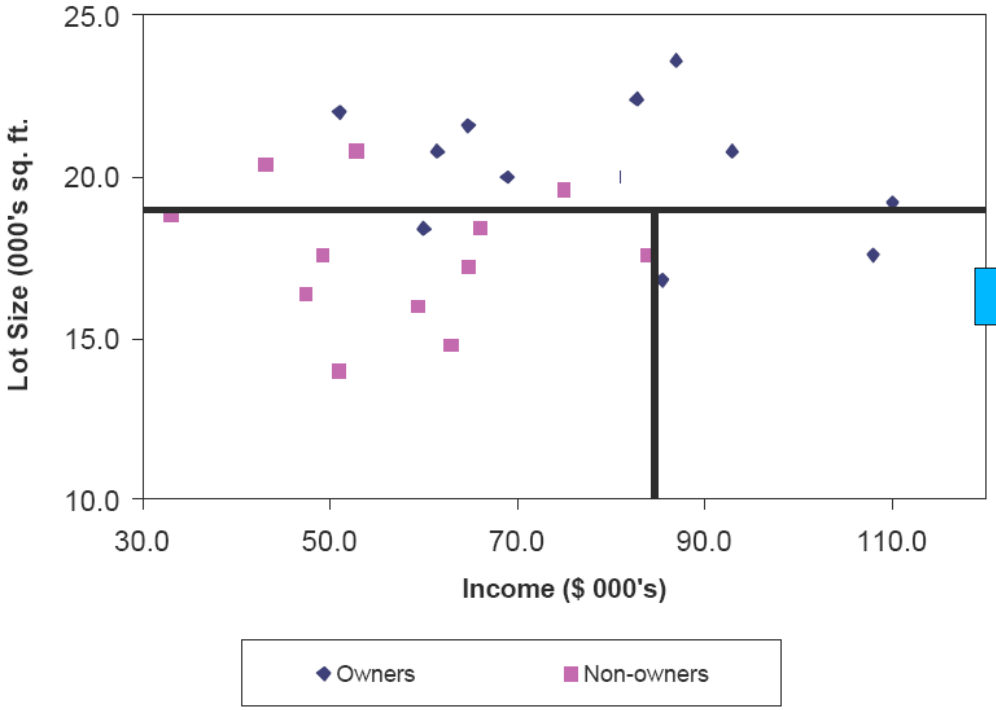
Jak doszło do podziału?

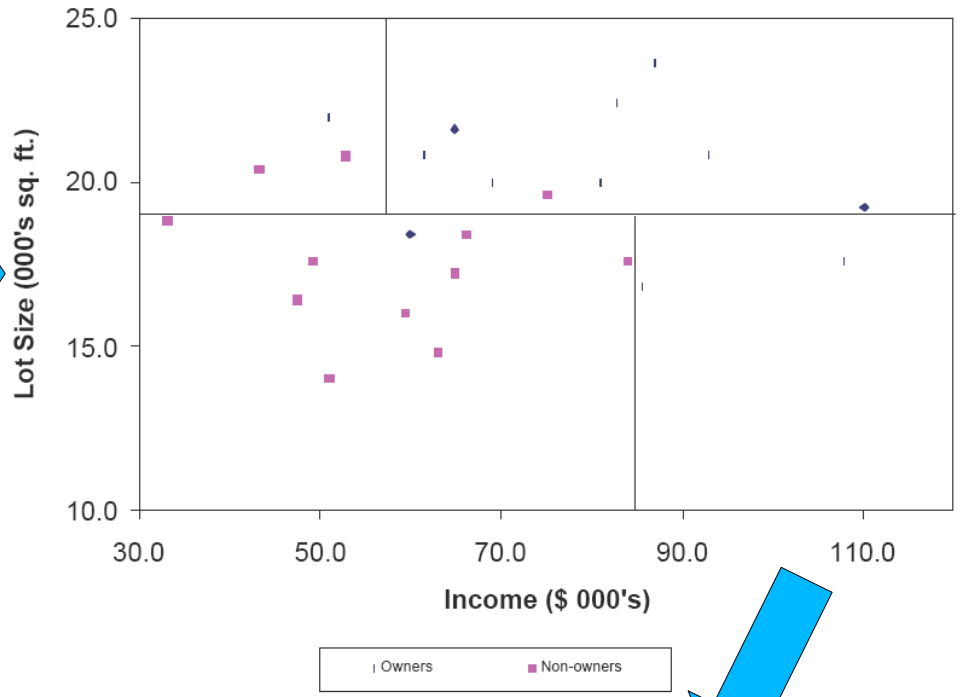
- Sprawdzamy wszystkie możliwe miejsca podziału dla każdej zmiennej (czyli wymiaru)
- Możliwe miejsca podziału to punkty pomiędzy dwoma kolejnymi punktami zrzutowanymi na daną oś
- Ocena jakości podziału = Zanieczyszczenie prostokąta przed podziałem – zanieczyszczenie prostokątów powstałych po podziale
- Miara zanieczyszczenia:
  - Np. indeks Gini:

