Przetwarzanie wstępne danych

Projekt Eksploracji Danych

JERZY STEFANOWSKI
Instytut Informatyki
Politechnika Poznańska

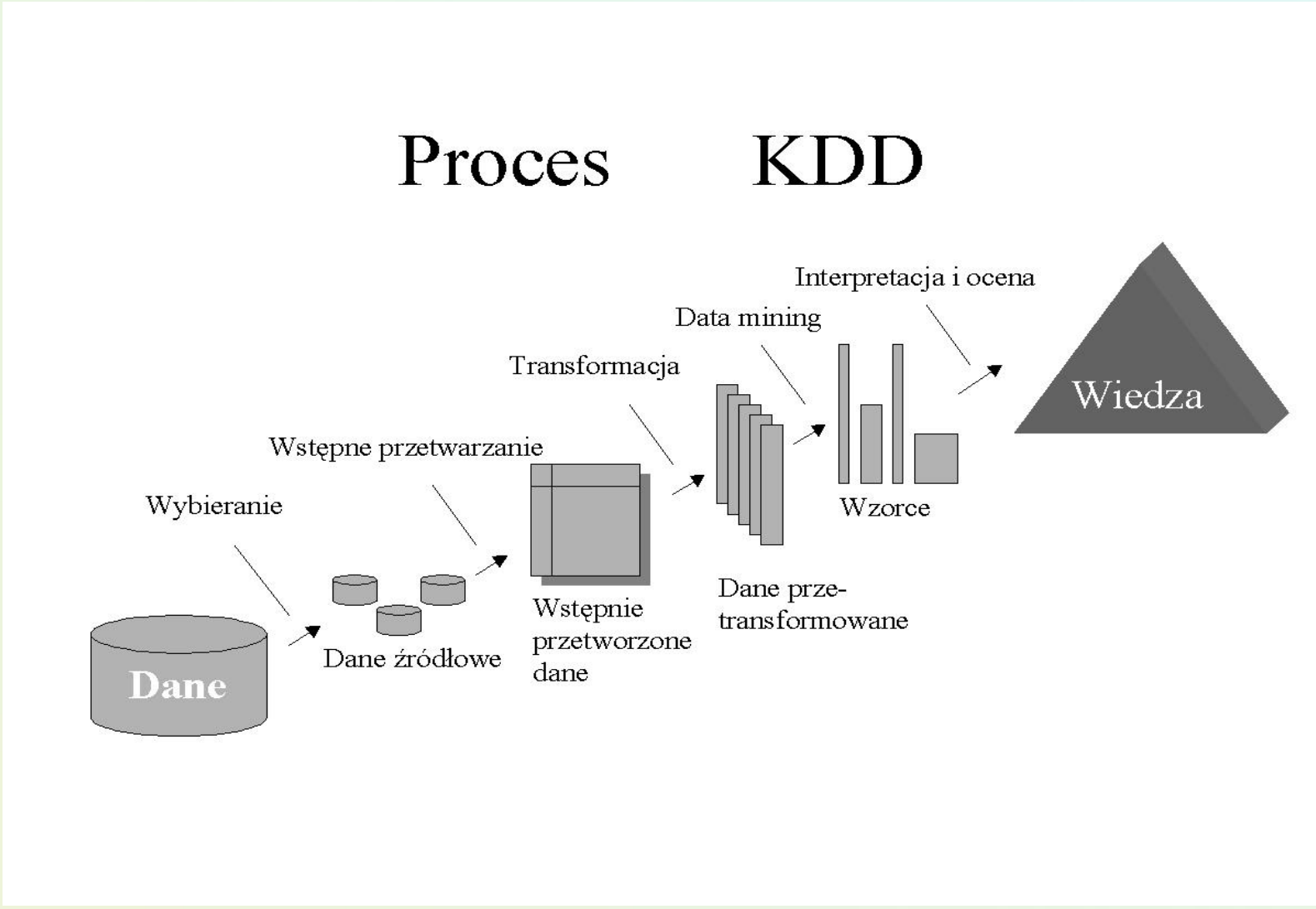
Wykład 2
2020

Plan wykładu

1. Miejsce przetwarzania wstępnego danych w procesie KDD
2. Związki z integracją danych
3. Oczyszczanie danych
   - Nieznane wartości atrybutów
   - Identyfikacja obserwacji odstających
4. Redukcja rozmiarów danych
   - Selekcja atrybutów
   - Wybrane metody

