Zaawansowana Eksploracja Danych

Przetwarzanie wstępne danych

JERZY STEFANOWSKI
Instytut Informatyki
Politechnika Poznańska

Wykład 3

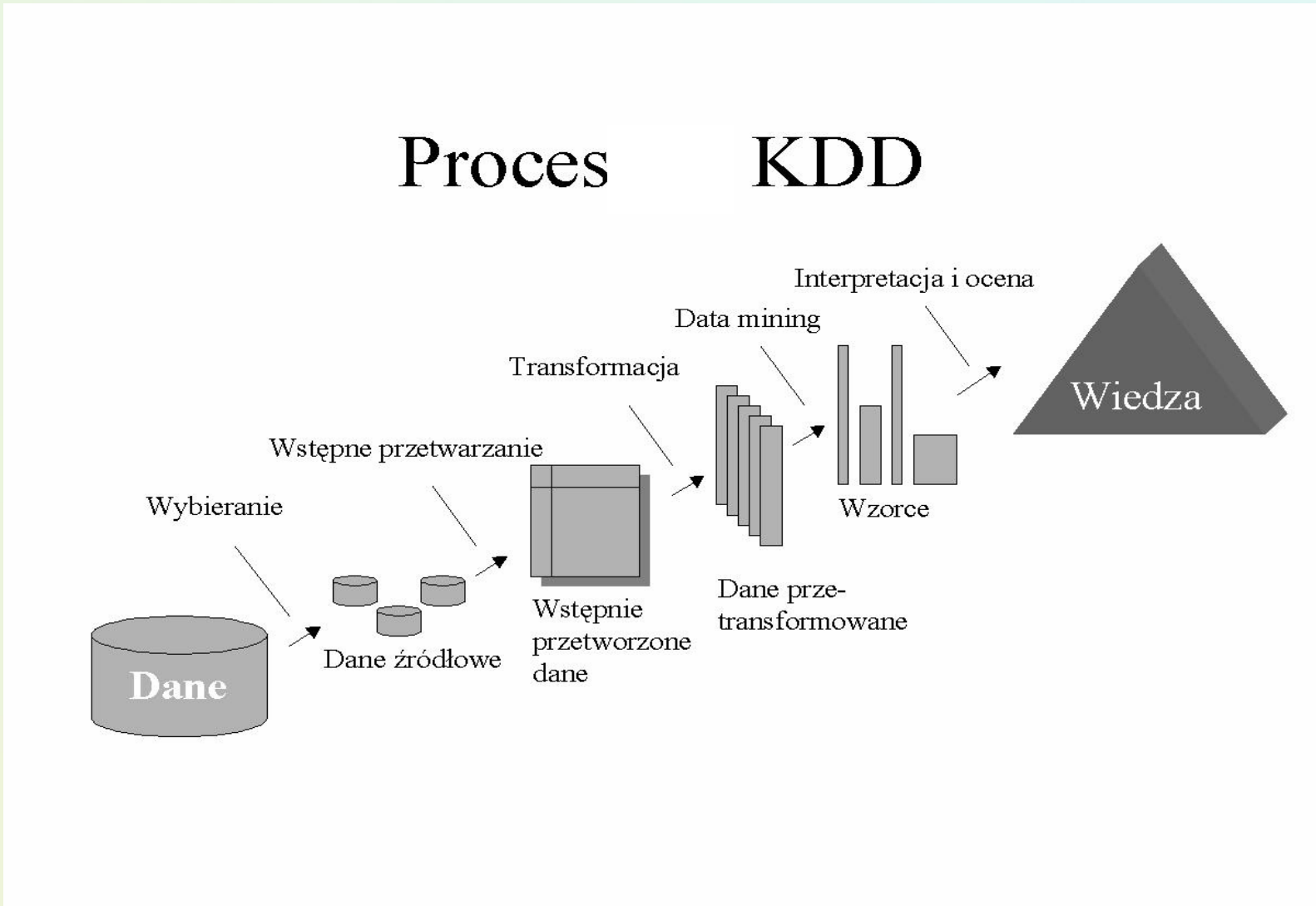
TPD – Zaawansowana eksploracja danych
2008/2009
Aktualizacja 2010

Plan wykładów

1. Miejsce przetwarzania wstępnego danych w procesie KDD
2. Typy przetwarzanych danych
3. Związki z integracją danych
4. Oczyszczanie danych
   - Wykrywanie błędów
   - Nieznane wartości atrybutów
   - Identyfikacja obserwacji odstających
5. Transformacje atrybutów
6. Dyskretyzacja atrybutów liczbowych
7. Redukcja rozmiarów danych
   - Selekcja atrybutów
   - Wybór obiektów

- Slajdy – niektóre częściowo oparte na materiałach od Han i Tan, Steinbach, Kumar

Proces odkrywanie wiedzy i etapy początkowe

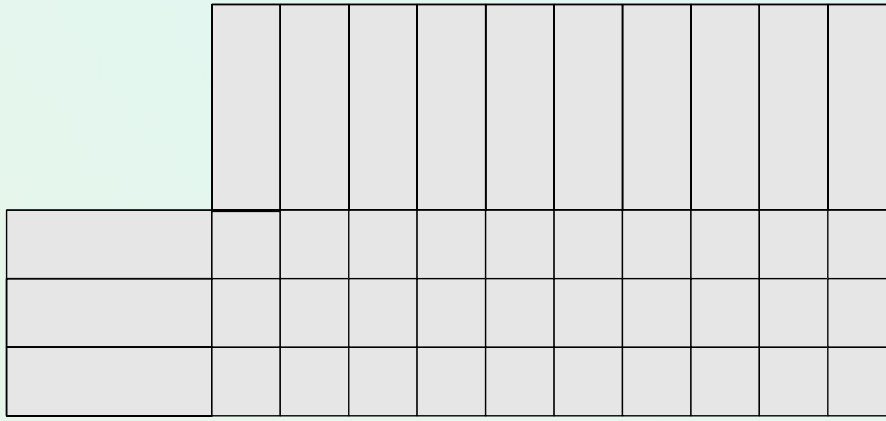


Kilka pytań:

- Jakie źródła danych są związane z zadaniem / zastosowaniem?
- Które z dostępnych danych są adekwatne do celów zastosowania (data relevant)?
- Czy mamy dostęp do innych źródeł danych?
- Jakiej wielkości są dane historyczne (obiekty i atrybuty)?
- Kto dobrze zna posiadane dane (who is data expert)?

Zróżnicowanie typów danych → Han's book

- Records (tablice danych)
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered events
  - Spatial data: maps
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data



TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Tabela danych – typowa reprezentacja

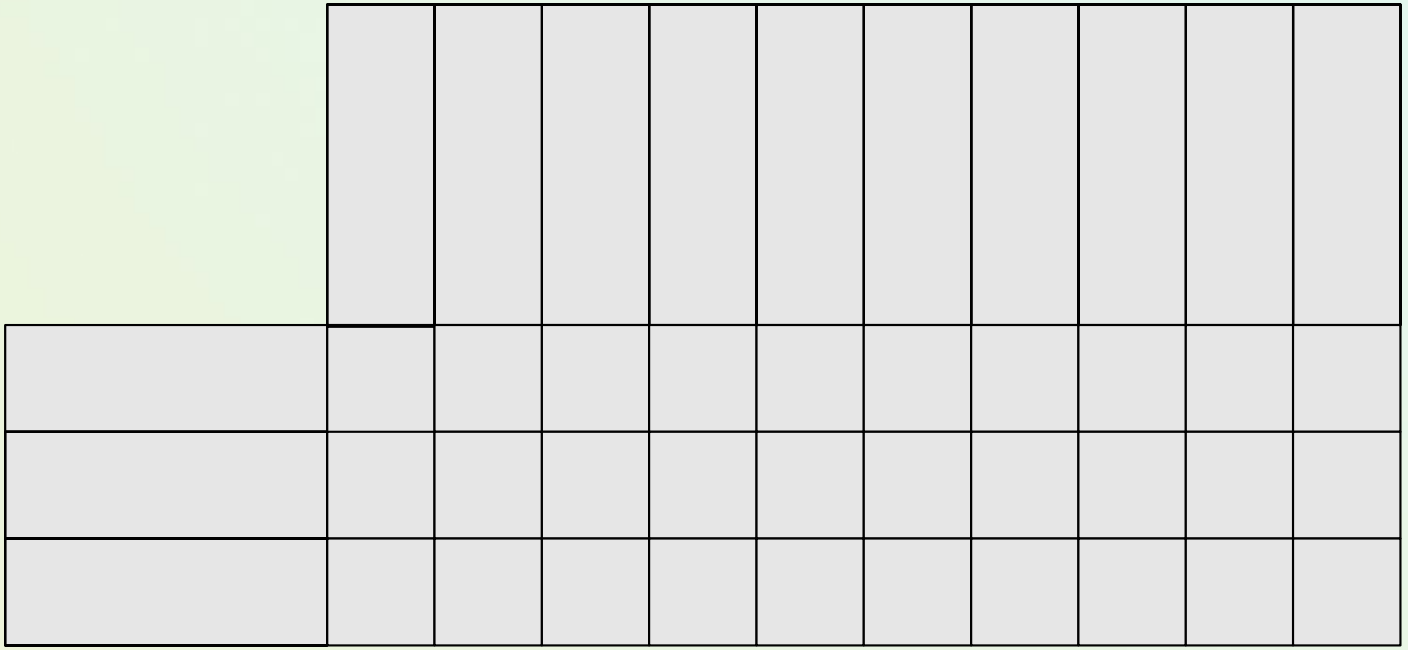
- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an $m \times n$ matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Analiza tekstów

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.



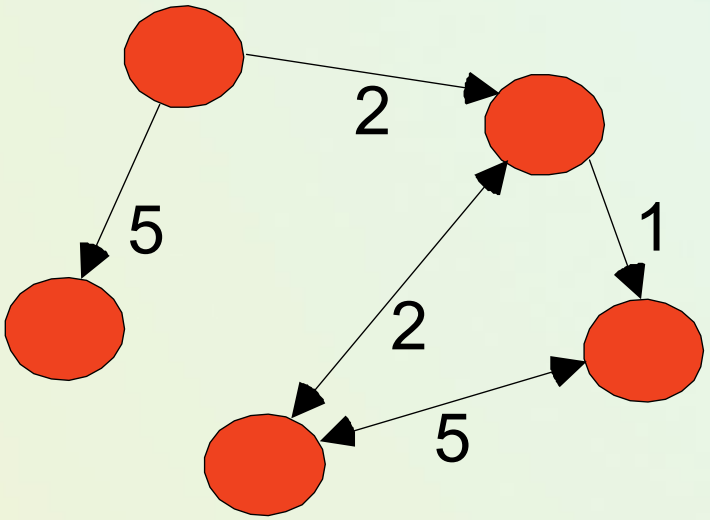
Dane transakcyjne

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

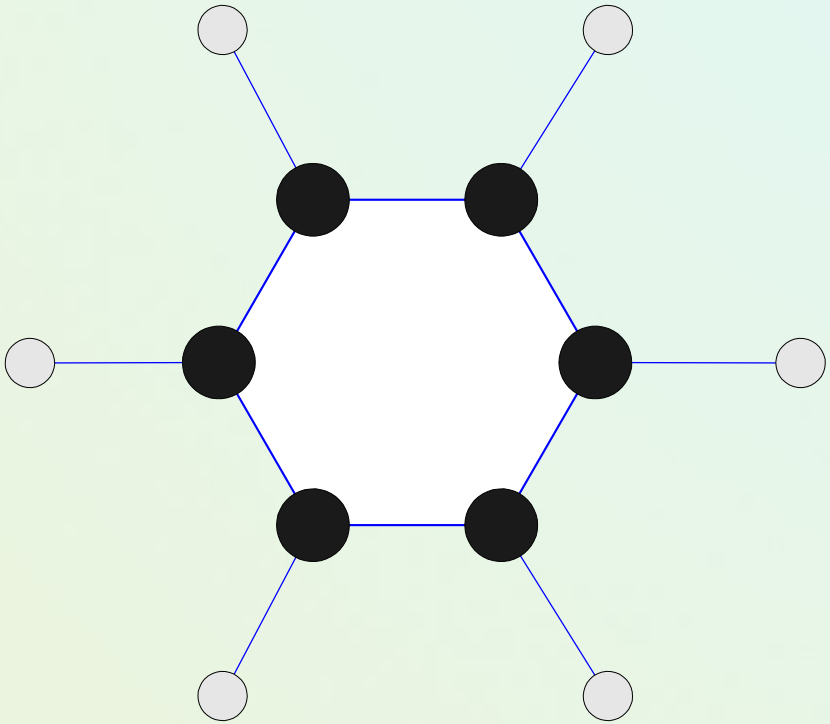
- Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

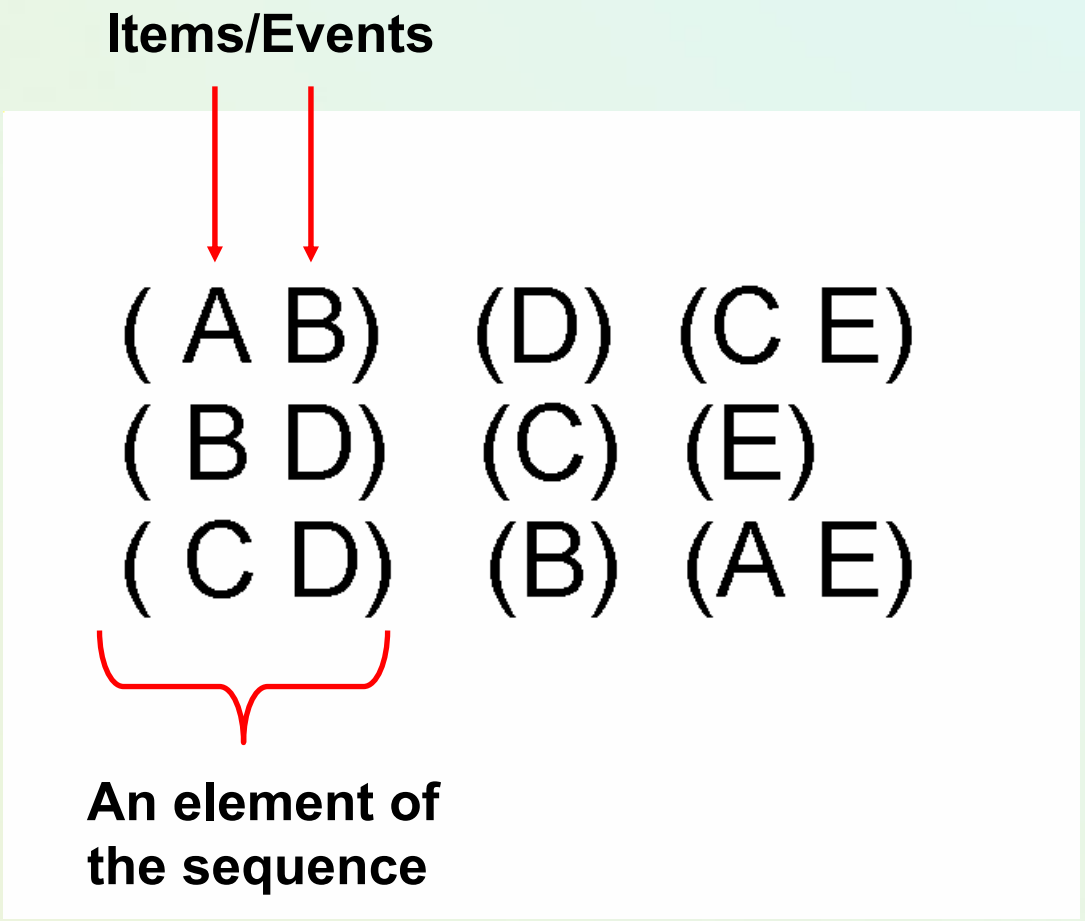
Chemical Data

- Benzene Molecule: C_6H_6



Sequence Ordered Data

- Sequences of transactions



Inne rozumienie sekwencji

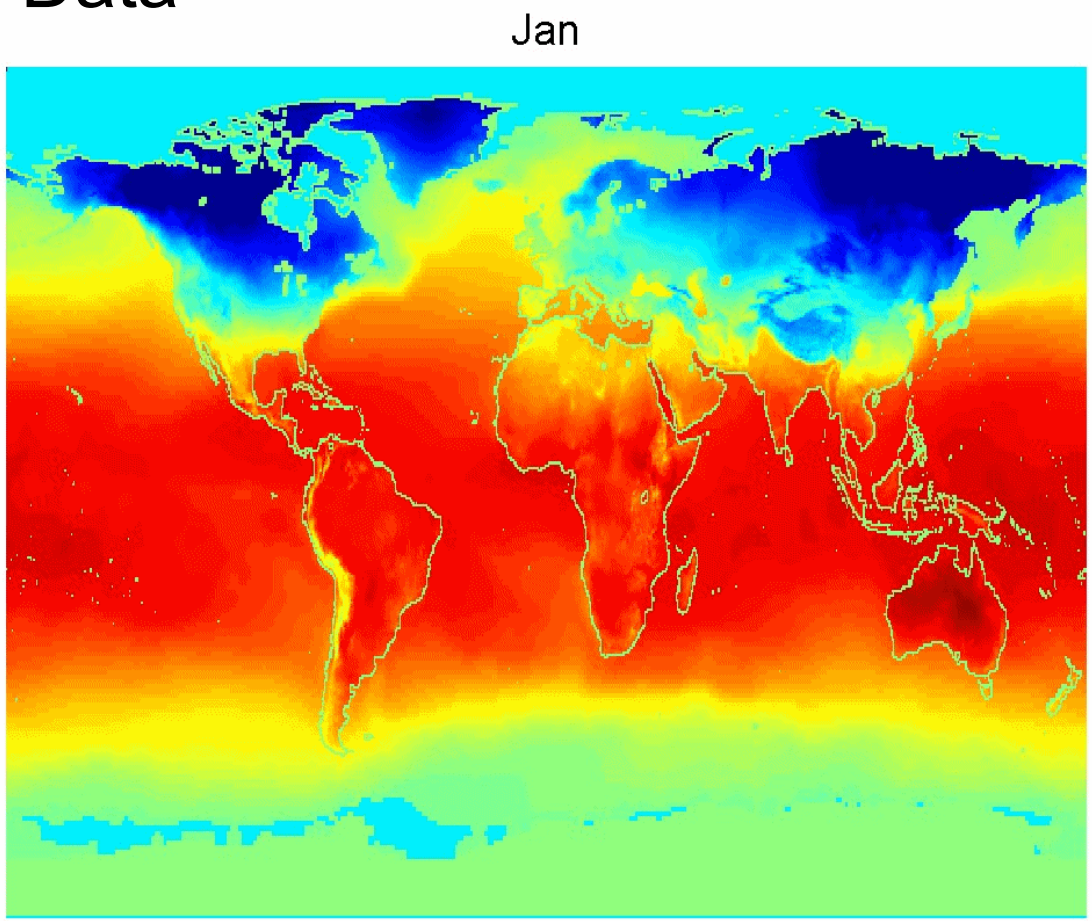
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Złożony typ danych

- Spatio-Temporal Data

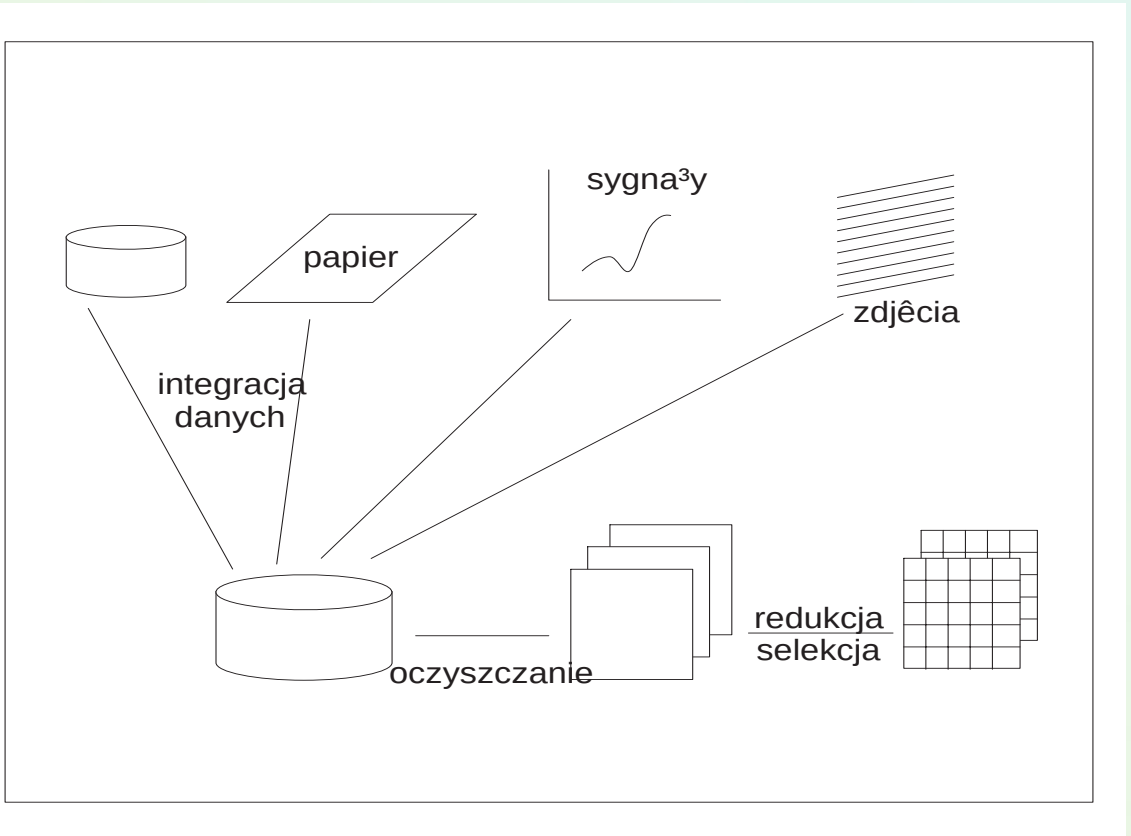
Average Monthly Temperature of land and ocean



Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Motywacje do …

- Rzeczywiste dane mogą być niespójne, niekompletne i obarczone różnego rodzaju zaburzeniami.
- Ponadto mogą występować różne trudności związane z ekstrakcją oraz integracją danych pochodzących z różnych źródeł.