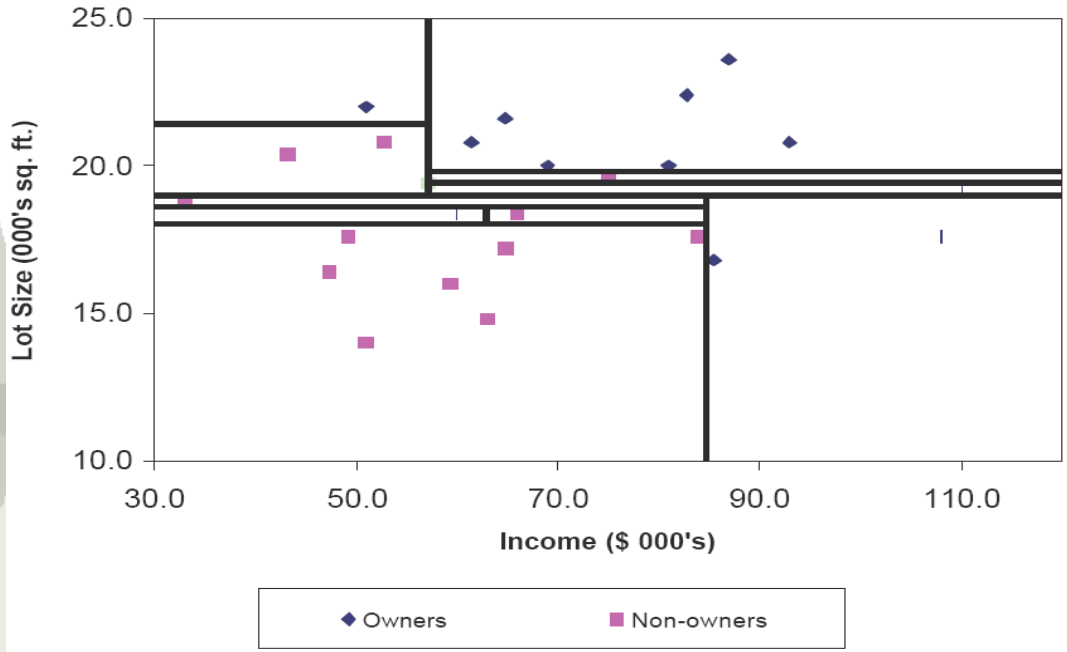
$$I(A) = 1 - \sum_{k=1}^{C} p_k^2$$

Gdzie:
A – badany prostokąt
C – liczba klas
p_k - ułamek obserwacji w A, które należą do klasy k

Podziały cd..