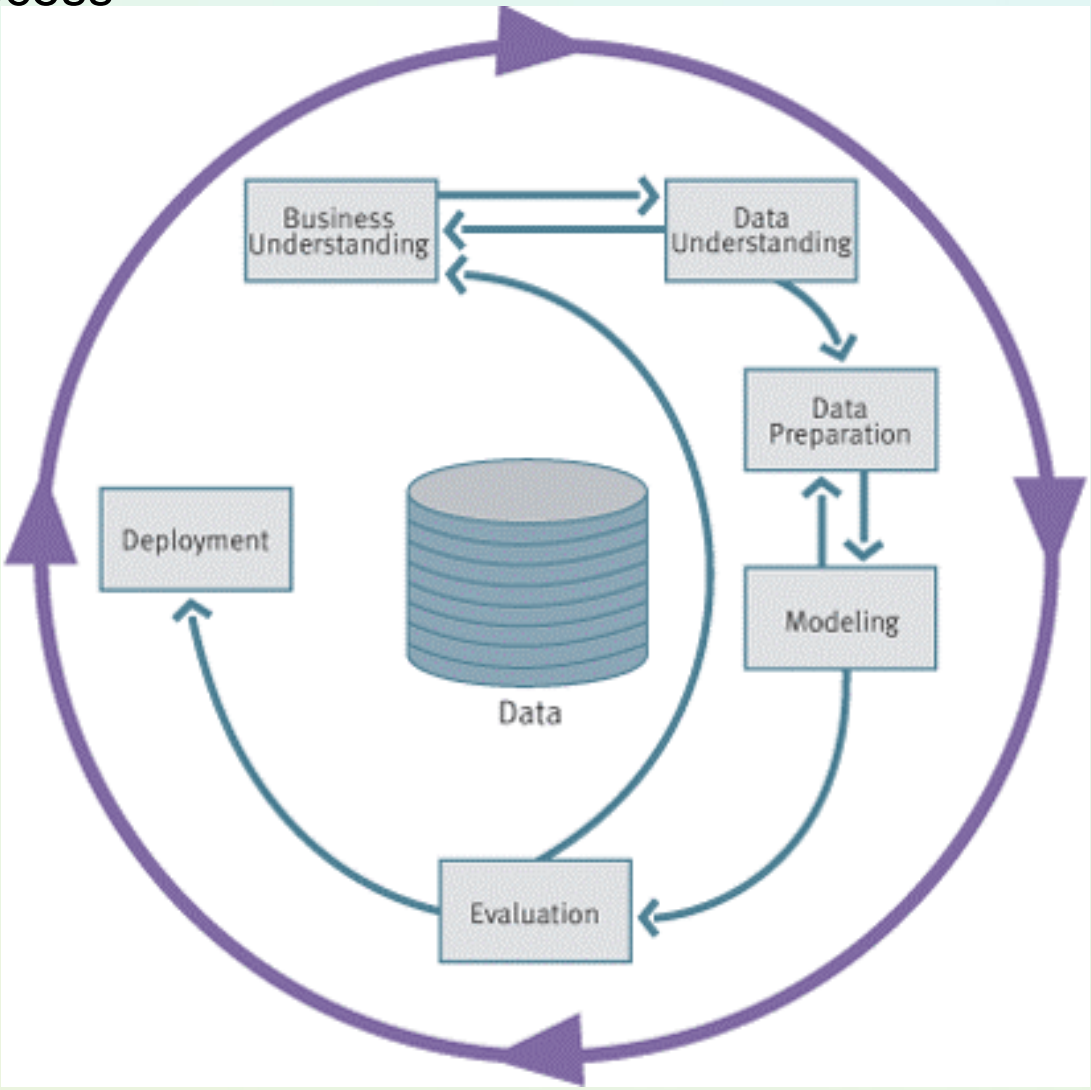
Slajdy – niektóre częściowo oparte na materiałach od Han i Tan, Steinbach, Kumar

Proces odkrywanie wiedzy i etapy początkowe



CRISP-DM: Próba standaryzacji postępowania

• Cross-Industry Standard Process for Data Mining (CRISP-DM)



Etapy i czynności wg. CRISP-DM

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

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

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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