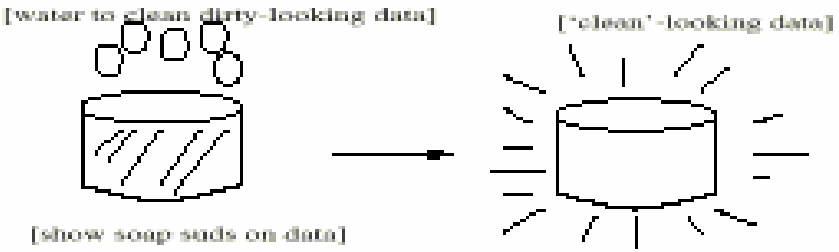


Problemy w przygotowaniu danych

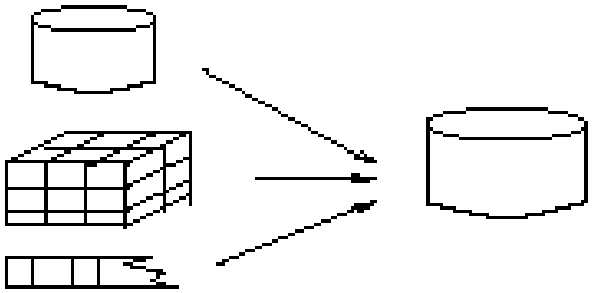
1. Integracja danych (Hurtownie danych)
2. Oczyszczanie
   - Błędy
   - Missing attribute values
   - Noisy Data
   - Outliers
3. Transformacje atrybutów
4. Dyskretyzacja atrybutów liczbowych
5. Redukcja rozmiarów
   - Selekcja atrybutów
   - Wybór obiektów

Różne czynności w przetwarzaniu wstępnym danych [rys. za J.Han]

Data Cleaning



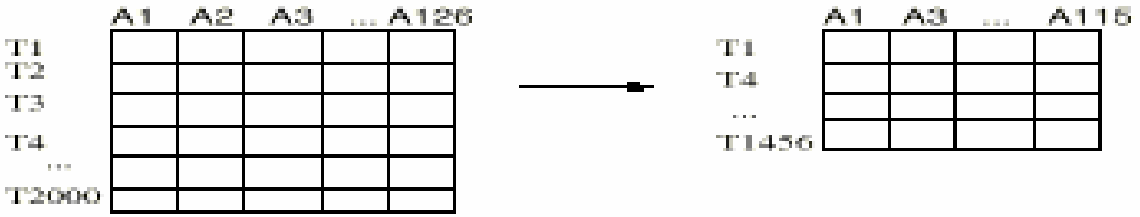
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



ata Transformation and Cleaning Software

- Ab Initio, provides high-performance software library and graphical environment for data transformation
- AMADEA, data Extraction, Transformation, and Real Time Reporting software
- BioComp iManageData(tm), Accesses, cleans, filters, converts and transforms data from files, Excel, Oracle, SQL Server, process control systems and more.
- ChoiceMaker 2.2 data quality and database record matching, merging, & deduplication software based on patented AI and machine learning techniques.
- COMGEN - Disk, tape and data conversion and data recovery experts, Commercial and General Systems.
- Data Manager, windows GUI application for data transformation and cleansing before data mining.
- DataFlux, provides Data Management solutions including Data profiling, Data quality, Data integration and Data augmentation
- Datatect, a powerful program for generating realistic test data to ASCII flat files or directly to RDBMS including Oracle, Sybase, SQL Server, and Informix.
- Dataskope, department-level tools to map, transform, alarm, output and view high volumes of binary or ASCII input data.
- DQ Now, profiling, cleansing, and dedup tools, providing a clear view of the data
- DQ Global, data cleansing, data management software, including de-duplication, merge/purge, address correction and suppression.
- GritBot, for identifying anomalies in data (compatible with See5 and Cubist).
- Hummingbird ETL, powerful data integration solution.
- IBM Datajoiner, allows you to view IBM, multi-vendor, relational, nonrelational, local, remote, and now geographic data as local and access and join tables without knowing the source location.
- MiningMart platform, for the preparation of relational data for Knowledge Discovery, free for research and non-commercial applications.
- NewView from SPSS
- proMISS, imputes missing values in databases.
- Relational Tools streamline application testing by allowing moving, editing and comparing referentially intact sets of complex relational data.
- Sagent, provides a suite of data transformation and loading tools
- Syncsort, fast high-volume sorting, filtering, reformatting, aggregating, and more
- The TrueData COMponent, functions to programmatically standardise your data, process it phonetically, and output a match key.
- WinPure, powerful data cleaning software, including duplication removal, email suggestions, statistics and more.

KDnuggets recomendations for data transformation and cleaning software

More at

- http://www.kdnuggets.com/software/index.html

Data Cleaning Tools – za Handbook of Data mining

Table 2.1. Industrial data cleansing tools circa 2004

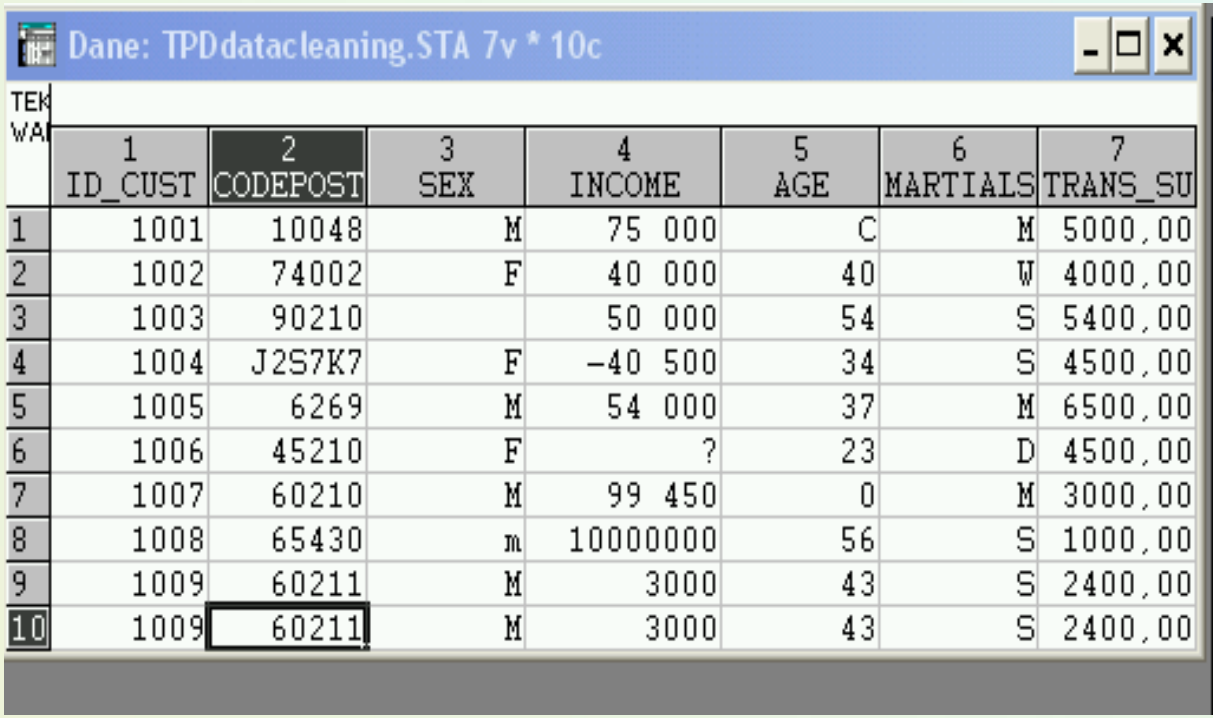
Tool	Company
Centrus Merge/Purge	*Qualitative Marketing Software*, http://www.qmsoft.com/
Data Tools Twins	*Data Tools*, http://www.datatools.com.au/
DataCleanser DataBlade	*Electronic Digital Documents*, http://www.informix.com
DataSet V	*iNTERCON http://www.ds-dataset.com*
DeDuce	*The Computing Group*
DeDupe	*International Software Publishing*
dfPower	*DataFlux Corporation*, http://www.dataflux.com/
DoubleTake	*Peoplesmith*, http://www.peoplesmith.com/
ETI Data Cleanse	*Evolutionary Technologies Intern*, http://www.evtech.com
Holmes	*Kimoce*, http://www.kimoce.com/
i.d.Centric	*firstLogic*, http://www.firstlogic.com/
Integrity	*Vality*, http://www.vality.com/
matchIT	*helpIT Systems Limited*, http://www.helpit.co.uk/
matchMaker	*Info Tech Ltd*, http://www.infotech.ie/
NADIS Merge/Purge Plus	*Group1 Software*, http://www.g1.com/
NoDupes	*Quess Inc*, http://www.quess.com/nodupes.html
PureIntegrate	*Carleton*, http://www.carleton.com/products/View/index.htm
PureName PureAddress	*Carleton*, http://www.carleton.com/products/View/index.htm
QuickAdress Batch	*QAS Systems*, http://207.158.205.110/
reUnion and MasterMerge	*PitneyBowes*, http://www.pitneysoft.com/
SSA-Name/Data Clustering Engine	*Search Software America* http://www.searchsoftware.co.uk/
Trillium Software System	*Trillium Software*, http://www.trilliumsoft.com/
TwinFinder	*Omikron*, http://www.deduplication.com/index.html
Ultra Address Management	*The Computing Group*

Integracja danych

- Spojrzenie z punktu widzenia hurtowni danych:
  - Eksploracja danych jako bardziej zaawansowany krok analizy niż OLAP,
  - Lecz pamiętaj, że odkrywanie wiedzy nie musi wykorzystywać bezpośrednio hurtowni danych.
  - Pomimo tego można wykorzystać metody wspólne z projektowaniem i konstruowaniem hurtowni danych
- Spójrz na dodatkowy plik z kopia tradycyjnych slajdów

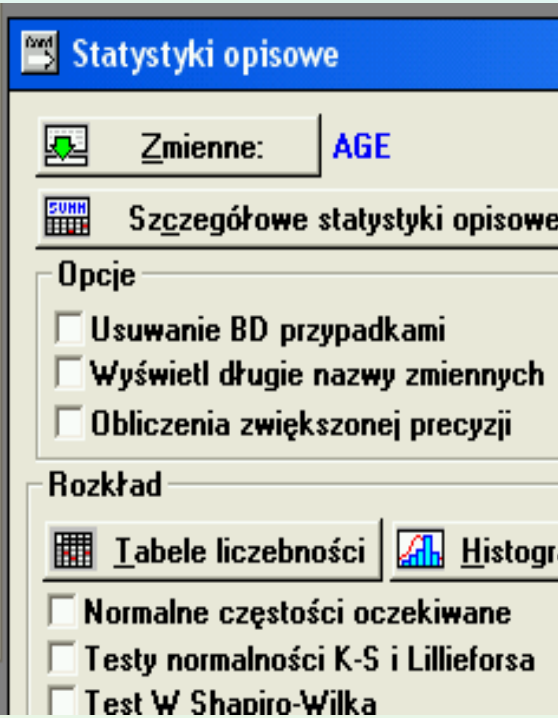
Niepoprawne / błędne wartości

- Co zauważasz podejrzanego w tym fragmencie danych?



Dane: TPDdatacleaning.STA 7v * 10c

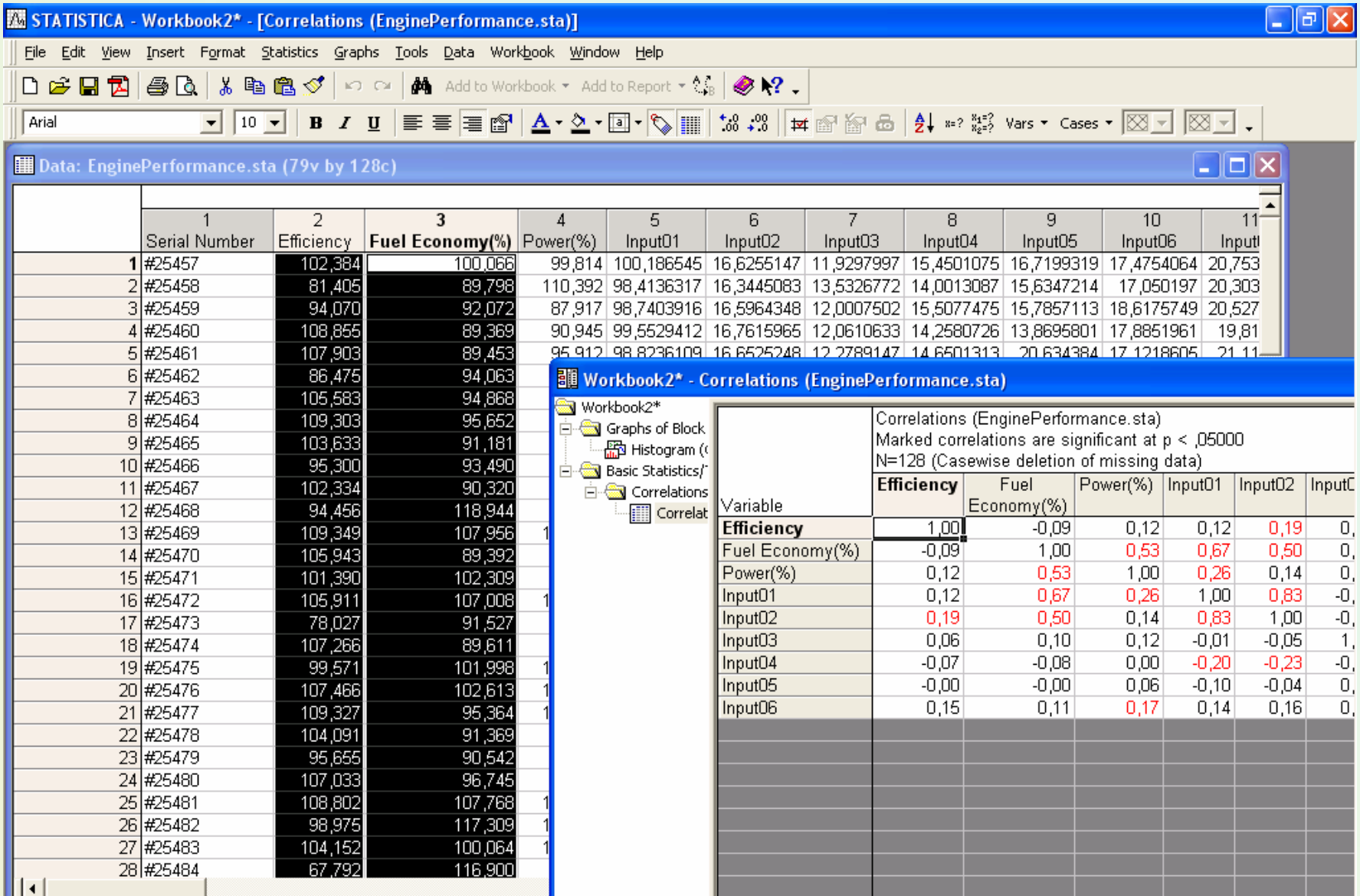
	1 ID_CUST	2 CODEPOST	3 SEX	4 INCOME	5 AGE	6 MARTIALS	7 TRANS_SU
1	1001	10048	M	75 000	C	M	5000,00
2	1002	74002	F	40 000	40	W	4000,00
3	1003	90210		50 000	54	S	5400,00
4	1004	J2S7K7	F	-40 500	34	S	4500,00
5	1005	6269	M	54 000	37	M	6500,00
6	1006	45210	F	?	23	D	4500,00
7	1007	60210	M	99 450	0	M	3000,00
8	1008	65430	m	10000000	56	S	1000,00
9	1009	60211	M	3000	43	S	2400,00
10	1009	60211	M	3000	43	S	2400,00



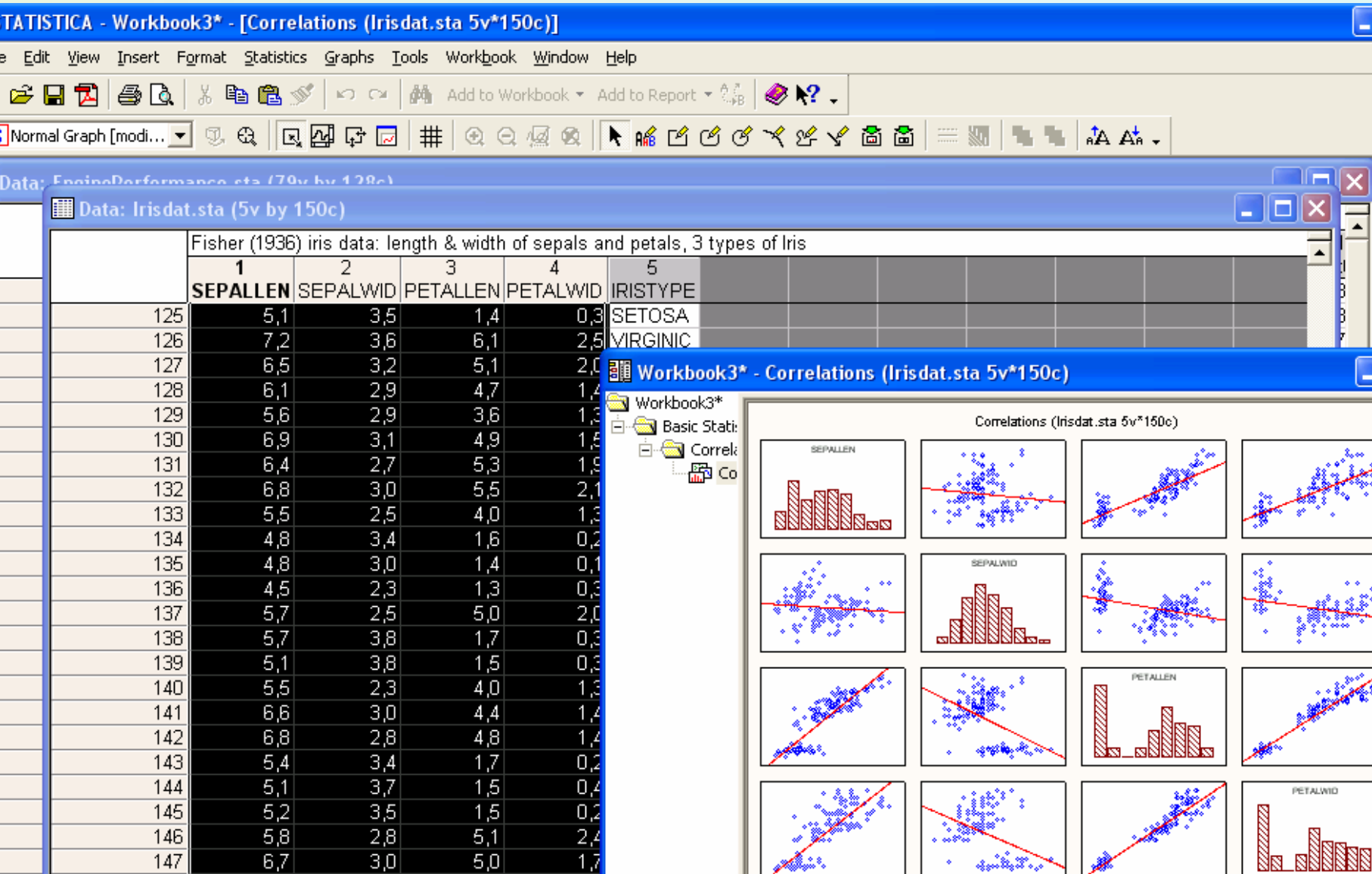
Duplicate Data – duplikaty rekordów

- Dane mogą zawierać informacje o tych samych lub prawie tak samo opisanych obiektach
- Przykłady:
  - Ta sama osoba z kilkoma adresami emaila
- Data cleaning
  - Usuń duplikaty
  - Lecz bądź ostrożny – czy zawsze?