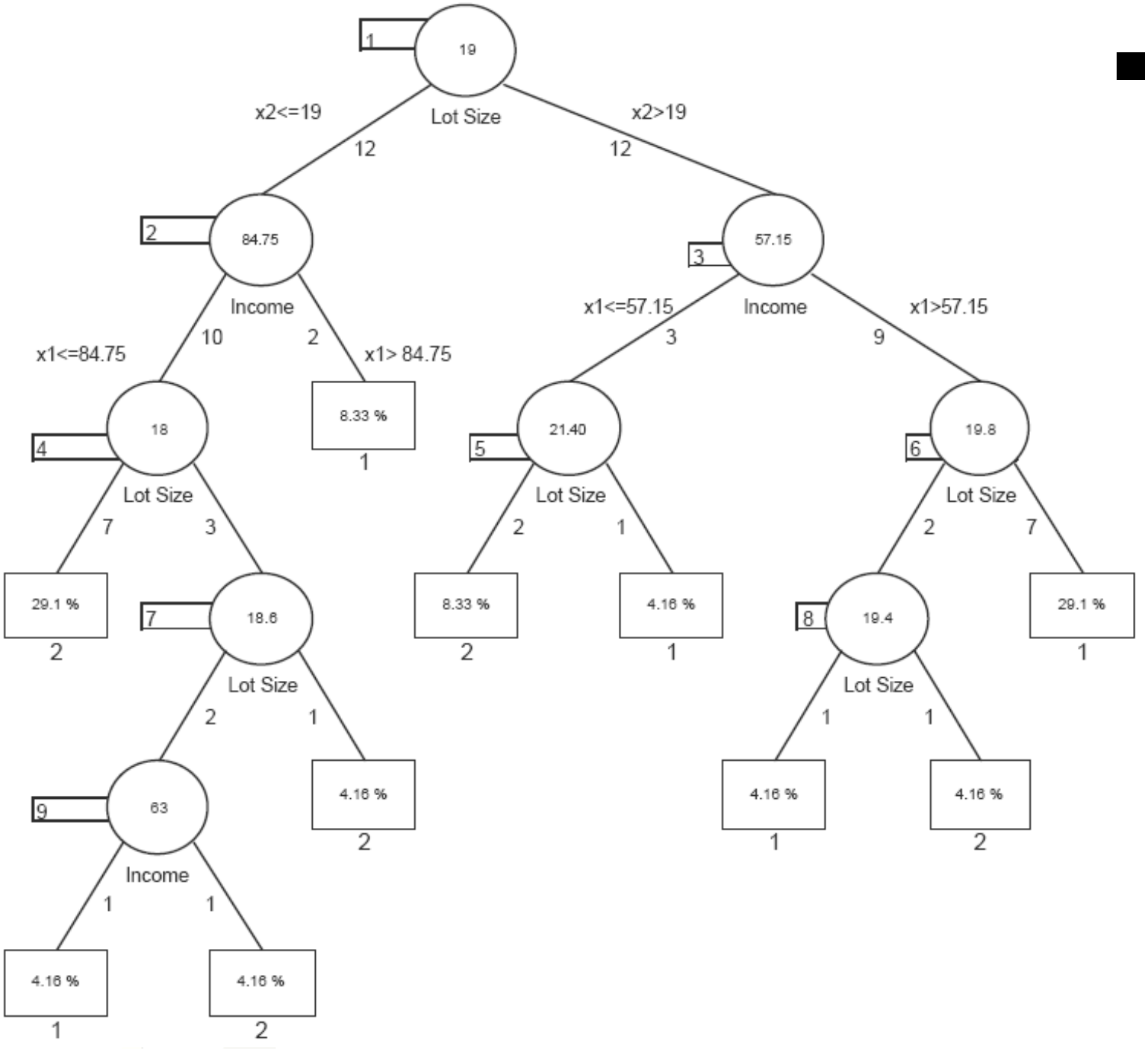




....

Rysunki z [3]