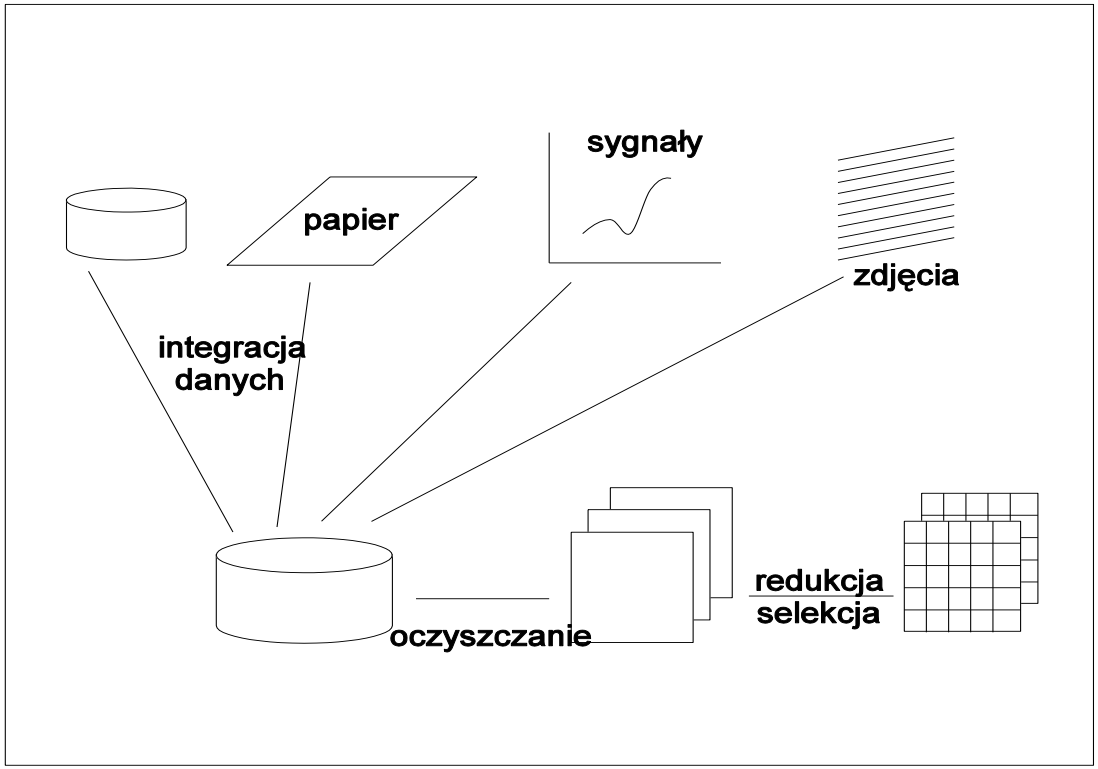
- Dane w innych formatach tranformowane do klasycznej reprezentacji tabeli (artrybut-wartość)

- Tabele danych - macierz $m \times n$, gdzie m wierszy odpowiadających instancjom, oraz n kolumn odpowiadających atrybutom

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

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.



Spojrzenie bazo-danowego data mining

1. Integracja danych (patrz Hurtownie danych)
2. Oczyszczanie
   - Błędy
   - Nieznane wartości atrybutów
   - „Szum” informacyjny
   - Obserwacje odstające
3. Transformacje dziedzin atrybutów
   - Dyskretyzacja atrybutów liczbowych
   - Inne przekształcenia
4. Redukcja rozmiarów danych
   - Selekcja atrybutów
   - Wybór obiektów

Za podręcznikiem J.Han

Trudności integracji danych

- *Integracja danych* – łączenie danych z różnych źródeł w jednolitą, spójną strukturę danych
- Podstawowe zagadnienia (hurtownie danych):
  - Integracja schematów,
  - Integracja danych wirtualnych,
  - Integracja danych zmaterializowanych

Integracja danych to nie tylko porównywanie i integracja schematów, ale również łączenie rzeczywistej zawartości źródłowych baz danych

Konflikty nazewnictwa, dziedzin, wartości, meta-danych,

Więcej o projektach metodyk i narzędziach w: Jarke i in., Hurtownie danych. Podstawy organizacji i funkcjonowania

Rozstrzyganie konfliktów - metodyki

- Tworzenie baz wiedzy zawierających hierarchie pojęć reprezentujących związki między atrybutami w różnych schematach.
- Zbiór transformacji w języku pierwszego rzędu wzbogaconym o reguły wyrażające więzy.
- Automatyczne moduły wnioskujące korzystające z baz wiedzy, ontologii terminologicznych i reguł.
- Narzędzia do porównywania schematów i rozstrzygania konfliktów:
- **Konflikty nazewnictwa** (różne schematy używają różnej terminologii odnośnie tych samych danych)
  - Customer_id w systemie 1 vs. Cust_np. w systemie 2,
    - homonimy,
    - synonimy.
- **Konflikty semantyczne**
  - Różni użytkownicy mają inny poziom zrozumienia, czego tak naprawdę dotyczy informacja przechowywana w systemie
- **Konflikty strukturalne**