Zbyt silna zależność między kolumnami – analiza korelacji

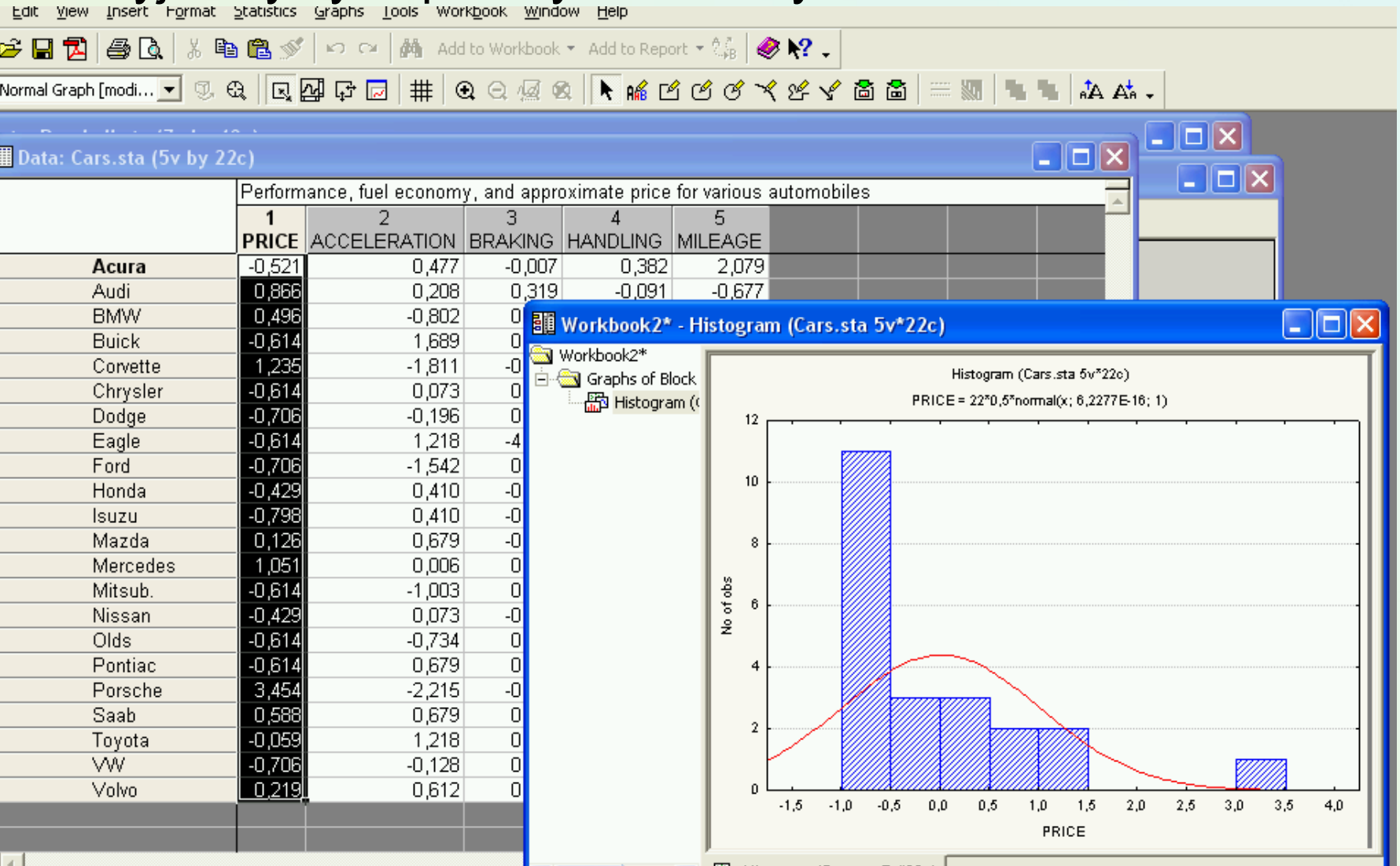


Scatterplot matrix / Macierz wykresów rozrzutu



Outliers – obserwacje oddalone

- Użyj statystyk opisowych oraz wykresów - Statistica

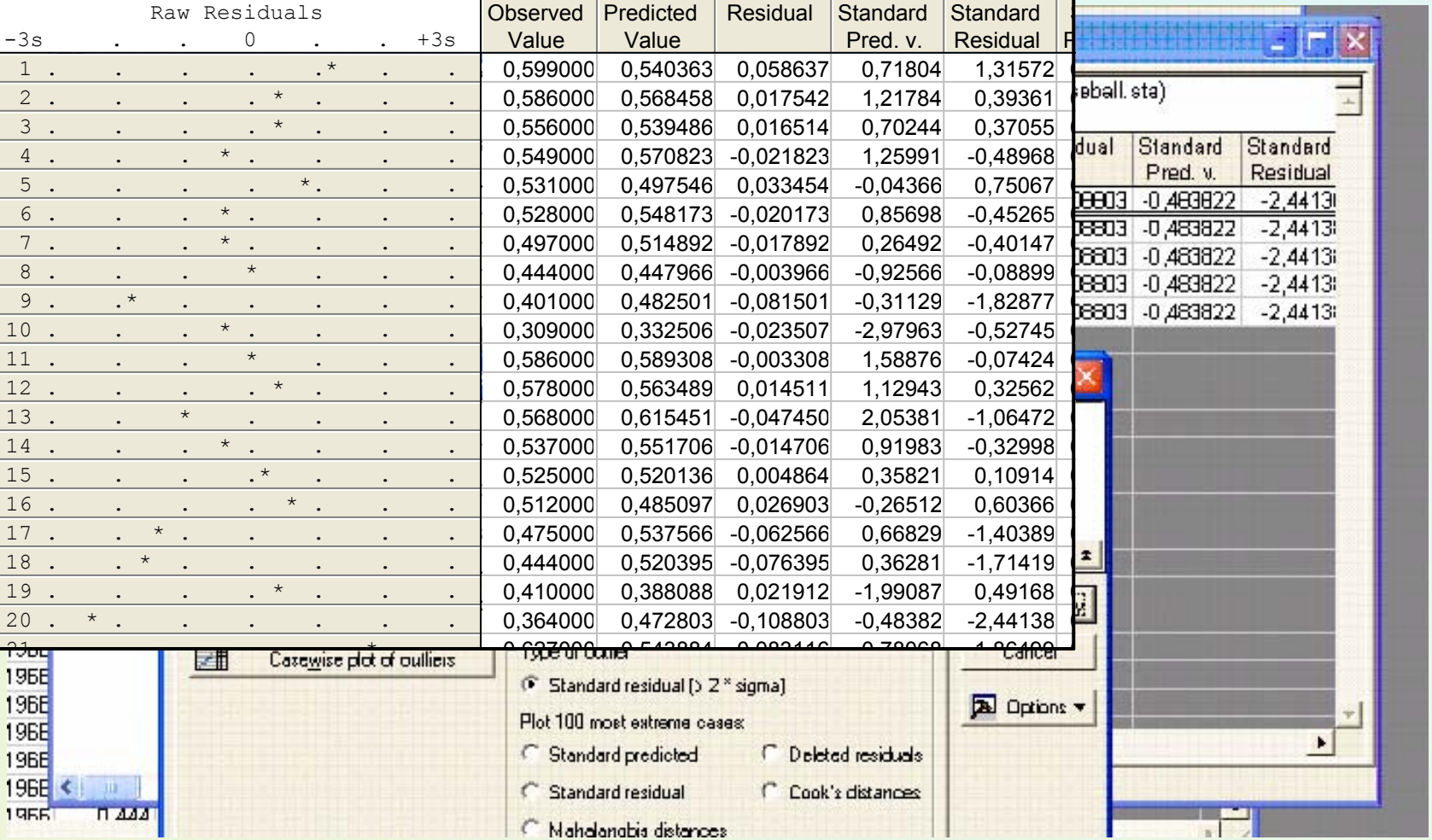


„Outiers” w wielowymiarowej regresji - analiza reszt

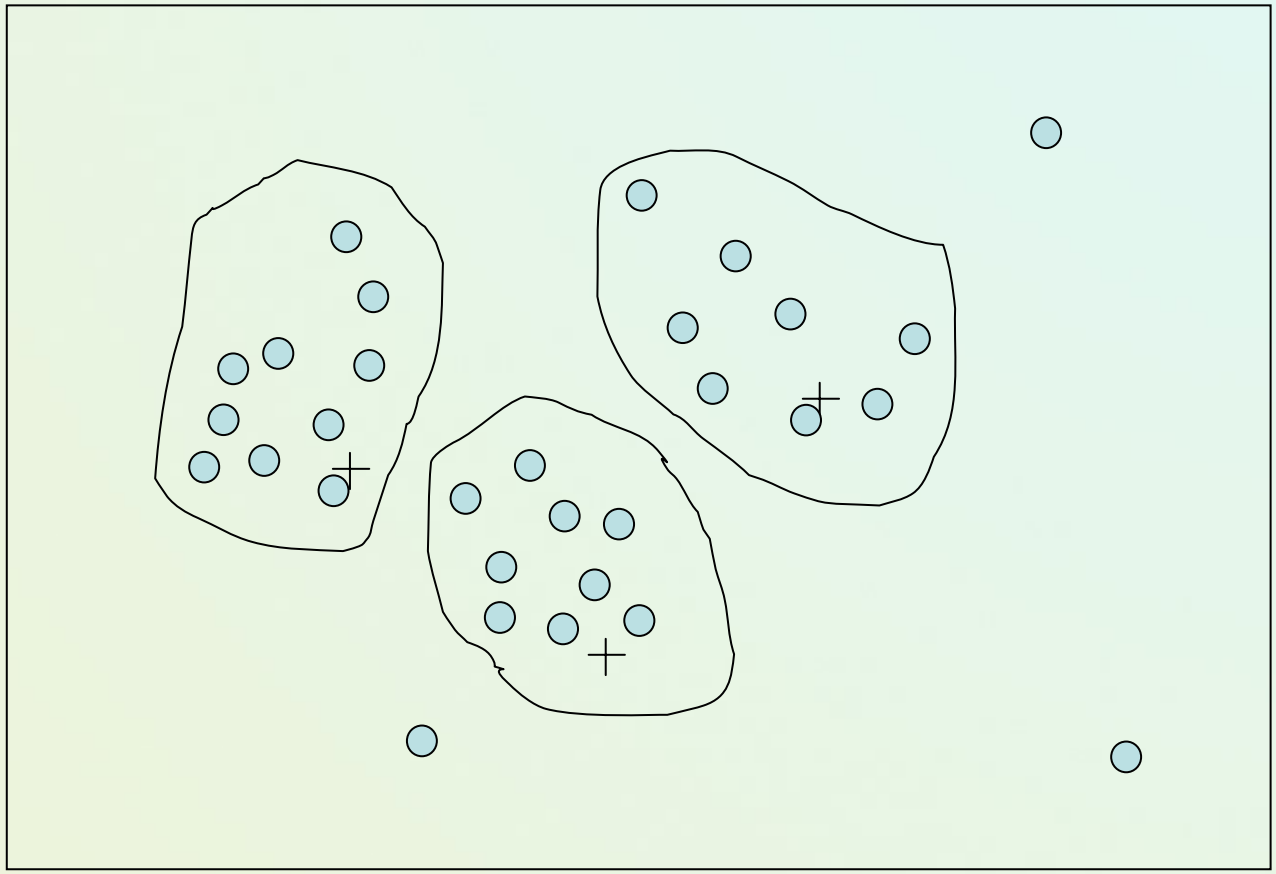
→ |standaryzowane reszty| ≤ 2

Raw Residual (Baseball.sta)
Dependent variable: WIN

Case	Raw Residuals -3s . . . 0 . . +3s	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual
1	. . . . .* . .	0,599000	0,540363	0,058637	0,71804	1,31572
2	. . . . * . . .	0,586000	0,568458	0,017542	1,21784	0,39361
3	. . . . * . . .	0,556000	0,539486	0,016514	0,70244	0,37055
4	. . . *	0,549000	0,570823	-0,021823	1,25991	-0,48968
5	. . . . *. . .	0,531000	0,497546	0,033454	-0,04366	0,75067
6	. . . *	0,528000	0,548173	-0,020173	0,85698	-0,45265
7	. . . *	0,497000	0,514892	-0,017892	0,26492	-0,40147
8	. . . *	0,444000	0,447966	-0,003966	-0,92566	-0,08899
9	. .*	0,401000	0,482501	-0,081501	-0,31129	-1,82877
10	. . . *	0,309000	0,332506	-0,023507	-2,97963	-0,52745
11	. . . *	0,586000	0,589308	-0,003308	1,58876	-0,07424
12	. . . . * . . .	0,578000	0,563489	0,014511	1,12943	0,32562
13	. . *	0,568000	0,615451	-0,047450	2,05381	-1,06472
14	. . . *	0,537000	0,551706	-0,014706	0,91983	-0,32998
15	. . . .* . . .	0,525000	0,520136	0,004864	0,35821	0,10914
16	. . . . * . . .	0,512000	0,485097	0,026903	-0,26512	0,60366
17	. . *	0,475000	0,537566	-0,062566	0,66829	-1,40389
18	. . *	0,444000	0,520395	-0,076395	0,36281	-1,71419
19	. . . . * . . .	0,410000	0,388088	0,021912	-1,99087	0,49168
20	. *	0,364000	0,472803	-0,108803	-0,48382	-2,44138



„Outliers” w analizaie skupień



Różne grupy rozwiązań

- Statystyczne (podstawowe statystyki + wizualizacje).
- Grupowanie (koszty obliczeniowe!).
- Pattern based identification (znajdź obserwacje, które nie potwierdzają wcześniej odnalezionych wzorców).
  - Reguły asocjacyjne

Przykłady wstępnej statystycznej analizy danych

- Prof. M.Lasek „Data mining” – banking data
- Larose D. Odkrywanie wiedzy z danych, PWN.
- Stanisz. Przystępny kurs statystyki (3 tomy), Statsoft.
- U. Fayyad, G.Gristen, A.Wierse, Information Visulization in Data Mining and Knowledge Discovery, Morgan Kaufmann Publisher.

Nieznane wartości atrybutów

Sposoby uwzględniania brakujących wartości:

- Stosowane w przetwarzaniu wstępnych (przekształć niekompletne dane w kompletne).
- Zintegrowane z algorytmami odkrywania wiedzy

Przetwarzanie wstępne:

- **Podejście naiwne**:
  - Zignorowanie przykładów opisanych nieznanymi wartościami.
- **Zastępowanie brakujących wartości poprzez**:
  - Użycie globalnej stałej wartości.
  - Zastąpienie najczęściej występującą wartością atrybutu nominalnego.
  - Zastąpienie wartością średnią atrybutu liczbowego.
  - Użycie najczęstszej lub średniej wartości atrybutu znajdowanej na podstawie rozkładu wartości wśród przykładów należących *tylko* do tej samej *klasy decyzyjnej* co analizowany przykład.
  - Użycie zbioru wszystkich możliwych wartości tego atrybutu.
  - Użycie podzbioru wartości atrybutu wraz z informacją o stopniach możliwości ich realizacji.
  - Wykonanie analizy zależności wartości atrybutu od atrybutów w pełni zdefiniowanych (regresja, drzewa i reguły decyzyjne).

„Closest Fit Approaches” [Grzymała 02]

- Definicja podobieństwa dla dwóch przypadków *e* i *e*’.

$$\sum_{i=1}^{n} \text{similarity}\,(e_i, e_i'),$$

where

$$\text{similarity}\,(e_i, e_i') = \begin{cases} 0 & \text{if } e_i \text{ and } e_i' \text{ are symbolic and } e_i \neq e_i', \text{ or } e_i = ? \text{ or } e_i' = ?, \\ 1 & \text{if } e_i = e_i', \\ 1 - \dfrac{|e_i - e_i'|}{|a_i - b_i|} & \text{if } e_i \text{ and } e_i' \text{ are numbers and } e_i \neq e_i', \end{cases}$$

Attributes		**Decision**
Abortions	Complications	Delivery
yes	none	fullterm
?	obesity	fullterm
no	alcoholism	fullterm
no	?	fullterm
yes	alcoholism	preterm

Porównanie kilku metod (regułowy klasyfikator LERS)

1. Most common value
2. Concept most com.
3. C4.5 method
4. All possible values
5. All in the concept
6. Ignoring tuples
7. Probabilistic even covering
8. LEM2 – omit
9. Treat as a special value.

Table 1. Description of data files

Name of Data Files	No. of Examples	No. of Attributes	No. of Concepts
Breast cancer	286	9	2
Echocardiogram	74	13	2
Hdynet	1218	73	2
Hepatitis	155	19	2
House	435	16	2
Im85	201	25	86
New-o	213	30	2
Primary tumor	339	17	21
Soybean	307	35	19
Tokt	6608	67	2

Table 2. Error rates of input data sets by using LERS new classification

	Methods								
Data file	1	2	3	4	5	6	7	8	9
Breast	34.62	34.62	31.5	28.52	31.88	29.24	34.97	33.92	32.52
Echo	6.76	6.76	5.4	—	—	6.56	6.76	6.76	6.76
Hdynet	29.15	31.53	22.6	—	—	28.41	28.82	27.91	28.41
Hepatitis	24.52	13.55	19.4	—	—	18.75	16.77	18.71	19.35
House	5.06	5.29	4.6	—	—	4.74	4.83	5.75	6.44
Im85	96.02	96.02	100	—	96.02	94.34	96.02	96.02	96.02
New-o	5.16	4.23	6.5	—	—	4.9	4.69	4.23	3.76
Primary	66.67	62.83	62.0	41.57	47.03	66.67	64.9	69.03	67.55
Soybean	15.96	18.24	13.4	—	4.1	15.41	19.87	17.26	16.94
Tokt	31.57	31.57	26.7	32.75	32.75	32.88	32.16	33.2	32.16

Missing and other absent values of attributes

- Wartości mogą być nieznane z różnych przyczyn:
- Różna semantyka nieznanych wartości atrybutów (ang. *unknown attribute values*):
  - brakujące wartości atrybutów (ang. *missing values*),
  - niedostępne wartości atrybutów (ang. *absent values*).
- In medical data, value for **Pregnant?** attribute for **Jane** or **Anna** is missing, while for **Joe** should be considered **Not applicable**
- Inna semantyka – absent values, don't care, i inne

Hospital Check-in Database

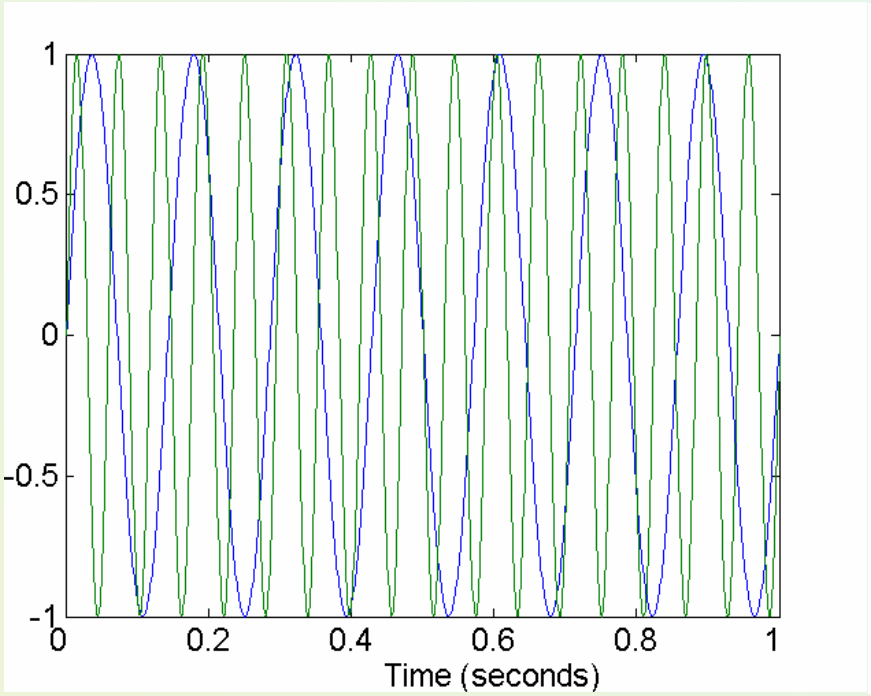
Name	Age	Sex	Pregnant	..
Mary	25	F	N	
Jane	27	F	?	
Joe	30	M	-	
Anna	2	F	?	

Agregacje atrybutów

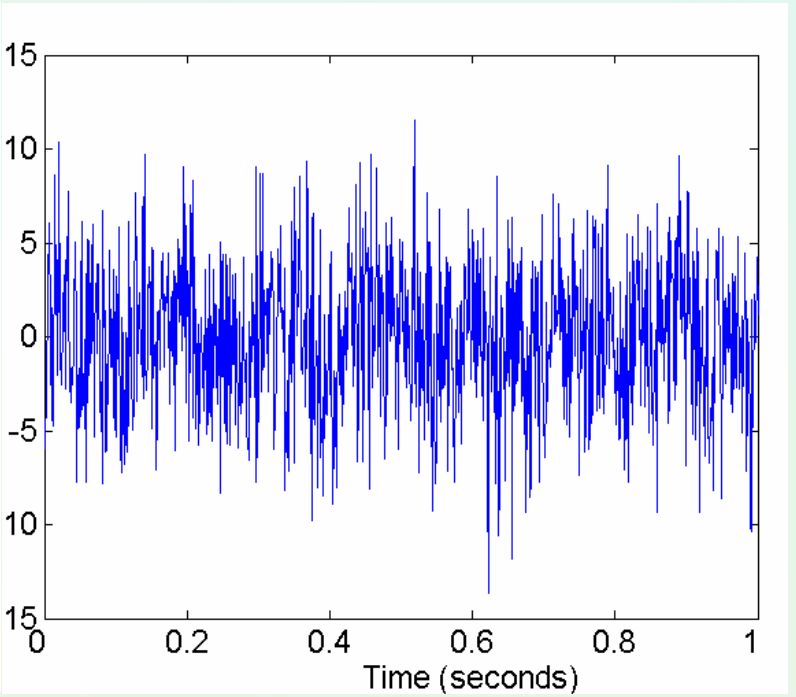
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Noise - szum

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



Two Sine Waves



Two Sine Waves + Noise

Szum w danych

- Przyczyny (?)
  - błędy urządzeń pobierających dane,
  - błędy transmisji,
  - błędy ludzkie,
  - ograniczenia technologiczne,
  - niespójności i niekonsekwencje nazewnictwa.
- Konsekwencje danych niekompletnych i sprzecznych.