Utworzone drzewo

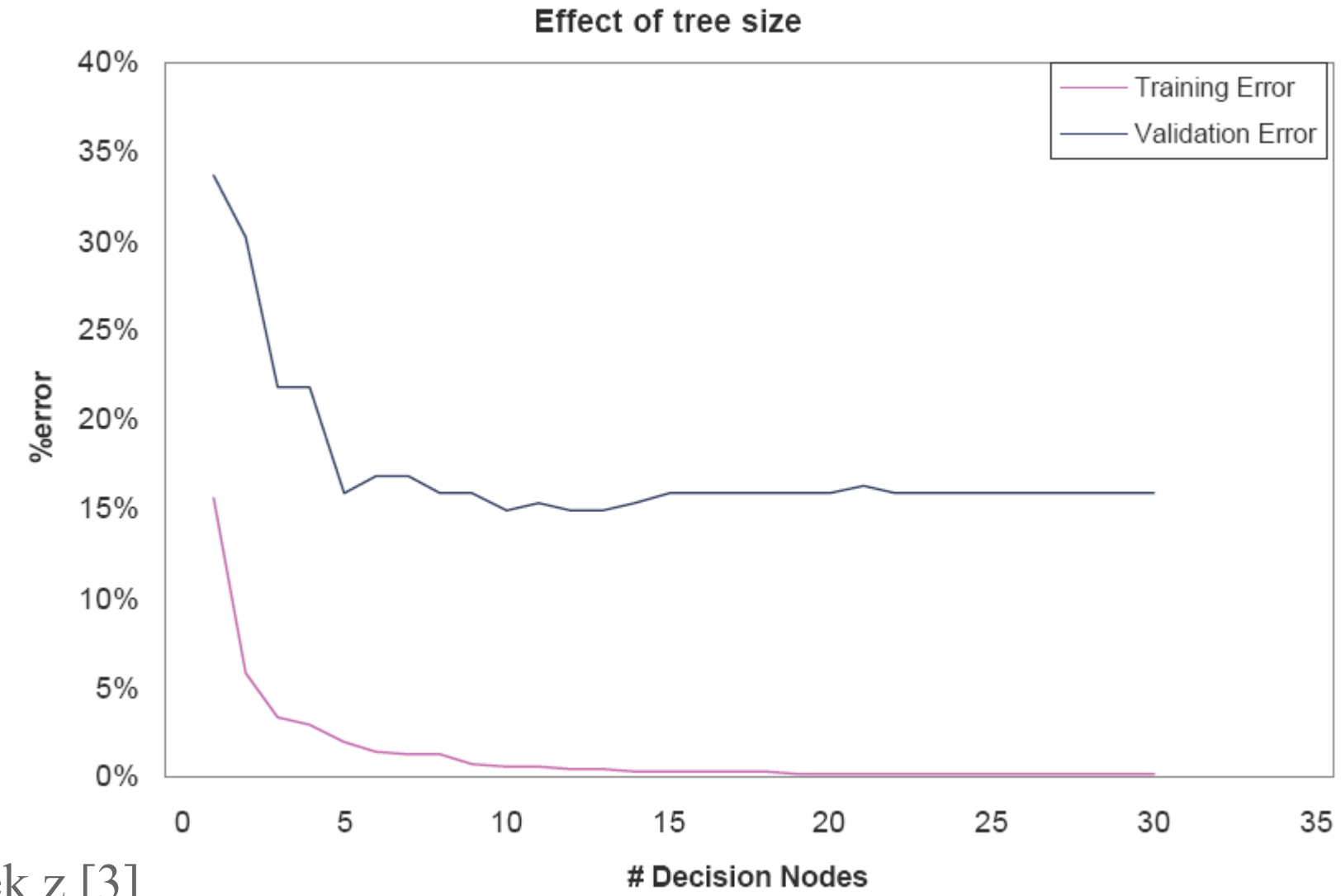


- Każdy podział, to nowy węzeł wewnętrzny

Rysunek z [3]

Obcinanie zbędnych gałęzi

- Po zbudowaniu ogromnego drzewa:
  - Wybieramy ten etap rozwoju drzewa, dla którego błąd klasyfikacji na zbiorze walidacyjnym był najmniejszy



Rysunek z [3]