Potrzeba specjalistycznego oprogramowania

Data 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

- Przegląd różnych systemów

Sprawdź

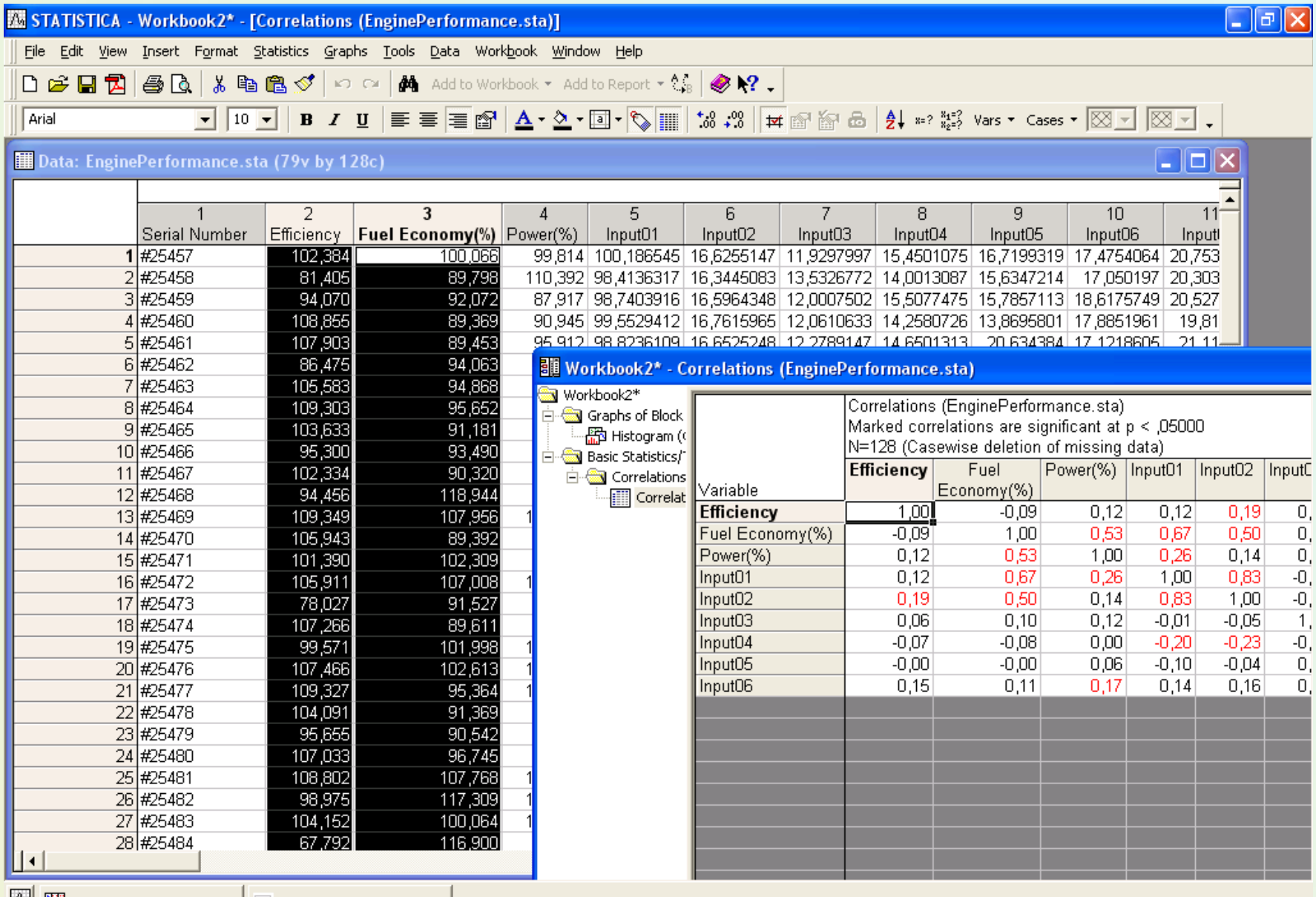
- http://www.kdnuggets.com/software/index.html

- Wykonaj b. aktualny „przegląd” źródeł!