Radzenie sobie z szumem

- Wygładzenie danych (ang. *smoothing techniques*).
- Tworzenie przedziałów (ang. *binning*).
- Algorytmy skupień i obiekty reprezentacji skupień
- Wykorzystywanie modeli predykcji (regresja)
- Konsultacja z użytkownikiem / ekspertem
- Inne …

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Normalizacja – transformacja danych

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

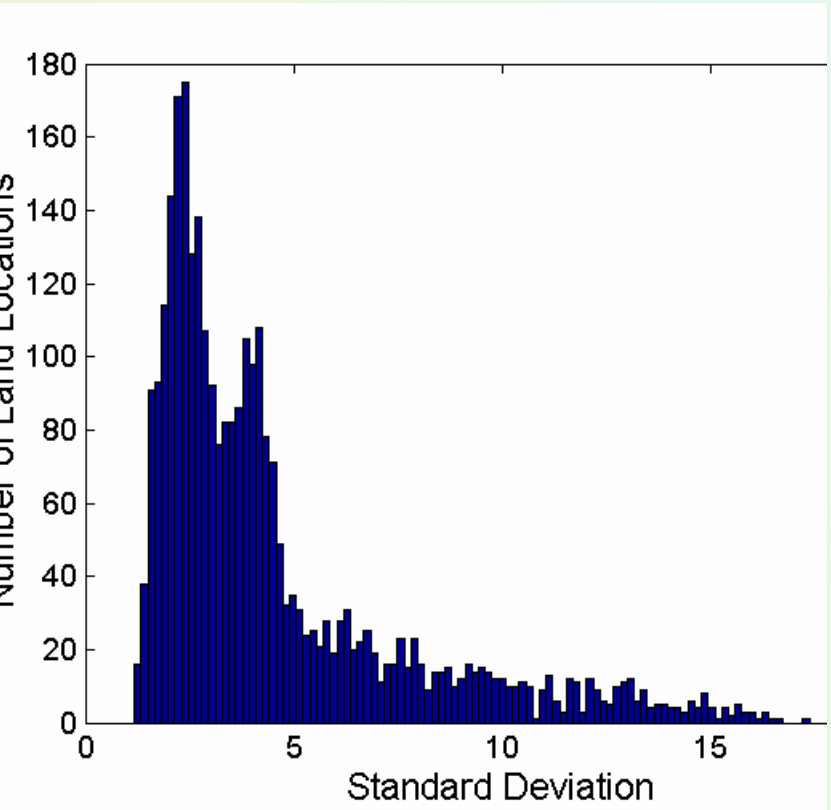
Where j is the smallest integer such that Max($|v'|$)<1

Transformacje dziedzin atrybutów liczbowych

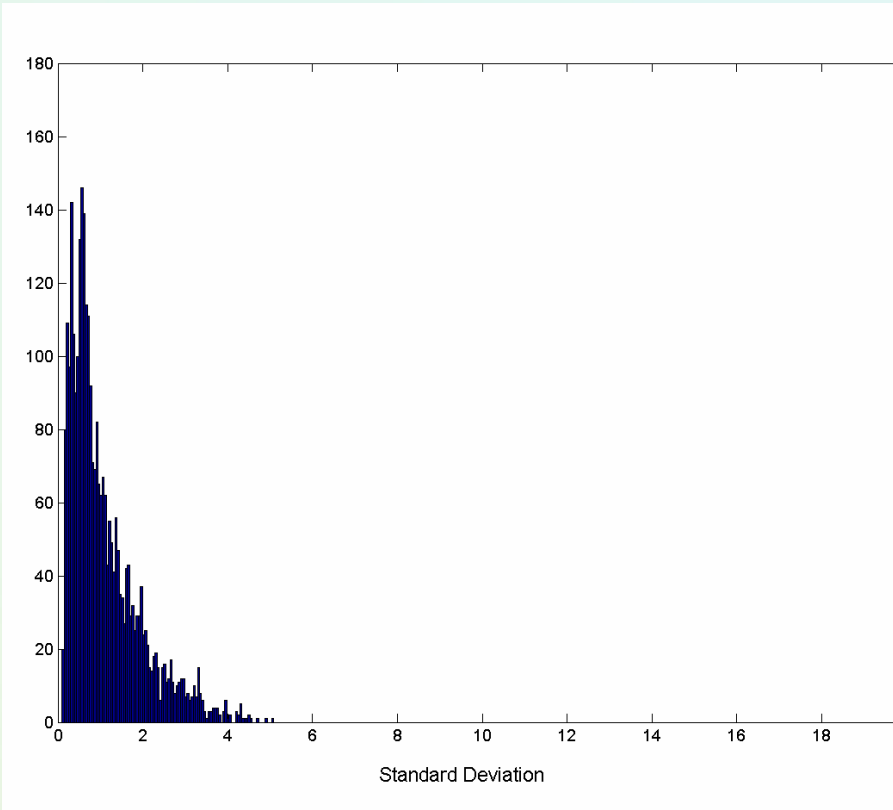
- Smoothing: remove „noise” from data
- Aggregation: summarization, …
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Discretization
- Attribute/feature construction
  - New attributes constructed from the given ones

Aggregation - przykład

Variation of Precipitation in Australia



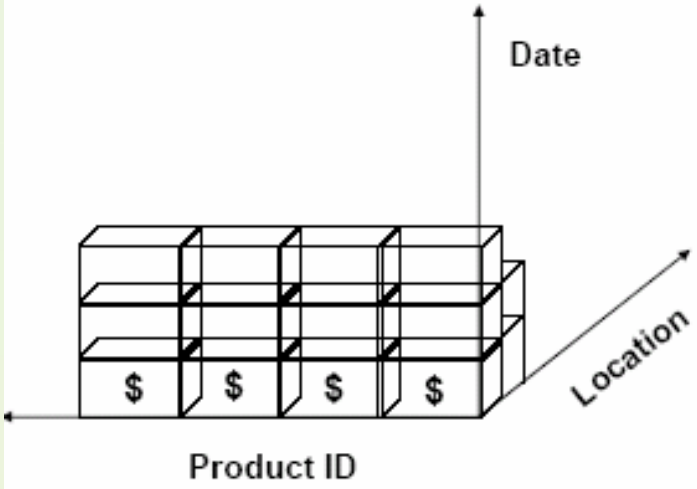
Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Agregacja w „kostkach danych” OLAP

- Combining two or more objects into a single object.



- Reduce the possible values of date from 365 days to 12 months.
- Aggregating the data per store location gives a view per product monthly.

Attribute “Location” is eliminated

Online Analytical Processing

(OLAP)

Transformacje danych: Dyskretyzacja

- Niektóre metody wymagają danych dyskretnych, np. Naïve Bayes, zbiory przybliżone, reguły asocjacyjne, wzorce sekwencji.
- Ponadto przydatne do podsumowania danych i redukcji rozmiarów.
- Dyskretyzacja jest:
  - procesem zamiany atrybutów liczbowych na atrybuty symboliczne typu porządkowego. Polega ona podziale oryginalnej dziedziny atrybutu liczbowego na pewną liczbę przedziałów i przypisaniu tym przedziałom kodów symbolicznych.

Przegląd w Handbook of Data Mining and Knowledge Discovery. Springer 2005

Chapter 6

DISCRETIZATION METHODS

Ying Yang
School of Computer Science and Software Engineering,
Monash University, Melbourne, Australia
yyang@mail.csse.monash.edu.au

Geoffrey I. Webb
Faculty of Information Technology
Monash University, Australia
geoff.webb@infotech.monash.edu

Xindong Wu
Department of Computer Science
University of Vermont, USA
xwu@cs.uvm.edu

Abstract Data-mining applications often involve quantitative data. However, learning from quantitative data is often less effective and less efficient than learning from qualitative data. Discretization addresses this issue by transforming quantitative data into qualitative data. This chapter presents a comprehensive introduction to discretization. It clarifies the definition of discretization. It provides a taxonomy of discretization methods together with a survey of major discretization methods. It also discusses issues that affect the design and application of discretization methods.

Keywords: Discretization, quantitative data, qualitative data.

Introduction

Discretization is a data-processing procedure that transforms quantitative data into qualitative data.

GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 47-58

Discretization Techniques: A recent survey

Sotiris Kotsiantis, Dimitris Kanellopoulos

Educational Software Development Laboratory
Department of Mathematics, University of Patras, Greece
sotos@math.upatras.gr, dkanellop@teipat.gr

Abstract. A discretization algorithm is needed in order to handle problems with real-valued attributes with Decision Trees (DTs), Bayesian Networks (BNs) and Rule-Learners (RLs), treating the resulting intervals as nominal values. The performance of these systems is tied to the right election of these intervals. A good discretization algorithm has to balance the loss of information intrinsic to this kind of process and generating a reasonable number of cut points, that is, a reasonable search space. This paper presents the well known discretization techniques. Of course, a single article cannot be a complete review of all discretization algorithms. Despite this, we hope that the references cited cover the major theoretical issues and guide the researcher to interesting research directions and suggest possible combinations that have to be explored.

Klasyfikacja metod

- Wiele różnorodnych podejść:
  - Nadzorowana vs. nienadzorowana,
  - Globalna vs. lokalna (z punktu widzenia atrybutów lub decyzji o zestawie przedziałów),
  - Dynamiczna vs. Statyczna (dobór parametrów).
  - Hierarchiczna vs. niehierarchiczna (przedziały dobiera się stopniowe)
  - Rozłączone vs. nakładające się przedziały

Wybrane metody

Table 6.2. Taxonomy of Discretization Methods

Method	Taxonomy (corresponding to Section 2)										
	0.	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Equal-width Equal-frequency Fixed-frequency	primary	unsupervised	parametric	non-hierarchical	univariate	disjoint	global	eager	time-insensitive	nominal	non-fuzzy
Multi-interval-entropy-minimization	primary	supervised	non-parametric	hierarchical	univariate	disjoint	global	eager	time-insensitive	nominal	non-fuzzy
ChiMerge StatDisc InfoMerge	primary	supervised	non-parametric	hierarchical	univariate	disjoint	global	eager	time-insensitive	nominal	non-fuzzy
Cluster-based	primary	unsupervised	non-parametric	hierarchical	multivariate	disjoint	global	eager	time-insensitive	nominal	non-fuzzy
ID3	primary	supervised	parametric	hierarchical	univariate	disjoint	local	eager	time-insensitive	nominal	non-fuzzy
Non-disjoint	composite	unsupervised	*	non-hierarchical	univariate	non-disjoint	global	eager	time-insensitive	nominal	non-fuzzy
Lazy	composite	*	*	*	univariate	non-disjoint	global	lazy	time-insensitive	nominal	non-fuzzy
Dynamic-qualitative	primary	unsupervised	non-parametric	non-hierarchical	univariate	disjoint	local	lazy	time-sensitive	nominal	non-fuzzy
Ordinal	composite	*	*	*	univariate	disjoint	global	eager	time-insensitive	ordinal	non-fuzzy
Fuzzy	composite	*	*	*	univariate	non-disjoint	global	eager	time-insensitive	nominal	fuzzy
Iterative-improvement	composite	supervised	*	hierarchical	multivariate	disjoint	global	eager	time-insensitive	nominal	non-fuzzy

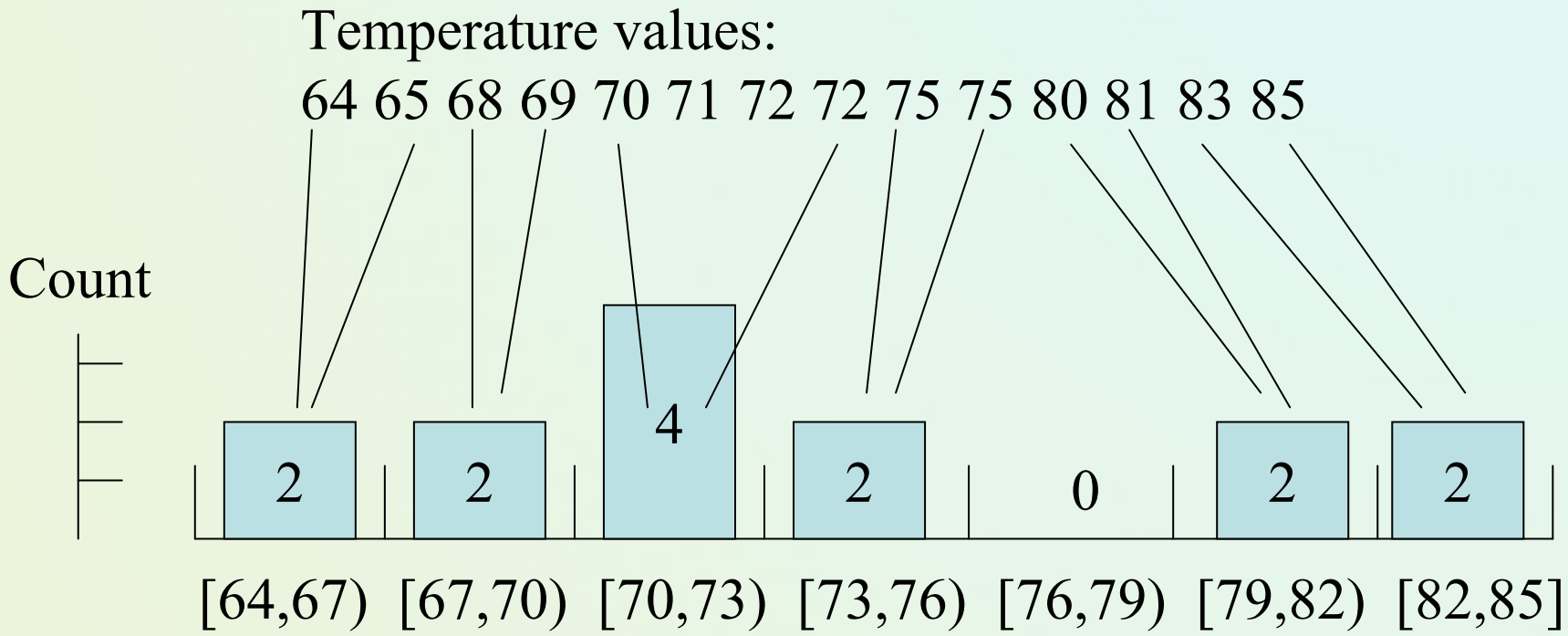
Note: each entry of the taxonomy is
0. primary *vs.* composite;
1. supervised *vs.* unsupervised;
2. parametric *vs.* non-parametric;
3. hierarchical *vs.* non-hierarchical;
4. univariate *vs.* multivariate;
5. disjoint *vs.* non-disjoint;
6. global *vs.* local;
7. eager *vs.* lazy;
8. time-sensitive *vs.* time-insensitive;
9. ordinal *vs.* nominal;
10. fuzzy *vs.* non-fuzzy.

An entry filled with '*' indicates that the corresponding method can be conducted in either way of the corresponding taxonomy entry. This often happens for composite methods, whose taxonomy depends on their primary methods.

Przykładowe popularne metody dyskretyzacji

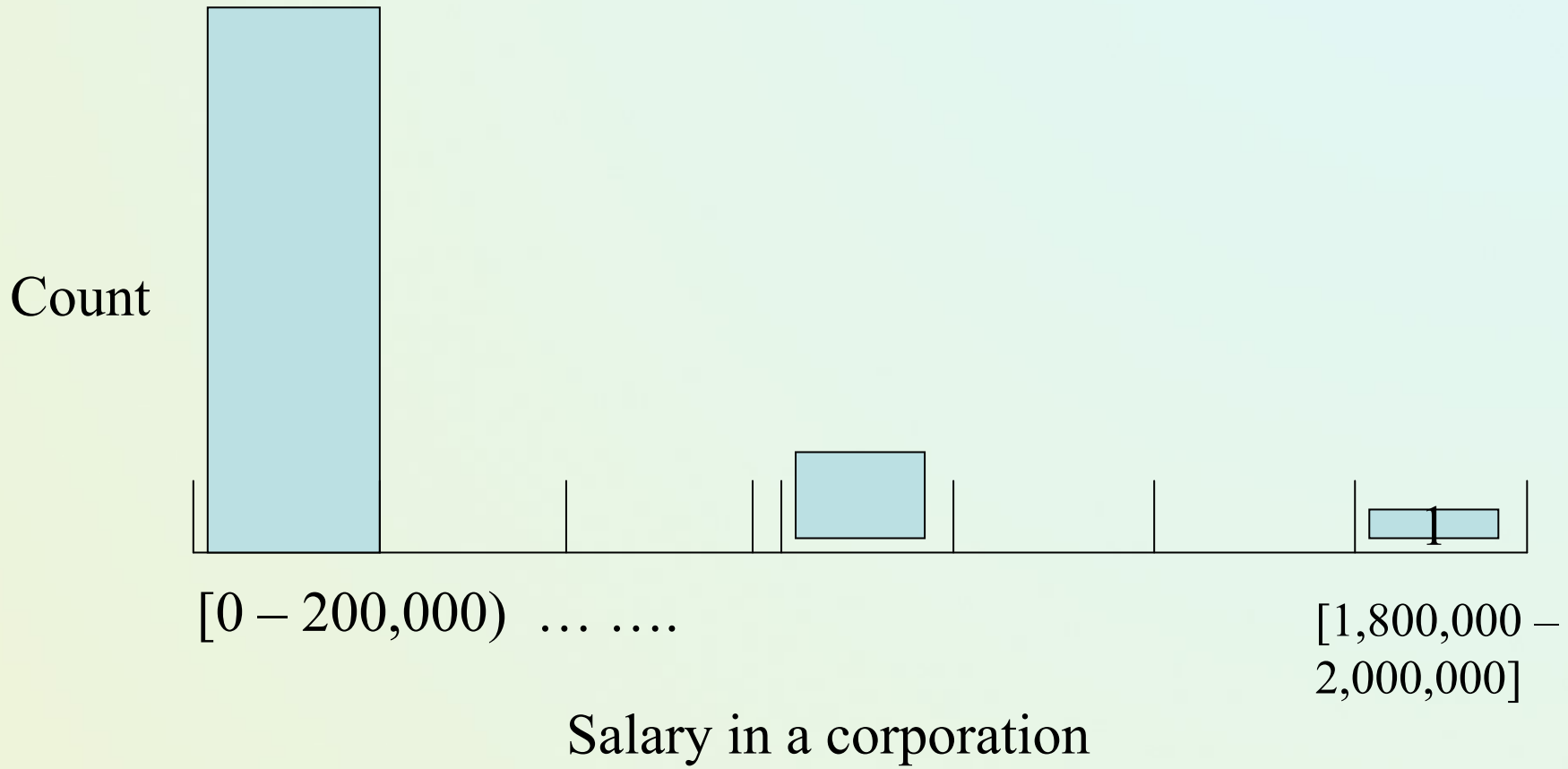
- Podział równymi przedziałami (*equal-width interval*)
  - Podziel zakres przedziału atrybutu na *N* podprzedziałów równej długości.
- Podział przedziałami o równej częstości (*equal-frequency interval*);
  - Podprzedziały zawierają w przybliżeniu taką samą liczbę obserwacji.
- *ChiMerge* –zachowuje podobieństwo względnych częstości klas decyzyjnych w podprzedziałach.
- Minimalizacja entropii warunkowej klas decyzyjnych (*Class Entropy discretization*);
  - Wersja lokalna, wersja wykorzystująca zasadę MDL, wersja globalizowana.
- Modyfikacje algorytmów analizy skupień (aglomeracyjne z warunkiem zatrzymania)

Discretization: Equal-Width (Length)