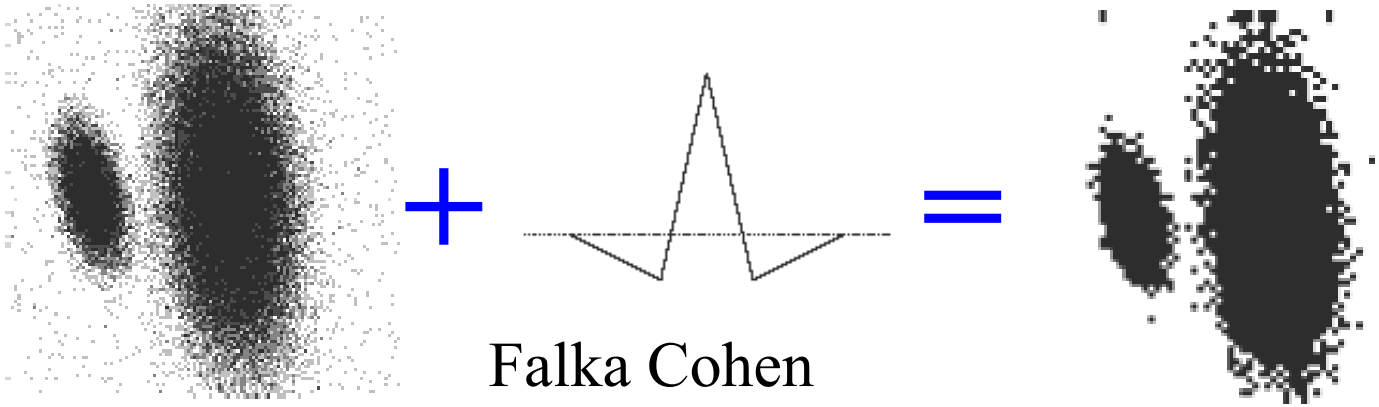
Grupowanie algorytmem WaveCluster

WaveCluster(1998)

- Wykorzystuje dyskretną tranformatę falkową (ang. discreete wavelet transform), która:
  - Dzieli 1-wymiarowy sygnał wejściowy na 2 pasma (zmniejszając dwukrotnie rozdzielczość):
    - Wysokiej częstotliwości – odpowiada brzegom grup
    - Niskiej częstotliwości – odpowiada wnętrzom grup
  - Sygnał 2-wymiarowy dzielimy stosując 2 razy transformatę 1-wymiarową. Otrzymujemy 4 pasma częstotliwości:
    - LL – niska-niska
    - LH – niska-wysoka
    - HL – wysoka-niska
    - HH – wysoka-wysoka

Przykład zastosowania falki

- Podziału sygnału dokonujemy stosując odpowiedni filtr-falkę:



Wejście

Falka Cohen
-Daubechies
-Feauveau(2,2)