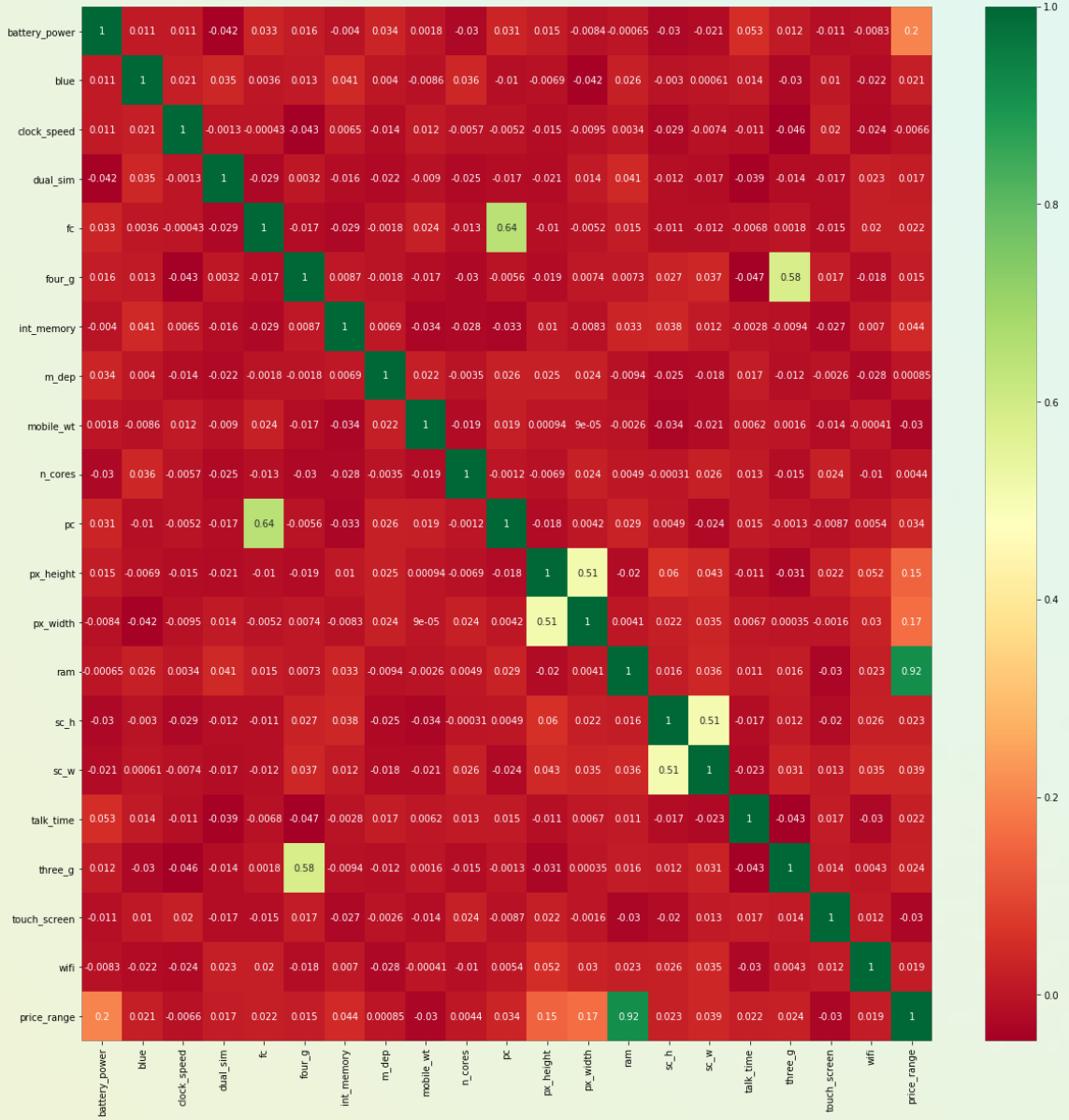
Duplicate Data – duplikaty rekordów

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

Zbyt silna zależność między kolumnami – analiza korelacji



Alternatywnie heatmaps

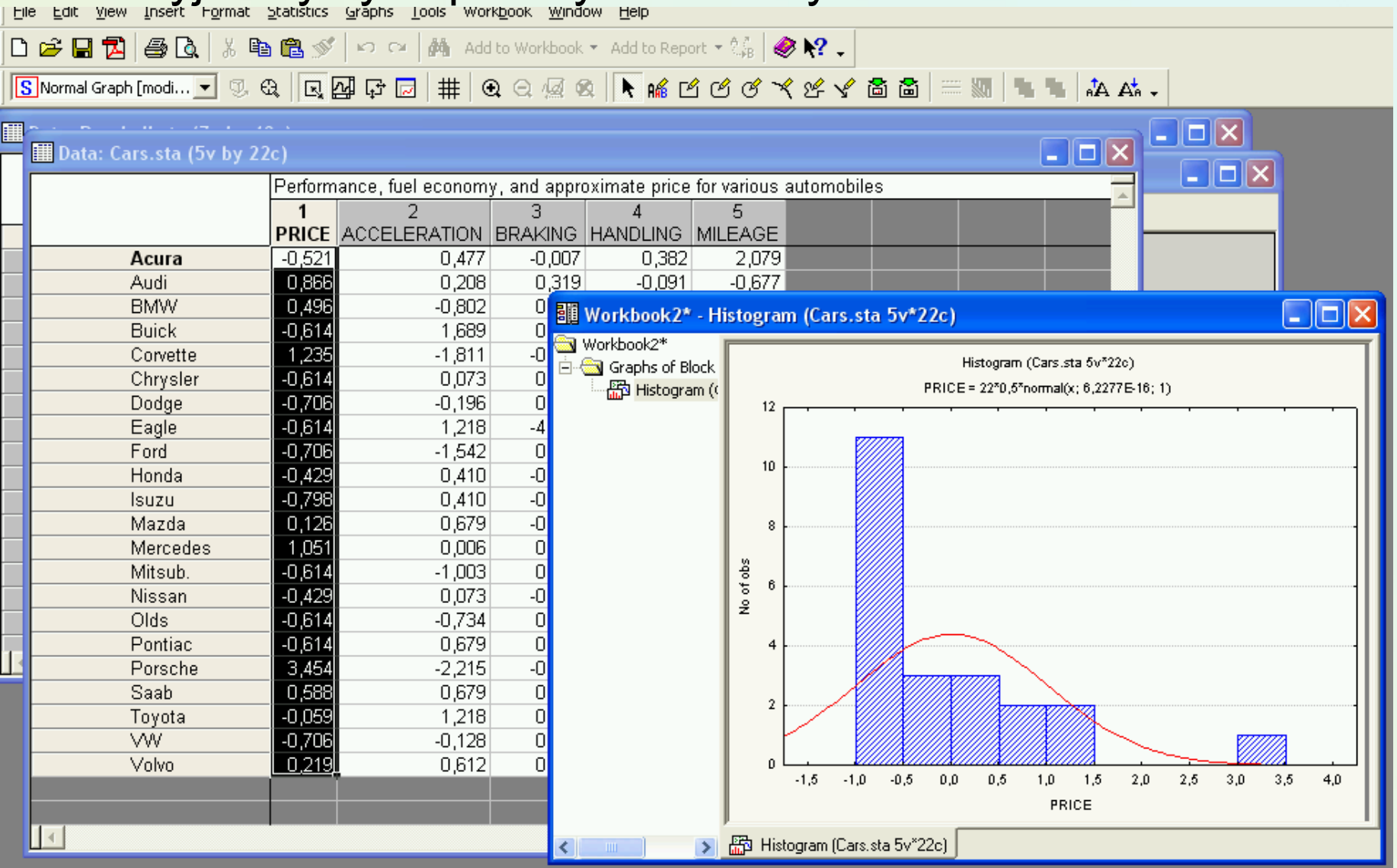



```
import pandas as pd
import numpy as np
import seaborn as sns


data = pd.read_csv("D://Blogs//train.csv")
X = data.iloc[:,0:20]  #independent columns
y = data.iloc[:,-1]    #target column i.e price range
#get correlations of each features in dataset
corrmat = data.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))
#plot heat map
g=sns.heatmap(data[top_corr_features].corr(),annot=True,cmap="RdYlGn
")
```