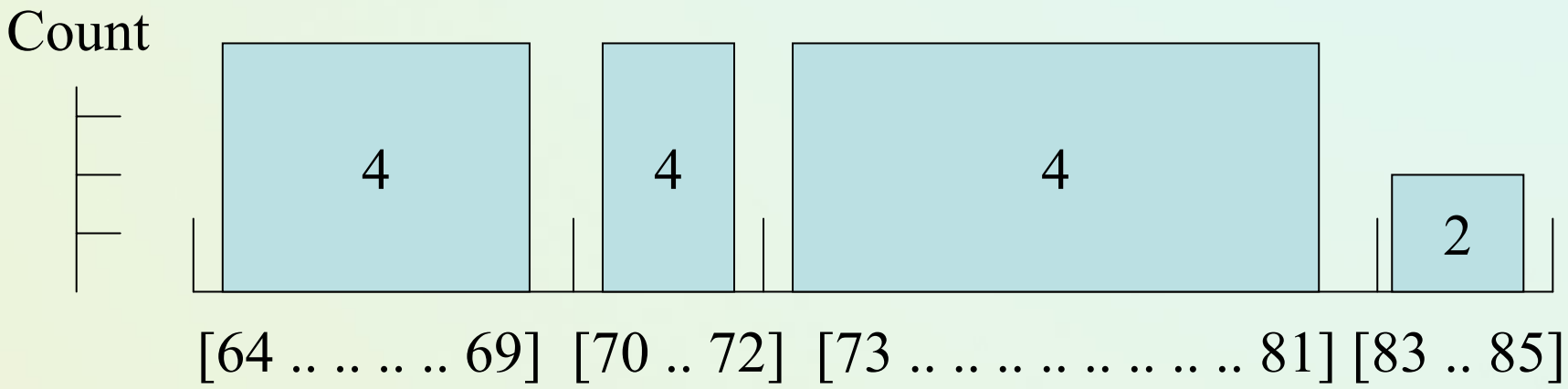
Equal Width, bins Low <= value < High

Discretization: Equal-Width may produce clumping



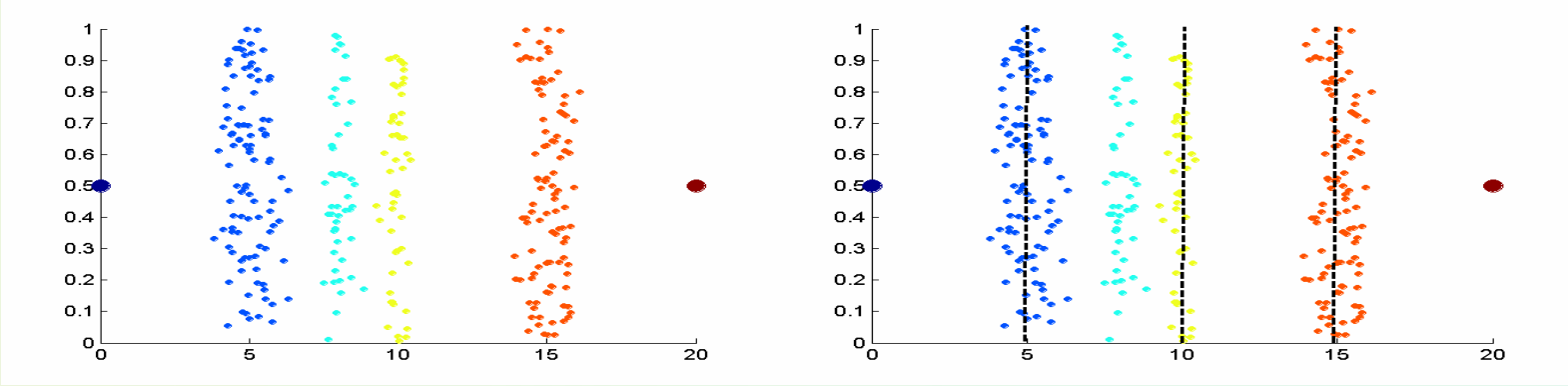
Discretization: Equal-Frequency

Temperature values:
64 65 68 69 70 71 72 72 75 75 80 81 83 85



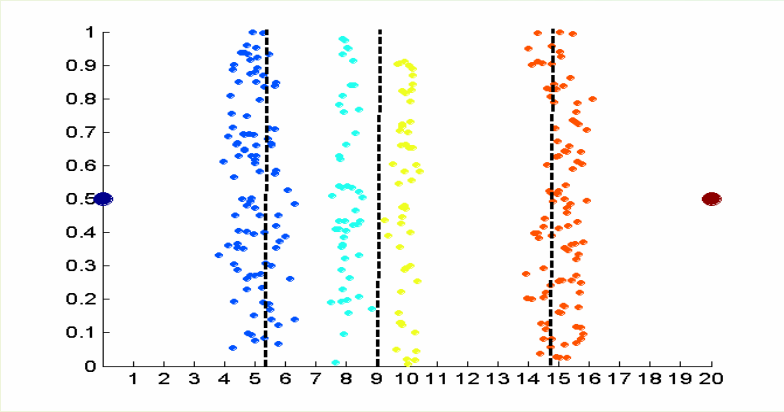
Equal Height = 4, except for the last bin

Dyskretyzacja nienadzorowana

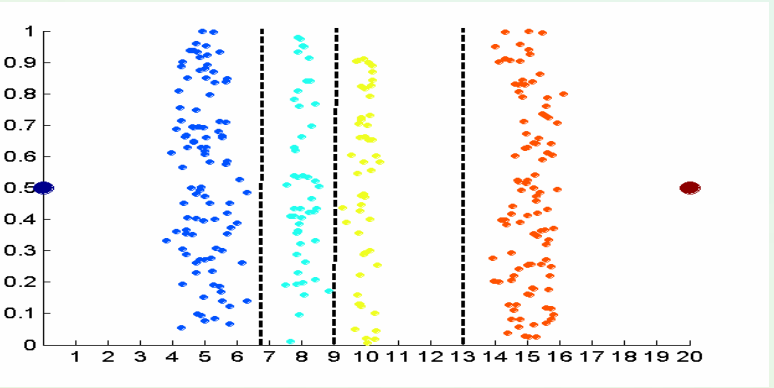


Data

Equal interval width



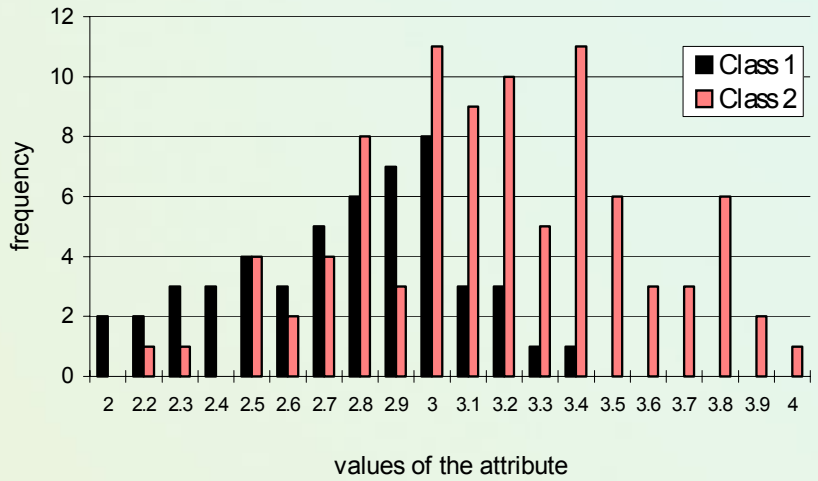
Equal frequency



K-means

Supervised (class) discretization

- Use information about attribute value distribution + class assignment.

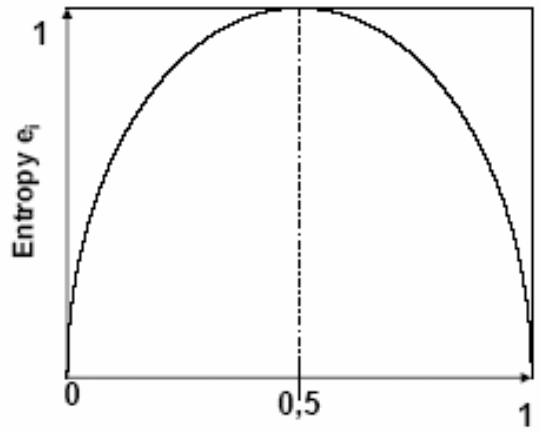


- Minimal entropy based approaches; Chi-Merge, others

Entropia informacji – przypomnienie właściwości

- Interpretacja entropii dla binarnych klas

$$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$



- Jak sytuacja jest najbardziej pożądana

Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two sub-intervals S1 and S2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

$$Ent(S) - E(T,S) > \delta$$

- Experiments show that it may reduce data size and improve classification accuracy

Przykład obliczeń

- Entropia przed podziałem $$Ent(S) = -\frac{3}{6}\cdot \lg\frac{3}{6} - \frac{3}{6}\cdot \lg\frac{3}{6} = 1$$
- Atrybut *IQ* i punkt graniczny T=107

105	107	107	109	113	115
yes	no	no	no	yes	yes

$$Ent(S \mid T) = \frac{1}{6}(-1\cdot \lg 1) + \frac{5}{6}(-\frac{3}{5}\cdot \lg\frac{3}{5} - \frac{2}{5}\lg\frac{2}{5}) = 0.811$$

- Inny punkt graniczny T=113
  Ent(S|T) = 0.541 - najlepszy możliwy.
- Właściwość Fayyad'a i Iraniego

Przykład dyskretyzacji [Grzymala 97]

Attributes			Decision
Caliber	Length	Weight	Recoil
5.56	45	55	light
6.5	55	120	light
6.5	55	142	medium
7	57	100	medium
7.5	55	150	medium
7.62	39	123	light
7.62	63	150	heavy
7.62	63	168	heavy
8	57	198	heavy

If (Weight=55) then (Decision=light)
If (Weight=120) then (Decision=light)
If (Weight=100) then (Decision=medium
If (Weight=142) then (Decision=medium

…

If (Length=63) then (Decision=heavy)

Attributes			Decision
Caliber	Length	Weight	Recoil
5.56..7.62	39..57	55..142	light
5.56..7.62	39..57	55..142	light
5.56..7.62	39..57	142..198	medium
5.56..7.62	57..63	55..142	medium
5.56..7.62	57..63	142..198	medium
7.62..8	39..57	55..142	light
7.62..8	57..63	142..198	heavy
7.62..8	57..63	142..198	heavy
7.62..8	57..63	142..198	heavy

If (length,39..57) & (weight,55..142) then (recoil,light)

If (caliber,5.56..7.62) & (weight,142..198) then (recoil,medium)

If (weight,55..142) & (length,57..63) then (recoil,medium)

If (caliber,7.62..8) & (length,57..63) then (recoil,heavy)

Porównanie różnych metod [Grzymała]

- Ocena eksperymentalna

Table 2. Accuracy Rate after Discretization

Data set	Equal interval width	Equal frequency per interval	Minimal class entropy	Cluster analysis
GM	68.0	59.0	73.0	69.0
rocks	57.5	54.2	55.6	53.0
iris	91.5	86.7	82.0	95.3
bank	77.3	95.5	84.9	97.0
hsv-r	42.5	35.8	46.7	48.3
bupa	41.9	39.7	41.3	42.5
glass	54.7	49.5	56.1	60.3
wave	99.4	99.4	99.4	99.8
image	69.0	70.0	73.8	77.6
cars	58.0	59.6	67.8	63.7

Dyskretyzacja ChiMerge (Kerber 92)

- Metoda lokalna, nadzorowana, dynamiczna.
  - Względna częstość przydziału obiektów do klas „wewnątrz” przedziału powinna być „w miarę” jednoznaczna, w przeciwnym razie przedział powinno podzielić się na podprzedziały.
  - Dwa sąsiednie przedziały nie powinny zawierać podobnego rozkładu częstości przydziału obiektów do klas decyzyjnych (w takim przypadku powinny być połączone).
- Wykorzystanie testu $\chi 2$ do badania zgodności rozkładów.
  - Jeśli zastosowanie testu $\chi 2$ wskazuje, że przydział obiektu do klas decyzyjnych jest niezależny od przedziału, to sąsiednie przedziały można połączyć.
  - Jeśli zastosowanie testu $\chi 2$ wskazuje, że przydział obiektu do klas decyzyjnych jest zależny od przedziału (tzn. różnica w rozkładach częstości klas w obu przedziałach jest statystycznie znacząca), to oba sąsiednie przedziały powinny pozostać niepołączone.

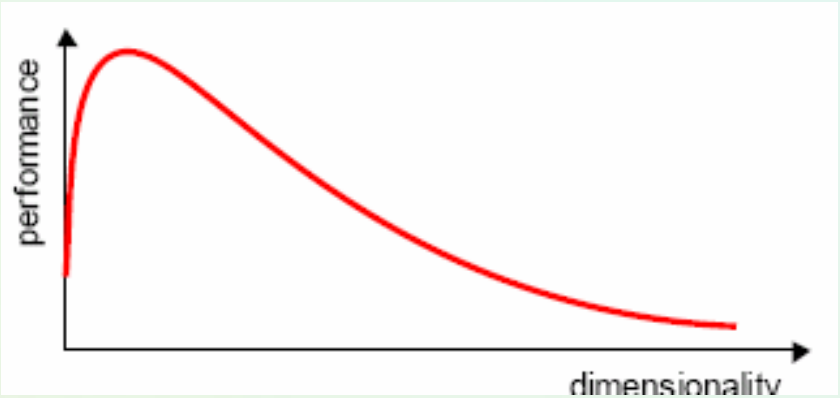
Przekleństwo wymiarowości

„Curse of dimensionality” [Bellman 1961]

- W celu dobrego przybliżenia funkcji (stworzenia klasyfikatora) z danych:
  - „the number of samples required per variable increases exponentially with the number of variables”
  - Liczba obserwacji żądanych w stosunku do zmiennej wzrasta wykładniczo z liczbą zmiennych.
- Oznacza to konieczność zdecydowanego wzrostu niezbędnych obserwacji przy dodawaniu kolejnych wymiarów.

Konsekwencje przekleństwa wymiarowości

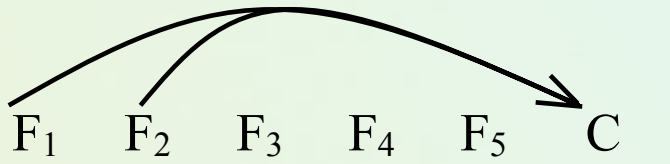
- For a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve!



Motywacje do redukcji danych

- Poprawa zdolności predykcyjnych
- Zwiększenie wydajności obliczeniowej
- Zmniejszenie wymagań dot. zbierania danych
  - Ułatwienia procedur rejestracji danych
  - Także koszty (np. testy diagnostyczne)
- Potencjalnie zmniejszają złożoność modelu (tzw. hypothesis complexity)
- Mogą zwiększyć czytelność reprezentacji
  - Inspekcja przez eksperta

Przykład problemu [D.Mladenic 2005]



F_1	F_2	F_3	F_4	F_5	C
0	0	1	0	1	0
0	1	0	0	1	1
1	0	1	0	1	1
1	1	0	0	1	1
0	0	1	1	0	0
0	1	0	1	0	1
1	0	1	1	0	1
1	1	0	1	0	1

- Data set
  - Five Boolean features
  - $C = F_1 \vee F_2$
  - $F_3 = \neg F_2$, $F_5 = \neg F_4$
  - Optimal subset:

    $\{F_1, F_2\}$ or $\{F_1, F_3\}$
- optimization in space of all feature subsets (2^F possibilities)

(tutorial on genomics [Yu 2004])

Przegląd w Handbook of Data Mining and Knowledge Discovery. Springer 2005

Chapter 5

DIMENSION REDUCTION AND FEATURE SELECTION

Barak Chizi
Tel-Aviv University

Oded Maimon
Tel-Aviv University

Abstract Data Mining algorithms search for meaningful patterns in raw data sets. The Data Mining process requires high computational cost when dealing with large data sets. Reducing dimensionality (the number of attributed or the number of records) can effectively cut this cost. This chapter focuses a pre-processing step which removes dimension from a given data set before it is fed to a data mining algorithm. This work explains how it is often possible to reduce dimensionality with minimum loss of information. Clear dimension reduction taxonomy is described and techniques for dimension reduction are presented theoretically.

Keywords: Dimension Reduction, Preprocessing

1. Introduction

Data Mining algorithms are used for searching meaningful patterns in raw data sets. Dimensionality (i.e., the number of data set attributes or groups of attributes) constitutes a serious obstacle to the efficiency of most Data Mining algorithms (Maimon and Last, 2000). This obstacle is sometimes known as the "curse of dimensionality" (Elder and Pregibon, 1996). Techniques quite efficient in low dimensions (e.g., nearest neighbors) cannot provide any meaningful results when the number of records goes beyond a 'modest' size of 10 attributes.

Data-mining algorithms are computationally intensive. Figure 5.1 describes the typical trade-off between the error rate of a Data Mining model and the cost of obtaining the model (in particular, the model may be a classification

Różne podejścia do redukcji wymiarów

- Selekcja cech, zmiennych (atrybutów, ..)
  - Wybierz podzbiór zmiennych $F' \subset F$
- Konstrukcja nowych cech
  - Metody projekcji – nowe cechy zastępują poprzednie; statystyczne PCA, dekompozycje macierzy
- Wykorzystanie wiedzy dziedzinowej
  - Wprowadzenie nowych cech (oprócz istniejących)
  - Indukcja konstruktywna

Selekcja cech vs. konstrukcja nowych cech

- Za A.Berge

- In general - two approaches for dimensionality reduction
  - Feature selection: choose a subset of the features

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_m} \end{bmatrix}$$

  - Feature extraction: create a subset of new features by combining existing features

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix}\right)$$

Proste podejście do selekcji atrybutów

1: Usuń kolumny o zbyt małej zmienności

Statystyczna miara zmienności Vx

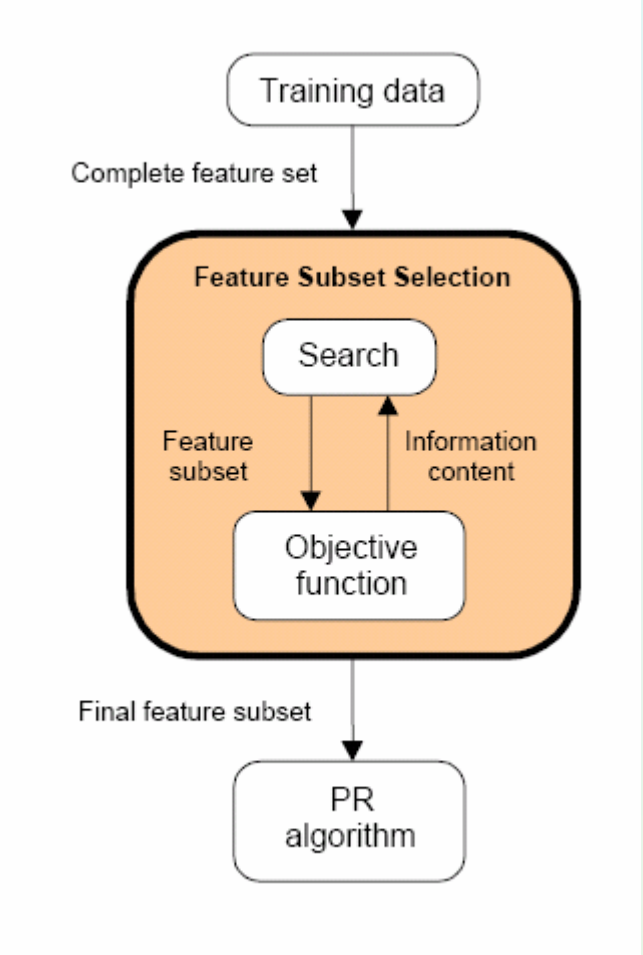
- Zbadaj liczbę różnych wartości w kolumnie
  - *Heurystyka:* pomiń kolumny zawierające taką samą lub prawie taką samą pojedynczą wartości (inne wartości $\leq$ minp).
  - *minp* może być mniej niż 5% przykładów / lub przykładów w najmniej licznej klasie.
- Bardziej wyrafinowane metody wykorzystują miary oceny znaczenia / informatywności atrybutów
  - WEKA zobacz attribute selection

Redukcja rozmiarów danych – Selekcja atrybutów

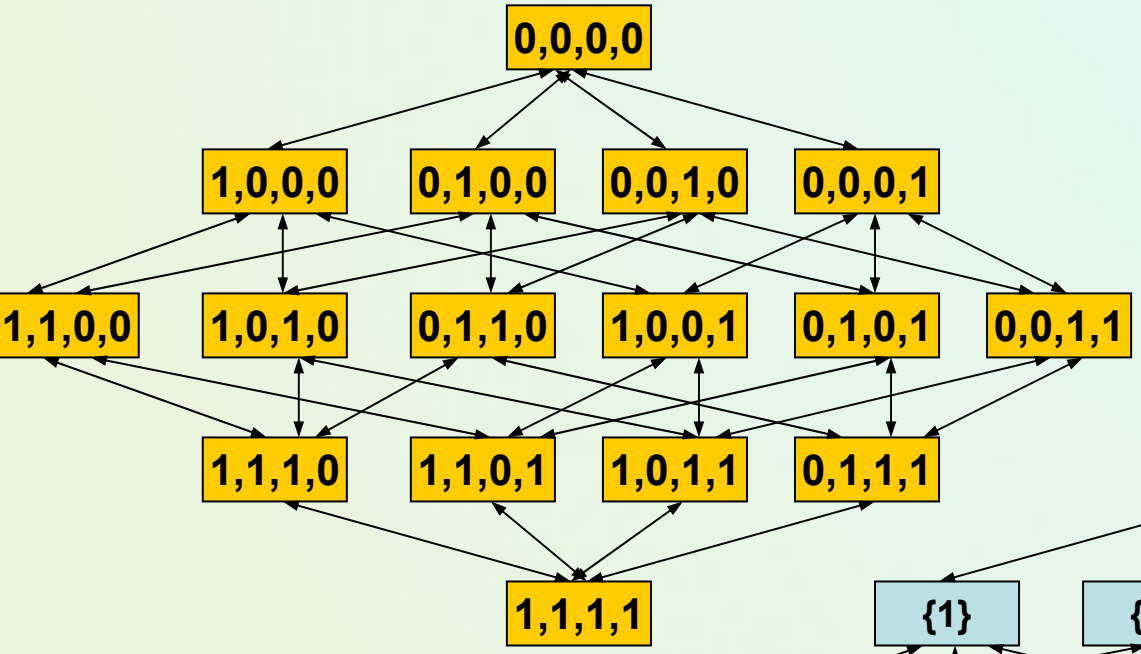
- Dany jest *n* elementowy zbiór przykładów (obiektów). Każdy przykład *x* jest zdefiniowany na $V1 \times V2 \times Vm$ gdzie *Vi* jest dziedziną *i*-tego atrybutu. W przypadku nadzorowanej klasyfikacji przykłady zdefiniowane są jako <*x*,*y*> gdzie *y* określa pożądaną odpowiedź, np. klasyfikację przykładu.
- **Cel selekcji atrybutów:**
  - *Wybierz minimalny podzbiór atrybutów, dla którego rozkład prawdopodobieństwa różnych klas obiektów jest jak najbliższy oryginalnemu rozkładowi uzyskanemu z wykorzystaniem wszystkich atrybutów.*
- **Nadzorowana klasyfikacja**
  - *Dla danego algorytmu uczenia i zbioru uczącego, znajdź najmniejszy podzbiór atrybutów dla którego system klasyfikujący przewiduje przydział obiektów do klas decyzyjnych z jak największą trafnością.*