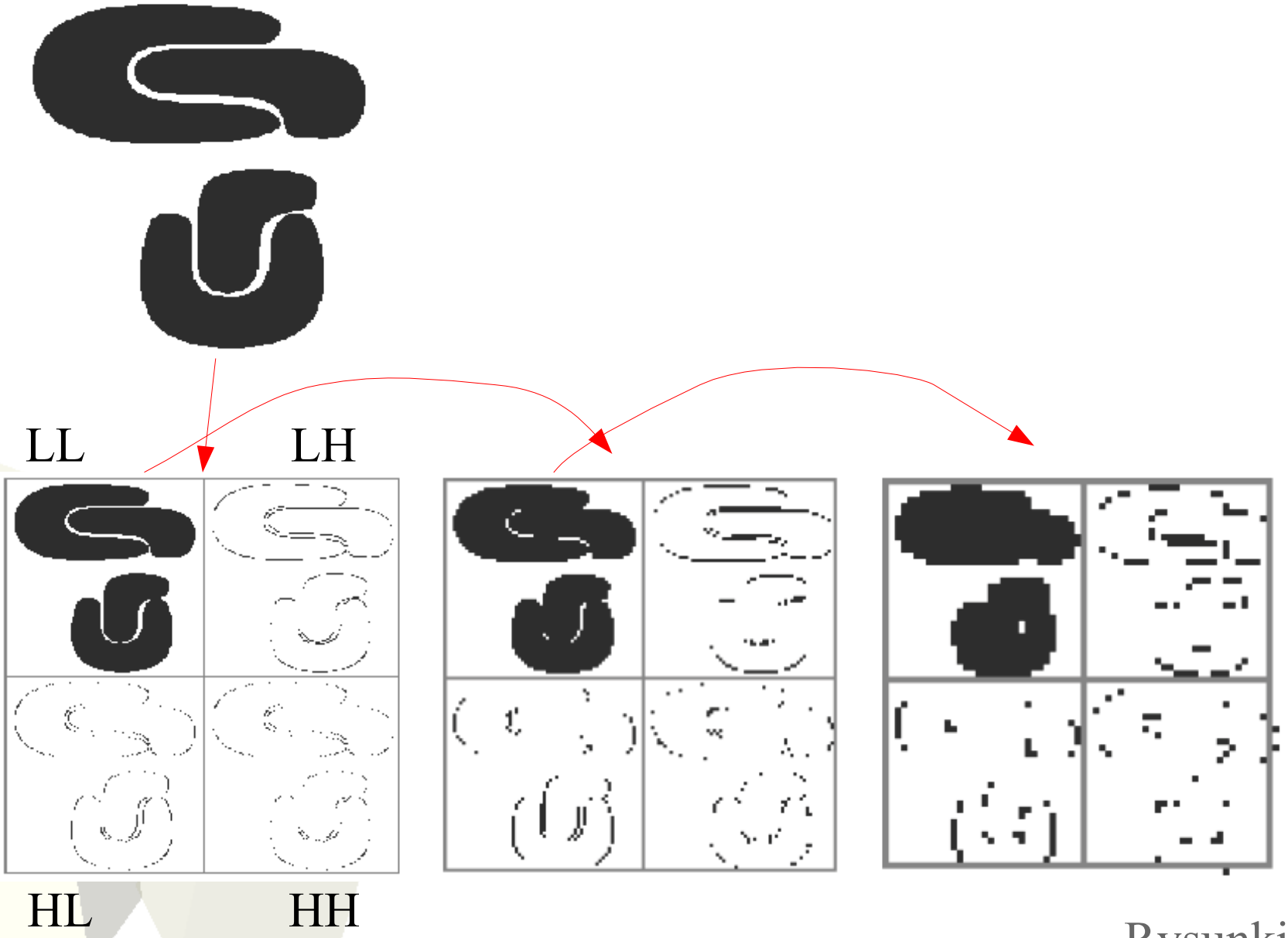
Powstała część LL

- Wyostrzyliśmy kształty i wyeliminowaliśmy szum

Rysunki z [4]