Outliers – obserwacje oddalone / odstające

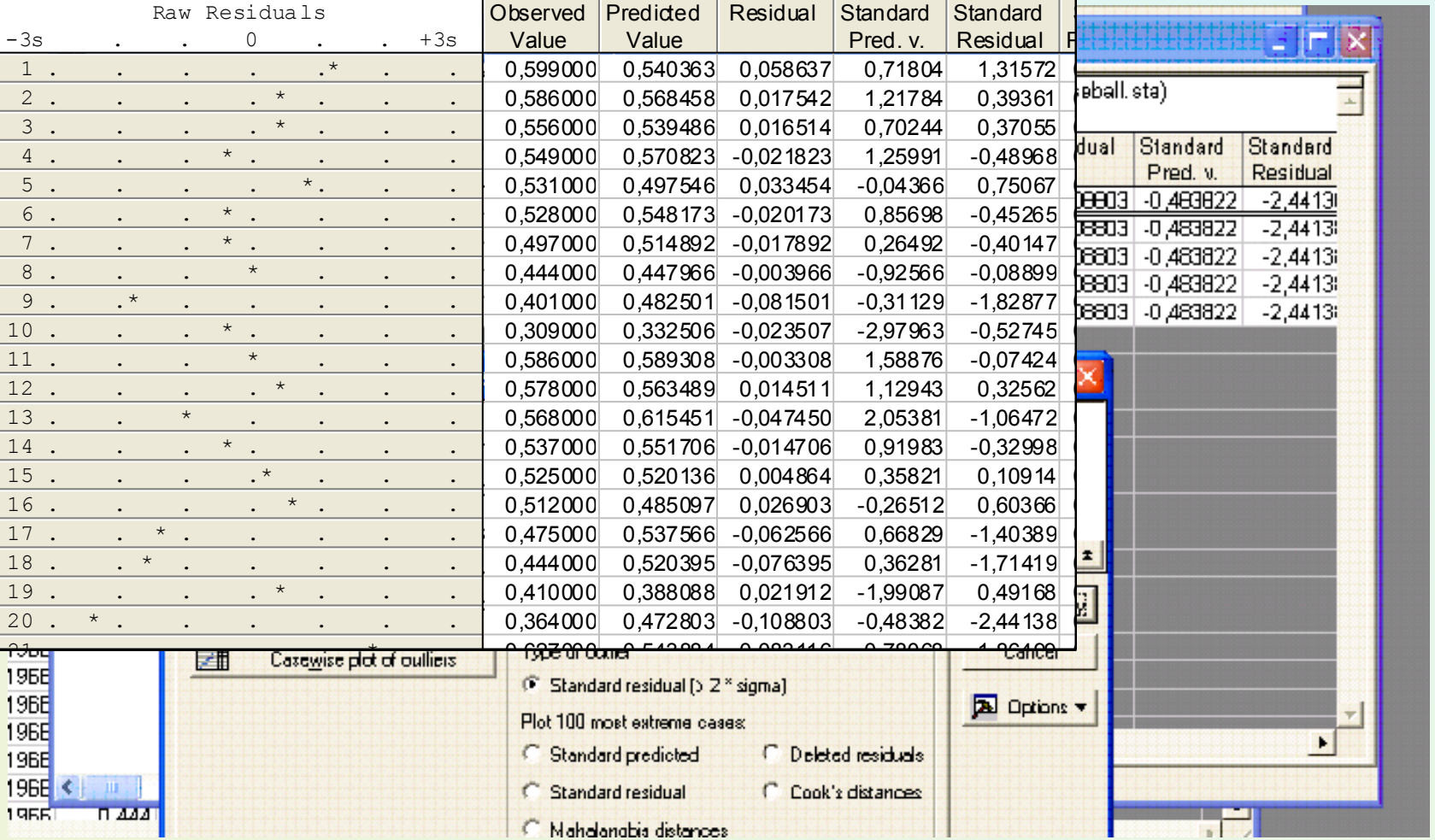
- Użyj statystyk opisowych oraz wykresów - Statistica