Problem selekcji atrybutów (Feature Selection)

- Optymalne rozwiązanie – NP.
- *n* liczba atrybutów
  $\rightarrow$ przegląd przestrzeni z 2^n stanami.
- Przestrzeń podzbiorów częściowo uporządkowana (ang. lattice)
- Strategia przeszukiwania oraz
- Miara oceny *J*



Przeszukiwanie przestrzeni podzbiorów



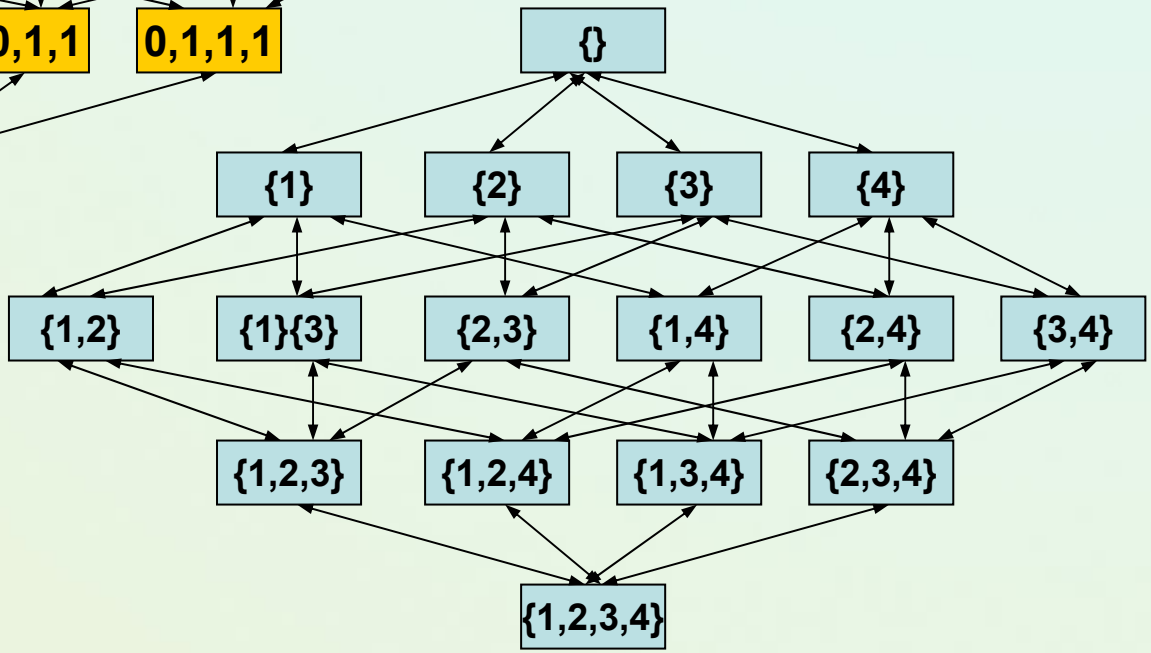
Subset Inclusion State Space

Poset Relation: Set Inclusion

***A* ≤ *B* = "B is a subset of A"**

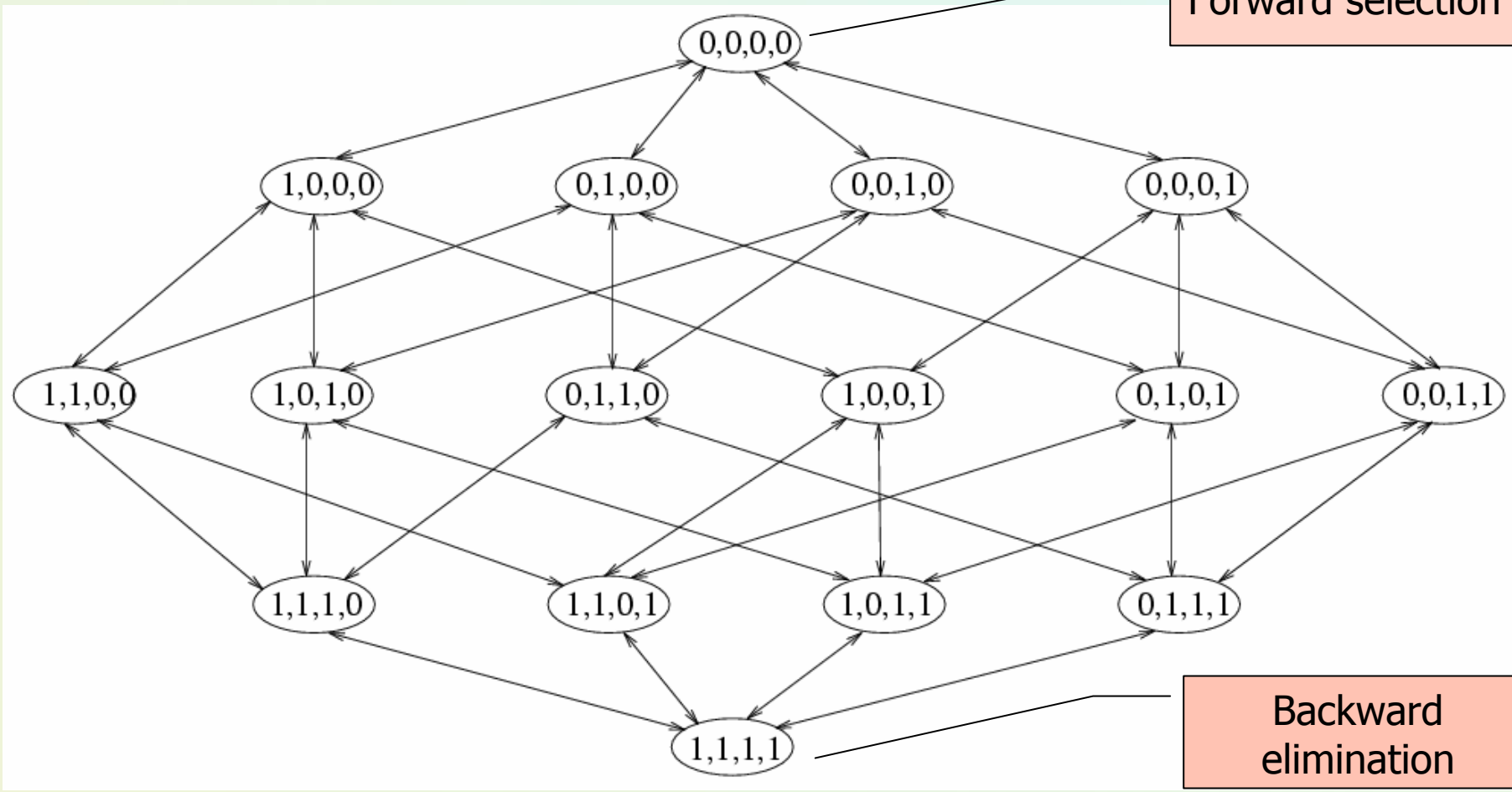
"Up" operator: DELETE

"Down" operator: ADD



Search for feature subset

- An example of search space (*John & Kohavi 1997*)



Feature subset selection

- commonly used search strategies:
  - **forward selection**
    - FSubset={}; greedily add features one at a time
  - **forward stepwise selection**
    - FSubset={}; greedily add or remove features one at a time
  - **backward elimination**
    - FSubset=AllFeatures; greedily remove features one at a time
  - **backward stepwise elimination**
    - FSubset=AllFeatures; greedily add or remove features one at a time
  - **random mutation**
    - FSubset=RandomFeatures;
    - greedily add or remove randomly selected feature one at a time
    - stop after a given number of iterations

Selekcja w trakcie wstępnego przetwarzania danych

- Ocena pojedynczych atrybutów:
  - testy $\chi 2$ i miary siły związku,
  - miary wykorzystujące względną entropie między atrybutem warunkowym a decyzyjnym (ang. *info gain*, *gain ratio*),
  - ...
- Ocena podzbiorów atrybutów (powinny być niezależne wzajemnie a silnie zależne z klasyfikacją):
  - Miara korelacji wzajemnych,
  - Statystyki λ Wilksa, T2-Hotellinga, odległości D2 Mahalanobisa,
  - Redukty w teorii zbiorów przybliżonych,
  - Techniki dekompozycji na podzbiory (ang. *data table templates*)
  - ...
- Model „filter” vs. „wrapper”

Inaczej o miarach oceny

Score *predictivness* of features, e.g.

Information theoretic analysis

- high information gain
- Breiman's Gini index (also a diversity/impurity idea)

Some other statistical tests

- e.g. chi-square statistic

Relief algorithm

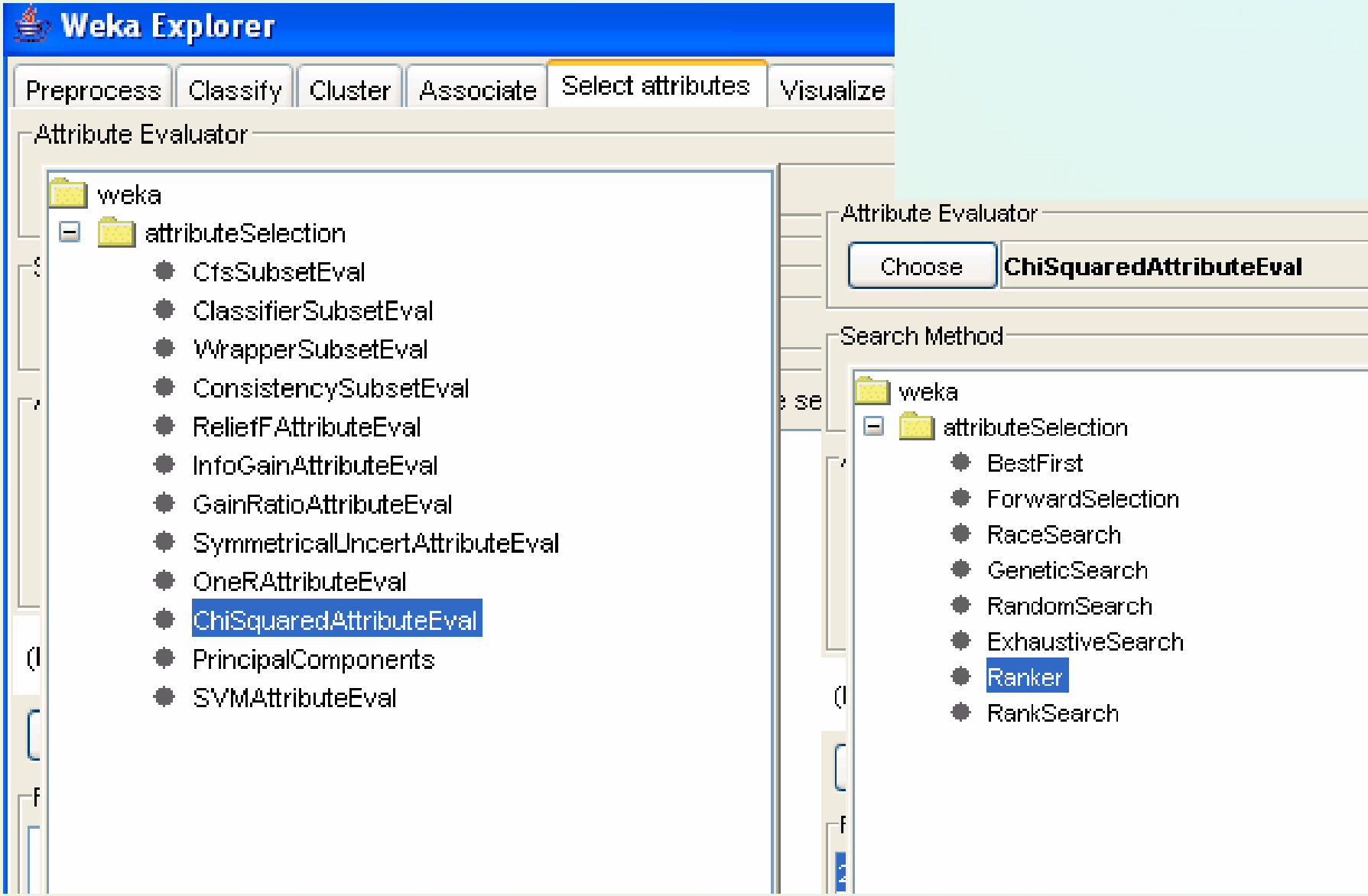
- Assign high scores to features that match on *near hits* and don't match on *near misses* (in the context of nearest neighbour classification) (Kira & Rendell, 1992)

A criterion for attribute selection

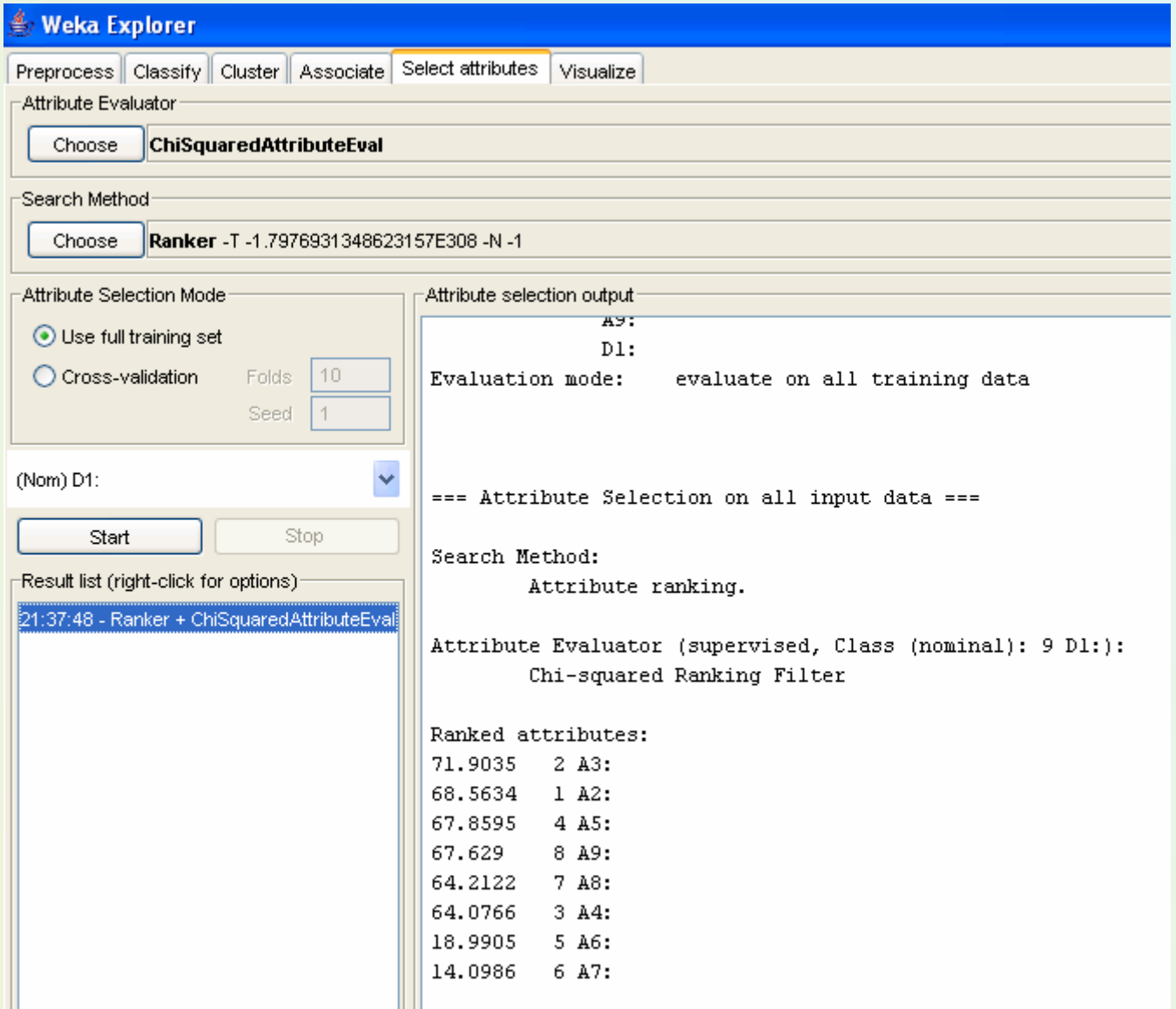
Impurity functions:

- Given a random variable *x* with *k* discrete values, distributed according to P={p1,p2,…pk}, a impurity function Φ should satisfies:
  - $\Phi(P) \geq 0$; $\Phi(P)$ is minimal if $\exists i$ such that pi=1;
    $\Phi(P)$ is maximal if $\forall i\ 1 \leq i \leq k$, pi=1/k
    $\Phi(P)$ is symmetrical and differentiable everywhere in its range
- The goodness of split is a reduction in impurity of the target concep after partitioning S.
- Popular function: *information gain*
  - Information gain increases with the average purity of the subsets that an attribute produces