Działanie algorytmu – przykład

- Przykład wykonywania kolejnych transformacji

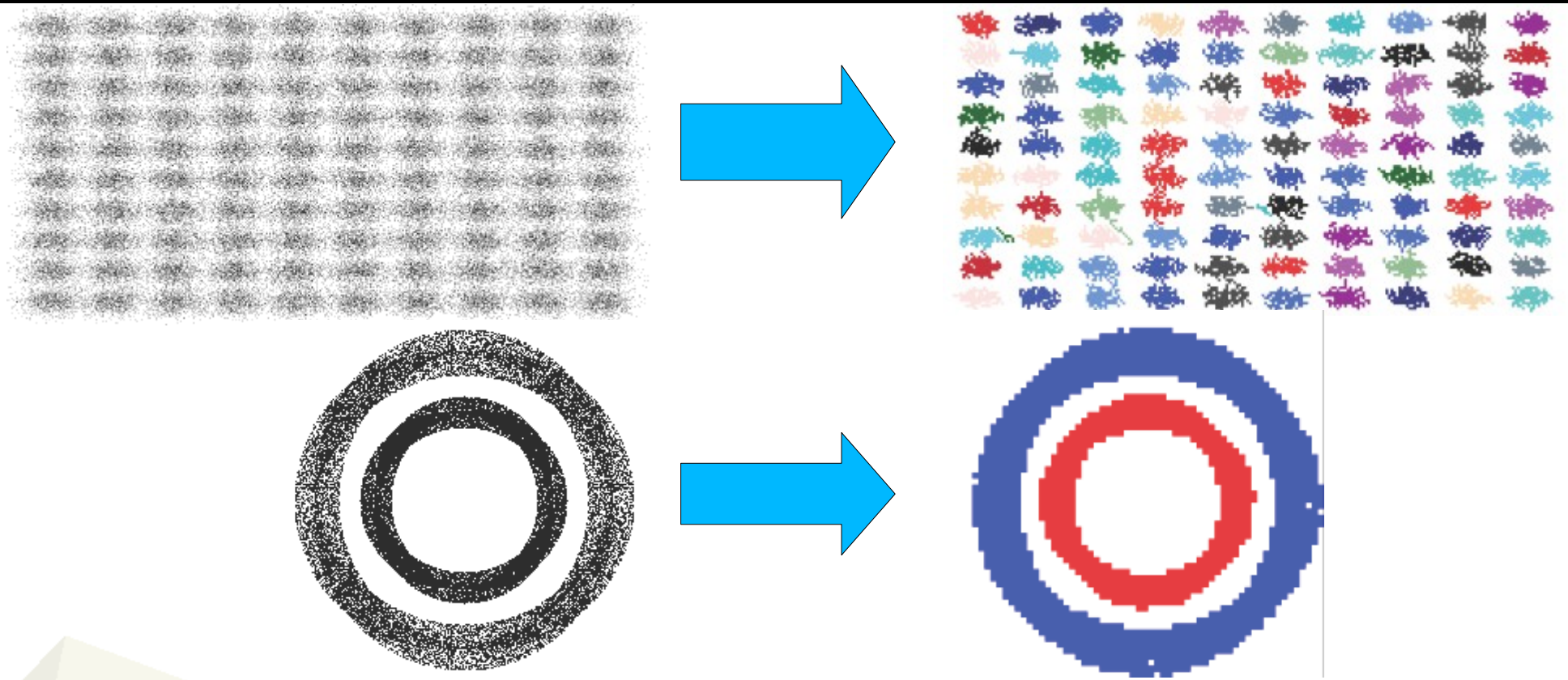


Rysunki z [4]

WaveCluster

- Algorytm (wejście: zbiór wielowymiarowych punktów(obiektów), wyjście: pogrupowane punkty)
  1. Podziel przestrzeń na jednostki (każda z jednostek sumuje informację punktów w niej zawartych)
  2. Zastosuj transformatę falkową na przestrzeni
  3. Znajdź połączone jednostki w przekształconej przestrzeni (określamy grupy)
  4. Przypisz przekształconym jednostkom etykiety grup
  5. Przejdź do zwykłej przestrzeni - dokonaj mapowania: jednostka przekształcona→ zwykłe jednostki
  6. Przypisz punkty do klastrów
- Operację powtarzamy aż do uzyskania zadowalającej rozdzielczości (a raczej zadowalającego rozmycia)

Przykłady znalezionych grup (za [4])



Znalezione grupy dla różnych rozdzielczości

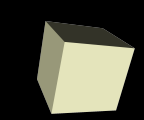
Rysunki z [4]

Zalety i wady algorytmu

- Nie trzeba podawać trudnych do określenia parametrów (jak np. w k-means, k-medoids)
  - Ale trzeba podać:
    - Wymiar jednostki (hiper-prostokąta), za pomocą której dzielimy przestrzeń
    - Ilość zastosowań transformaty falkowej (szukana rozdzielczość)
- Znajduje grupy dowolnych kształtów
- Wydajny (złożoność O(n)), można zaimplementować równolegle
- Odporny na szumy
- Mamy dostępne wiele poziomów dokładności (wada i zaleta)
- Wada: Dobrze radzi sobie tylko z danymi niskowymiarowymi (do 20 wymiarów)

Bibliografia

[1] Jawei Han, Micheline Kamber, "Data Minig: Concepts & Techniques", Morgan Kaufmann, 2000
[2] Janson Frand, "What is Data Mining?" (http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm) (22-11-2005)
[3] Nitin Patel, materiały do kursu nt Data Mining-u "15.062 Data Mining, Spring 2003" przygotowane przez prof. Nitin Patel, http://ocw.mit.edu/OcwWeb/Sloan-School-of-Management/15-062Data-MiningSpring2003/CourseHome/index.htm, MIT (2005)
[4] Gholamhosein Sheikoholeslami, Surojit Chatterjee, Aidong Zhang, "WaveCluster: A Multi-resolution clustering approach for very large spatial databases", Proceedings of the 24th VLDB Conference, NY, USA, 1998



Dziękuję za uwagę!