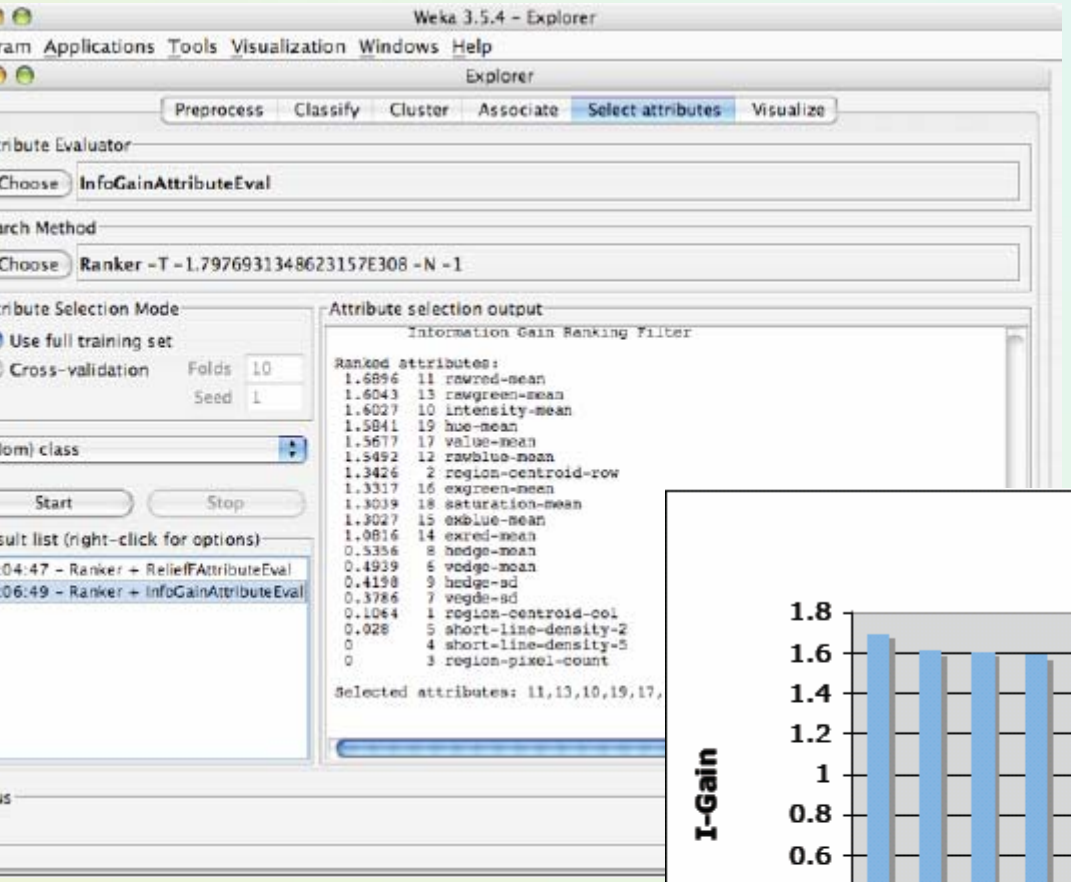
WEKA – attribute selection

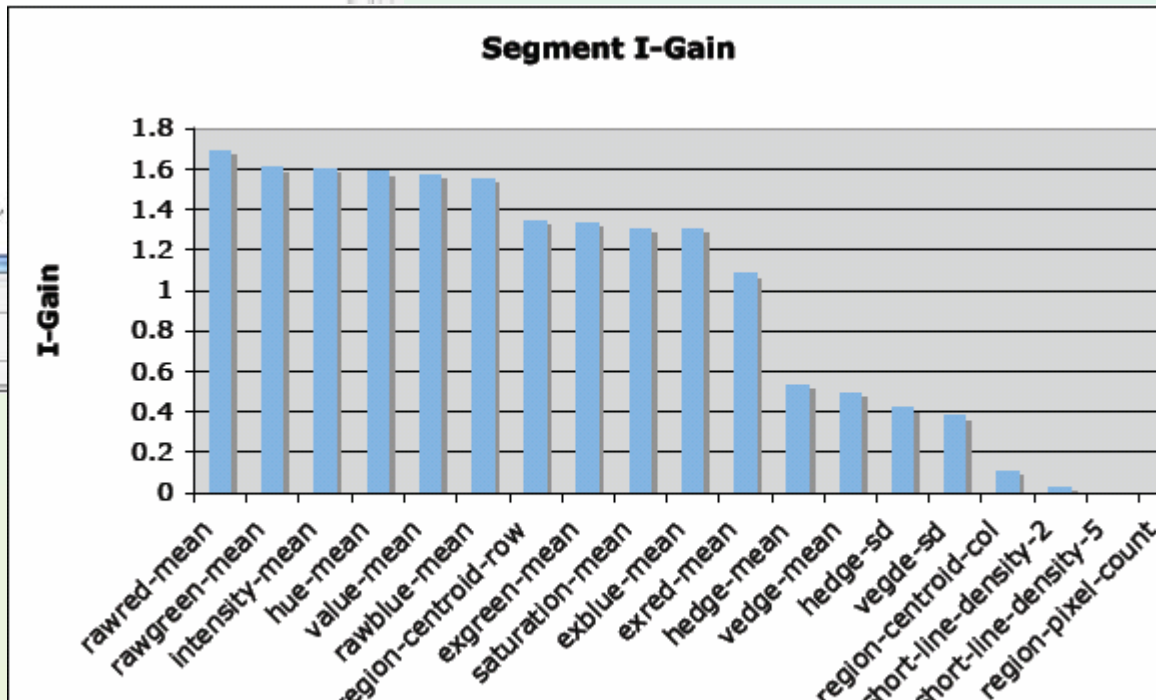


Ranking with …? WEKA



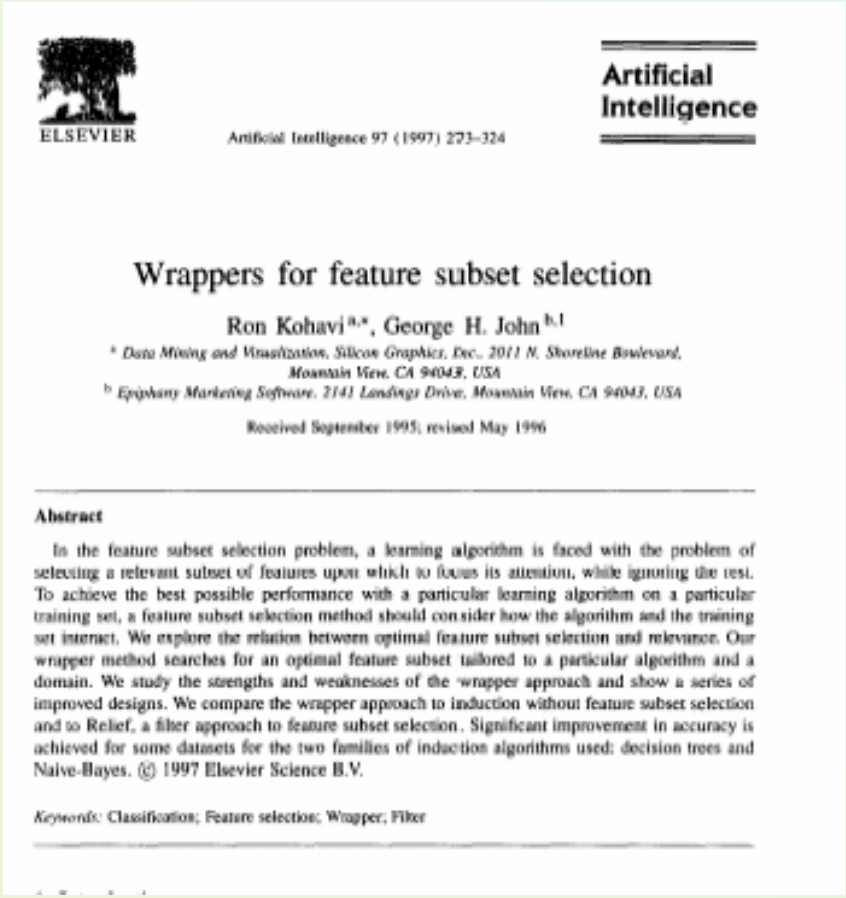
Jak wykorzystać ranking atrybutów





- Wybierz powyżej progu τ
- Mediana czy inny?

Inne podejścia



ELSEVIER

Artificial Intelligence 97 (1997) 273–324

Artificial Intelligence

Wrappers for feature subset selection

Ron Kohavi [a,*], George H. John [b,1]

[a] *Data Mining and Visualization, Silicon Graphics, Inc., 2011 N. Shoreline Boulevard, Mountain View, CA 94043, USA*

[b] *Epiphany Marketing Software, 2141 Landings Drive, Mountain View, CA 94043, USA*

Received September 1995; revised May 1996

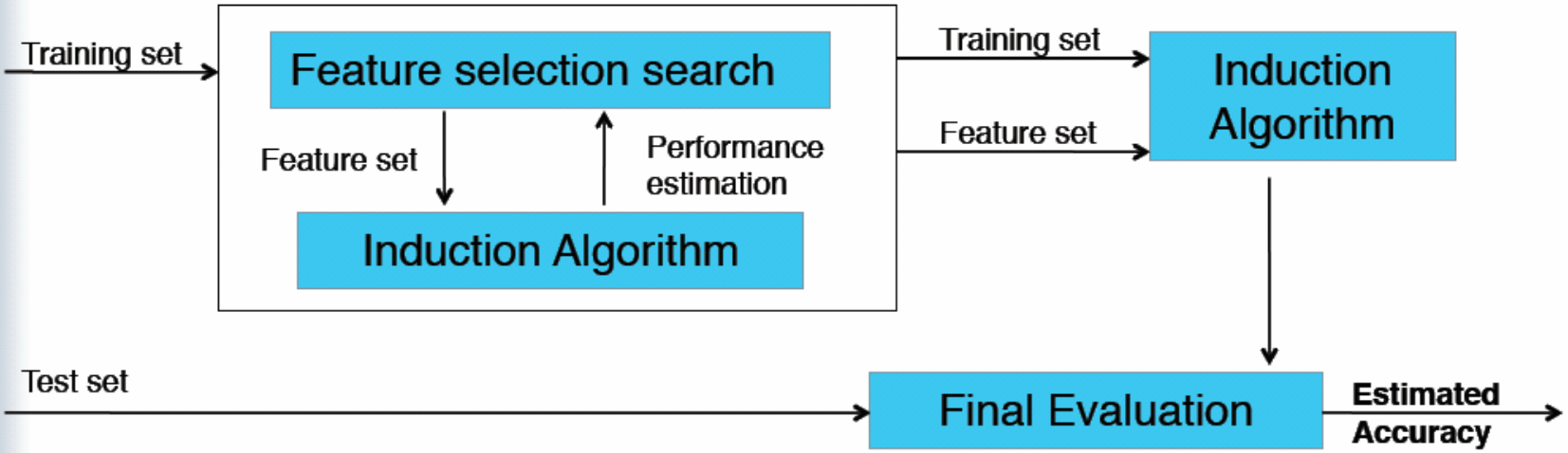
Abstract

In the feature subset selection problem, a learning algorithm is faced with the problem of selecting a relevant subset of features upon which to focus its attention, while ignoring the rest. To achieve the best possible performance with a particular learning algorithm on a particular training set, a feature subset selection method should consider how the algorithm and the training set interact. We explore the relation between optimal feature subset selection and relevance. Our wrapper method searches for an optimal feature subset tailored to a particular algorithm and a domain. We study the strengths and weaknesses of the wrapper approach and show a series of improved designs. We compare the wrapper approach to induction without feature subset selection and to Relief, a filter approach to feature subset selection. Significant improvement in accuracy is achieved for some datasets for the two families of induction algorithms used: decision trees and Naive-Bayes. © 1997 Elsevier Science B.V.

Keywords: Classification; Feature selection; Wrapper; Filter

Wrapper [Kohavi et al.]

Wrapper Approach



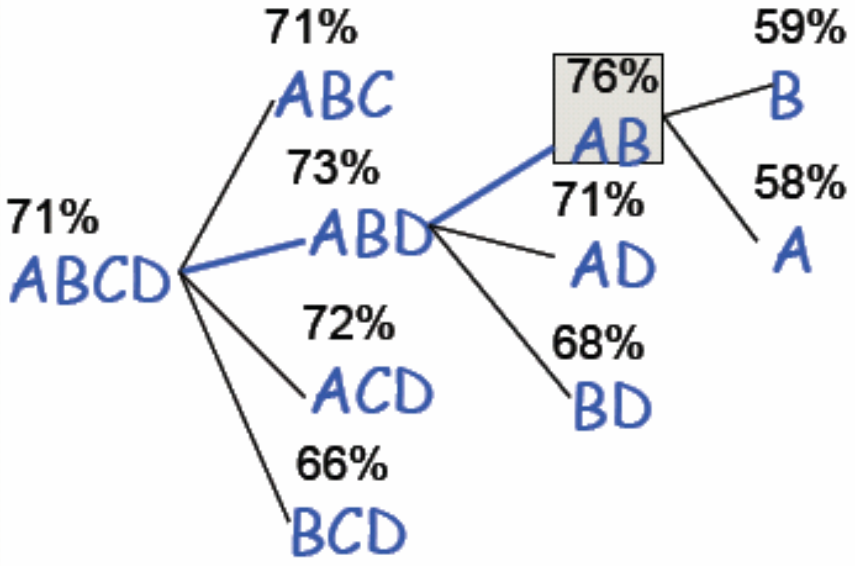
Induction Algorithm is wrapped in the selection mechanism

Strengths

- Takes bias of alg. into account
- Considers features in context

Wrapper – ocena trafności kieruje przeszukiwaniem

- Przykład



Backward Elimination:
Slightly better than FSS because it considers features in context

Przykład użycia – dane medyczne

Data set	Number of all features	Number of features selected by FBFS algorithm	Number of features selected by best-first algorithm	Accuracy for set of all features [%]	Accuracy for FBFS algorithm [%]	Accuracy for best-first choice algorithm [%]
HIST	67	17	11	87.510.47	89.100.75	85.380.53
COOC	69	25	22	61.650.72	66.310.70	64.270.78
DENS	96	16	9	19.920.50	27.040.65	22.810.76

Tabela – ocena zdolności rozpoznawania klas obrazów (algorytm typu K-NN i wrapper)

Za: J.Jelonek, J.Stefanowski: Feature subset selection for classification of histological images, *Artificial Intelligence in Medicine*, vol. 9, 1997, 227-239.

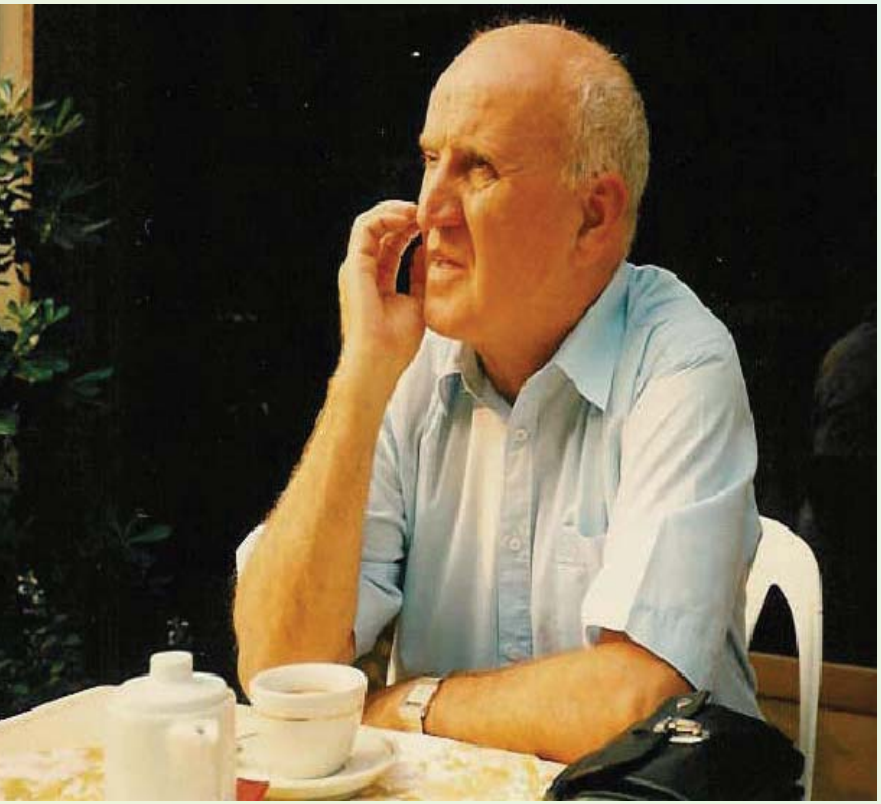
Wrapper vs. filter

- „Wrapper” uwzględniający konkretny algorytm uczenia klasyfikatora może być skuteczniejszy niż „filter”, lecz jest kosztowny obliczeniowo (np. > kilkadziesiąt atrybutów).
- Podejście pragmatyczne
  - Najpierw użyj „filter” dla wyboru większego zbioru atrybutów
  - Później wykorzystaj „wrapper”

Teoria zbiorów przybliżonych – ang. rough sets theory

- **Rough set theory** teoria zbiorów przybliżonych wprowadzona przez Zdzisława Pawlaka 1982.
- Podstawowe zastosowania – problemy klasyfikacyjne w tablicach decyzyjnych.
- Klasyczne prace:
  - Z. Pawlak, "Rough Sets", *International Journal of Computer and Information Sciences*, Vol.11, 341-356 (1982)
  - Z. Pawlak, *Rough Sets - Theoretical Aspect of Reasoning about Data*, Kluwer Academic Pubilishers (1991)

Zdzisław I. Pawlak (1926-2006)



- Najbardziej znany na świecie polsk naukowiec z zakresu informatyki
- Brał udział w konstrukcji polskich komputerów (latach 50-60)
- Pierwsze prace polskie prace naukowe publikowane w USA (1953)
- Logiczne podstawy informatyki
- Teoria automatów
- Maszyny bezadresowe
- Języki wyszukiwania informacji
- Teoria zbiorów przybliżonych

Przykład redukcji tablicy decyzyjnej oraz teoria zbiorów przybliżonych [Z.Pawlak 1992]

U	*Headache*	*Muscle pain*	*Temp.*	*Flu*
U1	Yes	Yes	Normal	No
U2	Yes	Yes	High	Yes
U3	Yes	Yes	Very-high	Yes
U4	No	Yes	Normal	No
U5	No	No	High	No
U6	No	Yes	Very-high	Yes

CORE **= {Headache,Temp}**
∩ {MusclePain, Temp}
= {Temp}

Reduct1 **= {Muscle-pain,Temp**

U	*Muscle pain*	*Temp.*	*Flu*
U1,U4	Yes	Normal	No
U2	Yes	High	Yes
U3,U6	Yes	Very-high	Yes
U5	No	High	No

Reduct2 **= {Headache, Temp.}**

U	*Headache*	*Temp.*	*Flu*
U1	Yes	Norlmal	No
U2	Yes	High	Yes
U3	Yes	Very-high	Yes
U4	No	Normal	No
U5	No	High	No
U6	No	Very-high	Yes

Discernibility Matrix (relative to positive region) C.Rauszer, A.Skowron 1992

- Let *T* = *(U, C, D)* be a decision table, with

$$U = \{u_1, u_2, ..., u_n\}.$$

By a discernibility matrix of *T*, denoted *M(T)*, we will mean $n \times n$ matrix defined as:

$$m_{ij} = \begin{cases} \{c \in C: c(u_i) \neq c(u_j)\} & \text{if } \exists d \in D [d(u_i) \neq d(u_j)] \\ \lambda & \text{if } \forall d \in D [d(u_i) = d(u_j)] \end{cases}$$

for i, j = 1,2,…,n such that u_i or u_j belongs to the *C*-positive region of *D.*

- m_{ij} is the set of all the condition attributes that classify objects u_i and u_j into different classes.

Discernibility Matrix (relative to positive region) (2)

- The equation of a boolean function is created as a conjunction taken over all non-empty entries (boolean variables corresponding to attributes) of *M(T)* corresponding to the indices *i, j* such that

  u_i or u_j belongs to the *C*-positive region of *D*.
- $m_{ij} = \lambda$ denotes that this case does not need to be considered. Hence it is interpreted as logic truth.
- All disjuncts of minimal disjunctive form of this function define the reducts of *T* (relative to the positive region).

Examples of Discernibility Matrix

No	a	b	c	d
u1	*a0*	*b1*	*c1*	*y*
u2	*a1*	*b1*	*c0*	*n*
u3	*a0*	*b2*	*c1*	*n*
u4	*a1*	*b1*	*c1*	*y*

$C = \{a, b, c\}$

$D = \{d\}$

$(a \vee c) \wedge b \wedge c \wedge (a \vee b)$

$= b \wedge c$

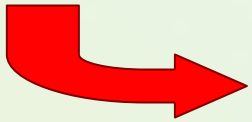
Reduct = $\{b, c\}$

In order to discern equivalence classes of the decision attribute *d*, to preserve conditions described by the discernibility matrix for this table

	u1	u2	u3
u2	a,c		
u3	b	λ	
u4	λ	c	a,b

Examples of Discernibility Matrix (2)

	a	*b*	*c*	*d*	*E*
u1	1	0	2	1	1
u2	1	0	2	0	1
u3	1	2	0	0	2
u4	1	2	2	1	0
u5	2	1	0	0	2
u6	2	1	1	0	2
u7	2	1	2	1	1



	u1	u2	u3	u4	u5	u6
u2	λ					
u3	b,c,d	b,c				
u4	b	b,d	c,d			
u5	a,b,c,d	a,b,c	λ	a,b,c,d		
u6	a,b,c,d	a,b,c	λ	a,b,c,d	λ	
u7	λ	λ	a,b,c,d	a,b	c,d	c,d

Core = {*b*}

Reduct1 = {*b*,*c*}

Reduct2 = {*b*,*d*}

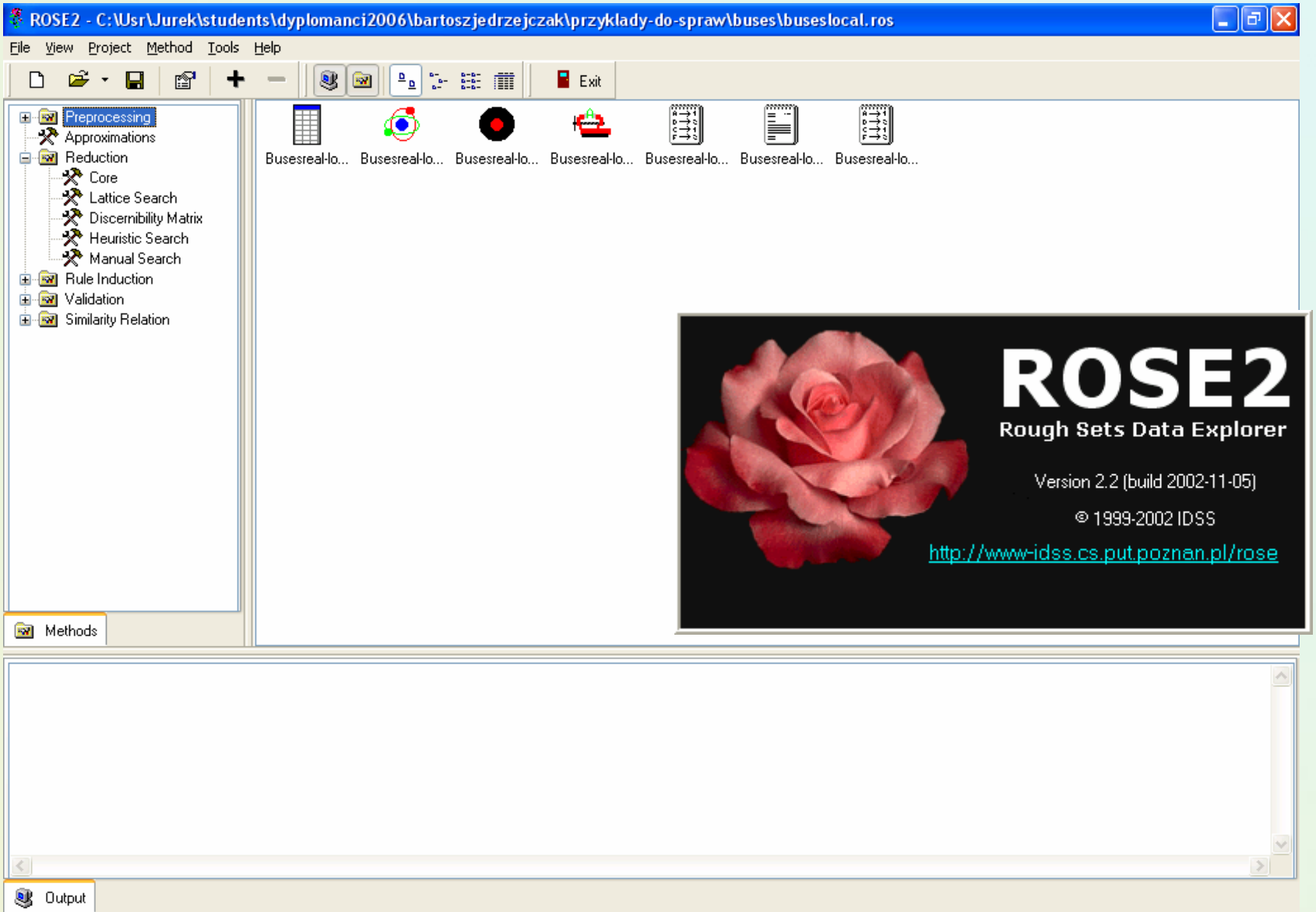
Quality of approximation of object classification

- Let $\chi=\{X_1,X_2,..,X_m\}$ be classification of objects from U, i.e. $\forall i,j \leq n : X_i \cap X_j = \varnothing$ and $\bigcup X_i = U,\ i=1,..,n.$
- The quality of approximation of object classification with respect to the subset of attributes $B \subseteq A$ is defined as:

$$\gamma_B(\chi)=\frac{\sum_{i=1}^{n} |\underline{B}X_i|}{|U|}$$

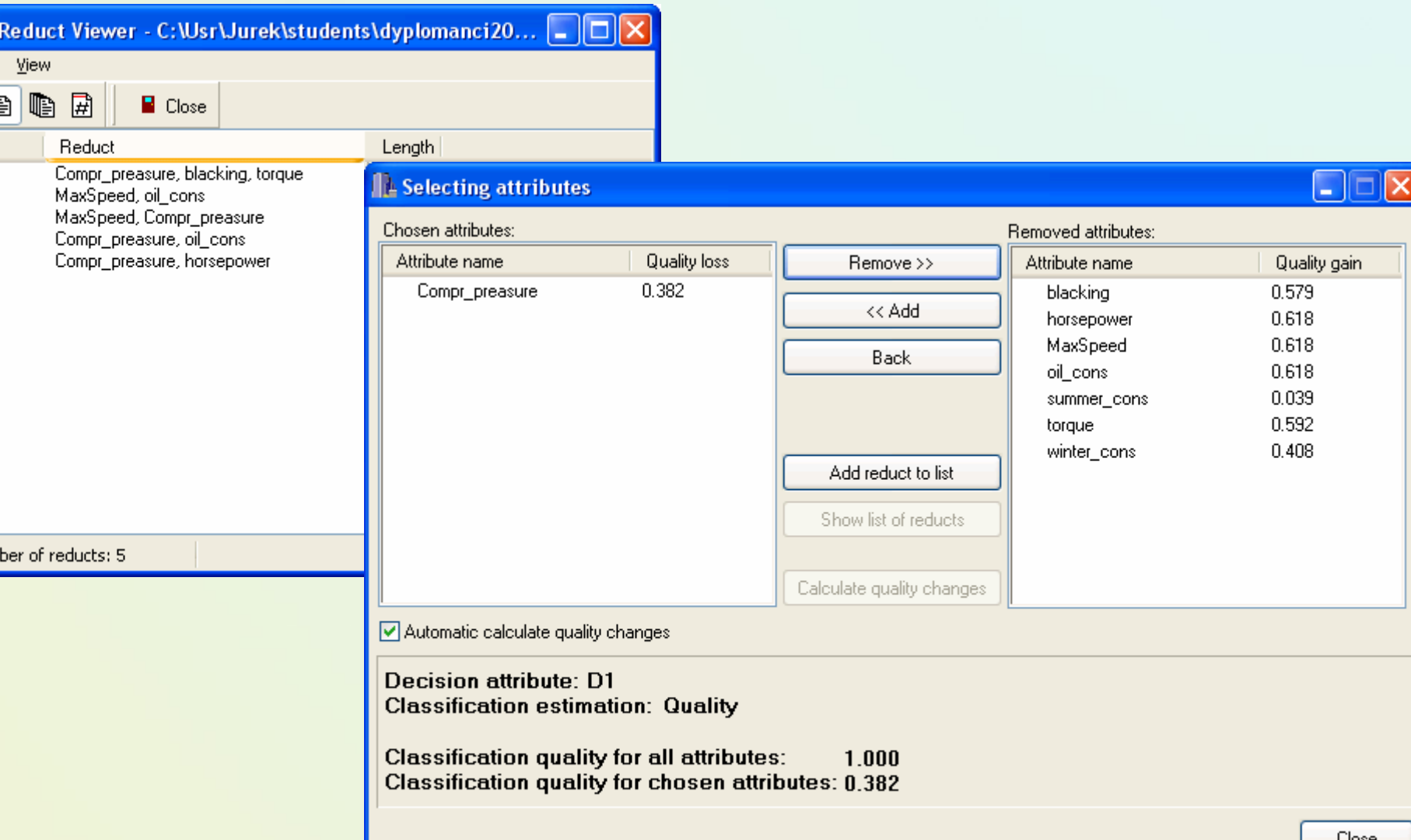
- Significance of attribute a $\rightarrow$ $\gamma_B(\chi) - \gamma_{B-a}(\chi)$

Oprogramowanie do TZP → ROSE (IDSS)



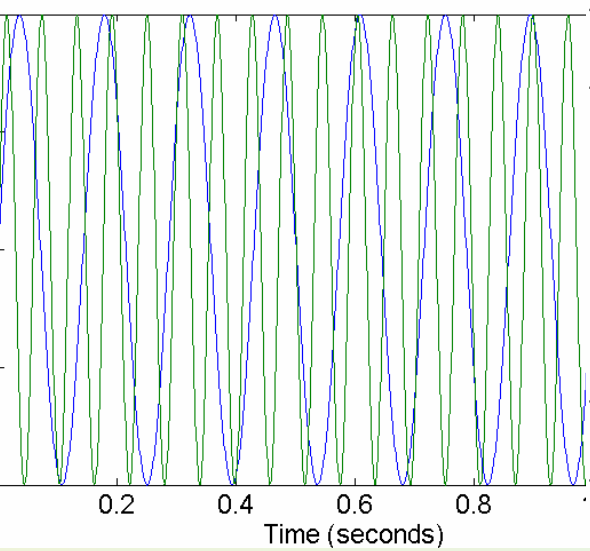
Rose → redukcja

- Redukt vs. significance for classification

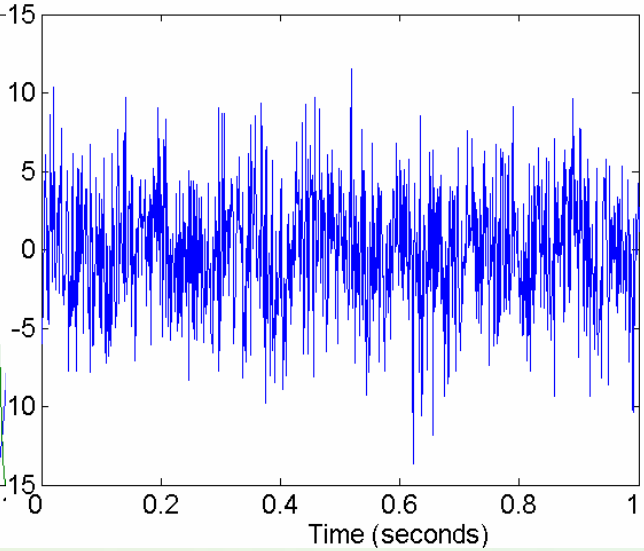


Mapping Data to a New Space

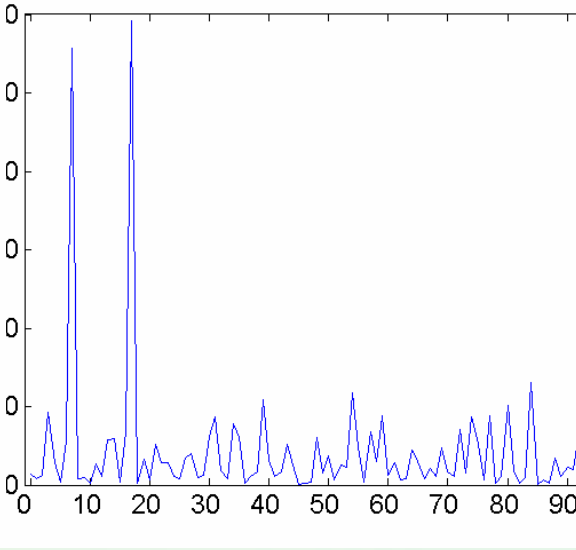
- **Fourier transform**
- **Wavelet transform**



Two Sine Waves



Two Sine Waves + Noise



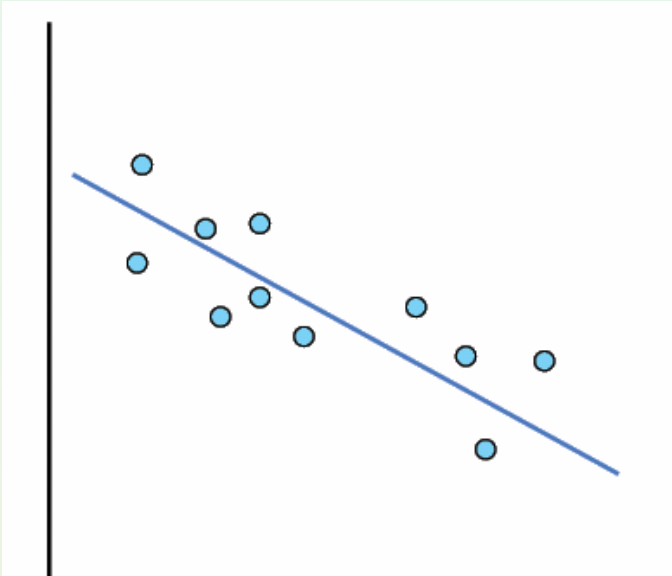
Frequency

Metody projekcji → tworzenie nowych atrybutów

- Principal Component Analysis PCA → Analiza składowych głównych / Pearson 1901
- Dla *n* obserwacji z *m* wymiarowej przestrzeni znajdź *c* ortogonalnych wektorów (*c* << *m*), które dobrze reprezentują właściwości oryginalnych danych.
- Jak dokonać dobrej konstrukcji / redukcji wymiarów bez znaczącej utraty informacji.
- Używane do kompresji danych i lepszej wizualizacji ogólnych prawidłowości w związkach między danymi
- Przeznaczone do danych liczbowych
- Oprogramowanie statystyczne → Dataminer Statsoft
- Trochę opisów i przykładów → książki Larose / Stanisz / Krzyśko,Wołyński

PCA – analiza składowych głównych (2)

- Istota to „ortogonalne” przekształcenie początkowych zmiennych w nowy zbiór nieskorelowanych zmiennych
  - Całkowita wariancja zmiennych jest równa sumie wariancji składowych głównych
- Każdy z nowych wektorów jest **kombinacją liniową** pewnych składowych głównych (odnoszących się do oryginalnych zmiennych)
- $\mathbf{x}' = \mathbf{W} \cdot \mathbf{x}$
- A co z nieliniowymi projekcjami?



PCA – trochę matematyki

- PCA simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance

The optimal* approximation of a random vector $\underline{x} \in \Re^N$ by a linear combination of M (M<N) independent vectors is obtained by projecting the random vector $\underline{x}$ onto the eigenvectors $\underline{v}_i$ corresponding to the largest eigenvalues λ_i of the covariance matrix of x (Σ_x)

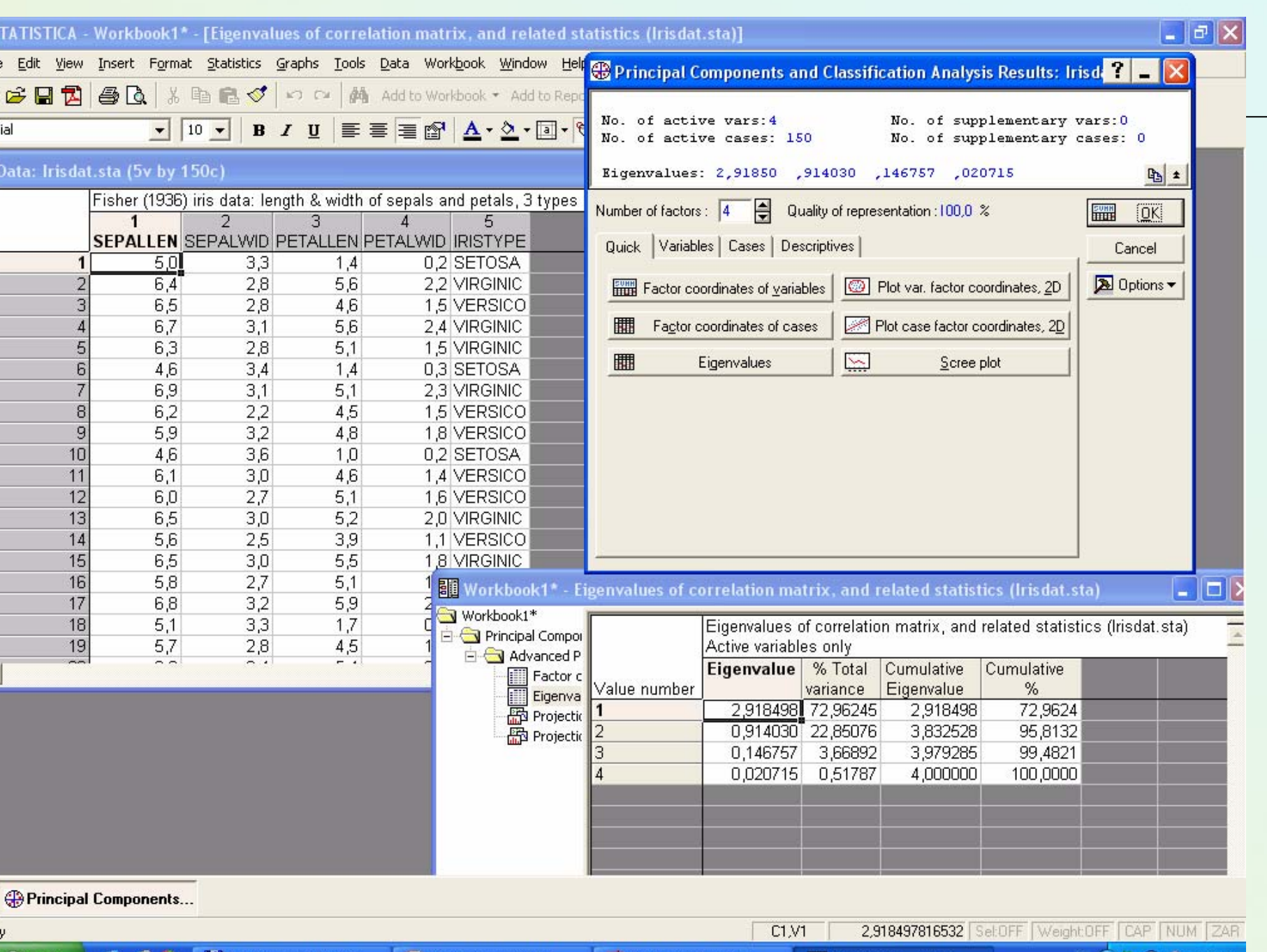
Principal Components Analysis

- **Summary**

$$\underline{x}' = y_1\underline{v}_1 + y_2\underline{v}_2 \cdots + y_M\underline{v}_M$$

$$\underline{x}' = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} \underline{v}_1^T \\ \underline{v}_2^T \\ \\ \underline{v}_M^T \end{bmatrix} \underline{x} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \\ v_{21} & v_{22} & & \\ \vdots & & \ddots & \\ v_{M1} & v_{M2} & \cdots & v_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}$$

- where $\underline{v}_k$ is the eigenvector corresponding to the k^{th} largest eigenvalue of the covariance matrix


TATISTICA - Workbook1* - [Eigenvalues of correlation matrix, and related statistics (Irisdat.sta)]
Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Add to Workbook
Add to Repo
Data: Irisdat.sta (5v by 150c)
Fisher (1936) iris data: length & width of sepals and petals, 3 types
1 SEPALLEN
2 SEPALWID
3 PETALLEN
4 PETALWID
5 IRISTYPE
1 5,0 3,3 1,4 0,2 SETOSA
2 6,4 2,8 5,6 2,2 VIRGINIC
3 6,5 2,8 4,6 1,5 VERSICO
4 6,7 3,1 5,6 2,4 VIRGINIC
5 6,3 2,8 5,1 1,5 VIRGINIC
6 4,6 3,4 1,4 0,3 SETOSA
7 6,9 3,1 5,1 2,3 VIRGINIC
8 6,2 2,2 4,5 1,5 VERSICO
9 5,9 3,2 4,8 1,8 VERSICO
10 4,6 3,6 1,0 0,2 SETOSA
11 6,1 3,0 4,6 1,4 VERSICO
12 6,0 2,7 5,1 1,6 VERSICO
13 6,5 3,0 5,2 2,0 VIRGINIC
14 5,6 2,5 3,9 1,1 VERSICO
15 6,5 3,0 5,5 1,8 VIRGINIC
16 5,8 2,7 5,1
17 6,8 3,2 5,9
18 5,1 3,3 1,7
19 5,7 2,8 4,5
Principal Components and Classification Analysis Results: Irisd
No. of active vars: 4
No. of active cases: 150
No. of supplementary vars: 0
No. of supplementary cases: 0
Eigenvalues: 2,91850 ,914030 ,146757 ,020715
Number of factors: 4
Quality of representation: 100,0 %
OK
Cancel
Options
Quick
Variables
Cases
Descriptives
Factor coordinates of variables
Plot var. factor coordinates, 2D
Factor coordinates of cases
Plot case factor coordinates, 2D
Eigenvalues
Scree plot
Workbook1* - Eigenvalues of correlation matrix, and related statistics (Irisdat.sta)
Workbook1*
Principal Compor
Advanced P
Factor c
Eigenva
Projectic
Projectic
Eigenvalues of correlation matrix, and related statistics (Irisdat.sta)
Active variables only
Value number
Eigenvalue
% Total variance
Cumulative Eigenvalue
Cumulative %
1 2,918498 72,96245 2,918498 72,9624
2 0,914030 22,85076 3,832528 95,8132
3 0,146757 3,66892 3,979285 99,4821
4 0,020715 0,51787 4,000000 100,0000
Principal Components...
C1,V1
2,918497816532
Sel:OFF
Weight:OFF
CAP
NUM
ZAR

Inne ciekawe transformacje …

Metody dekompozycji macierzy

- Singular Value Decomposition
- Powiązane z zastosowanie do przetwarzania dokumentów tekstowych / Latent Semantic Indexing

Sieci neuronowe Kohonena - SOM

Analiza Korespondencji

…

Indukcja konstruktywna [R.Michalski]

Definicja:

- *Przekształcanie przestrzeni hipotez uczenia w ten sposób, aby pojęcie docelowe mogło być w niej reprezentowane dokładnie i oszczędnie oraz aby możliwe było efektywne nauczenie się go za pomocą stosowanego algorytmu uczenia się.*
- Przestrzeń hipotez – zbiór możliwych hipotez dotyczących pojęcia docelowego, które można skonstruować w oparciu o wybrany algorytm uczenia oraz sposób reprezentacji danych.
- Rezygnacja z ustalonej i zadanej z góry przestrzeni hipotez może pozwolić na lepsze dostosowanie się do charakteru pojęcia docelowego.
- Najczęściej rozważa się przekształcenie języka reprezentacji w odniesieniu do atrybutów, np.:
  - usuwanie zbędnych atrybutów,
  - tworzenie nowych atrybutów zależnych funkcjonalnie od istniejących dotychczas atrybutów

Przykład indukcji reguł decyzyjnych

Nr.	wysokość	długość	szerokość	Klasa dec.
1	2	12	2	1
2	6	4	2	1
3	3	8	2	1
4	4	4	3	1
5	12	4	2	2
6	4	12	2	2
7	8	6	2	2
8	6	8	3	2

Zbiór reguł decyzyjnych otrzymanych za pomocą algorytmu LEM2
(długość =4) & (wysokość =6) ==> (decyzja = 1) {2}
(wysokość =4) & (długość =4) ==> (decyzja = 1) {4}
(wysokość =3) ==> (decyzja = 1) {3}
(wysokość =2) ==> (decyzja = 1) {1}
(długość =8) & (wysokość =6) ==> (decyzja = 2) {8}
(długość =12) & (wysokość =4) ==> (decyzja = 2) {6}
(wysokość =12) ==> (decyzja = 2) {5}
(wysokość =8) ==> (decyzja = 2) {7}

Przekształcenie przestrzeni atrybutów

Wprowadza się nowy atrybut – **iloczyn wysokości i długości**

Nr.	wysokość	długość	szerokość	**wysokość · długość**	Klasa dec.
1	2	12	2	**24**	1
2	6	4	2	**24**	1
3	3	8	2	**24**	1
4	4	4	3	**16**	1
5	12	4	2	**48**	2
6	4	12	2	**48**	2
7	8	6	2	**48**	2
8	6	8	3	**48**	2

(wysokość_·_długość =16) ==> (decyzja = 1), {4}

(wysokość_·_długość =24) ==> (decyzja = 1), {1,2,3}

(wysokość_·_długość =48) ==> (decyzja = 2), {5,6,7,8}