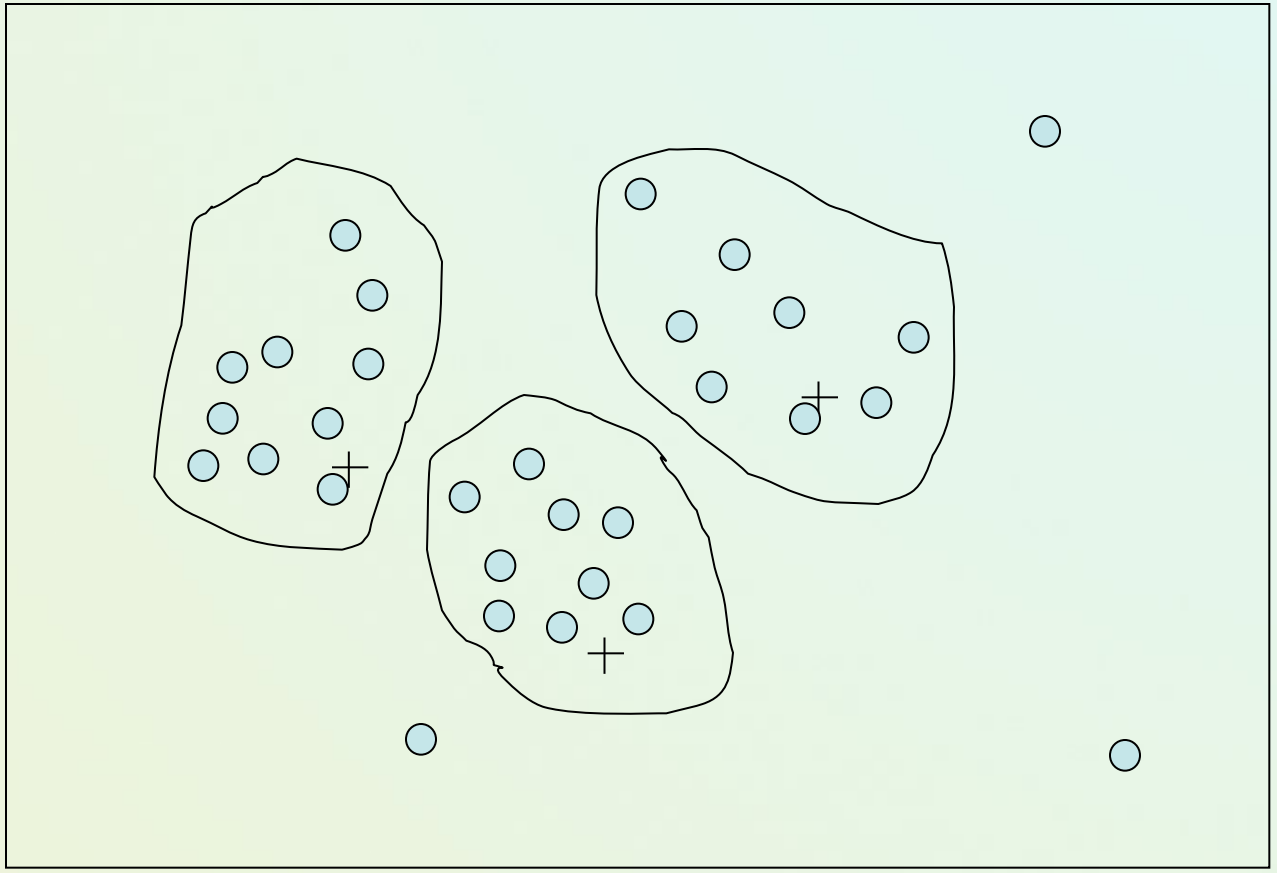
„Outiers” w wielowymiarowej regresji - analiza reszt → |standaryzowane reszty| ≤ 2

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 analizie skupień



Różne grupy rozwiązań

- Statystyczne (podstawowe statystyki + wizualizacje).
- Grupowanie (koszty obliczeniowe oraz problemy z parametryzacją algorytmów!).
- „Pattern based identification” (znajdź obserwacje, które nie potwierdzają wcześniej odnalezionych wzorców).
  - Reguły asocjacyjne lub inne lokalne wzorce

Przykłady wstępnej statystycznej analizy danych

- Prof. M.Lasek „Data mining” – banking data
- Larose D. Odkrywanie wiedzy z danych, PWN.
- 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

Różnorodna semantyka wartości niezannych

- 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

- Połącz dwa lub więcej atrybutów (obiektów) w pojedynczy atrybut (obiekt)
- Cele
  - Data reduction
    - Zmniejsz rozmiary danych
  - Zmień skale / granulacje informacji
    - Cities aggregated into regions, states, countries, ..
  - Zmniejsz wrażliwość danych - more“stable” data
    - Dane zagregowane charakteryzują miary stat. o mniejszej zmienności.

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

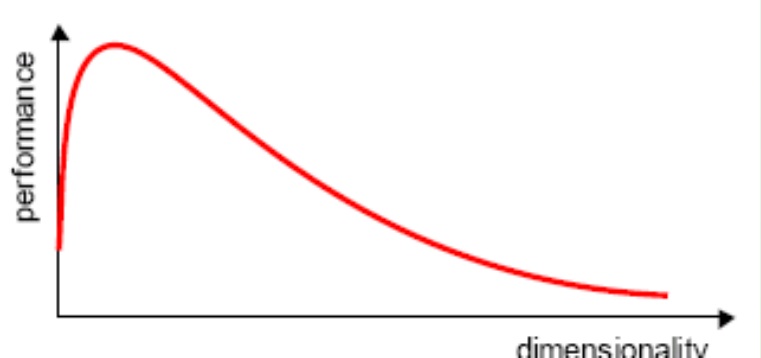
$$v' = \frac{v}{10^j}$$

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

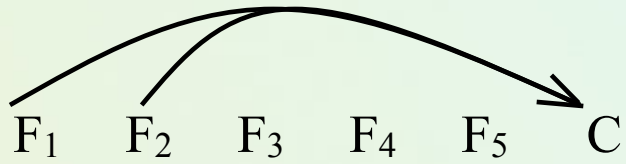
Przekleństwo wymiarowości

„Curse of dimensionality” [Bellman 1961]

- W celu dobrego przybliżenia funkcji (także 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
- For a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve!



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 (possibilities) 2^F

(tutorial on genomics [Yu 2004])

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

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

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

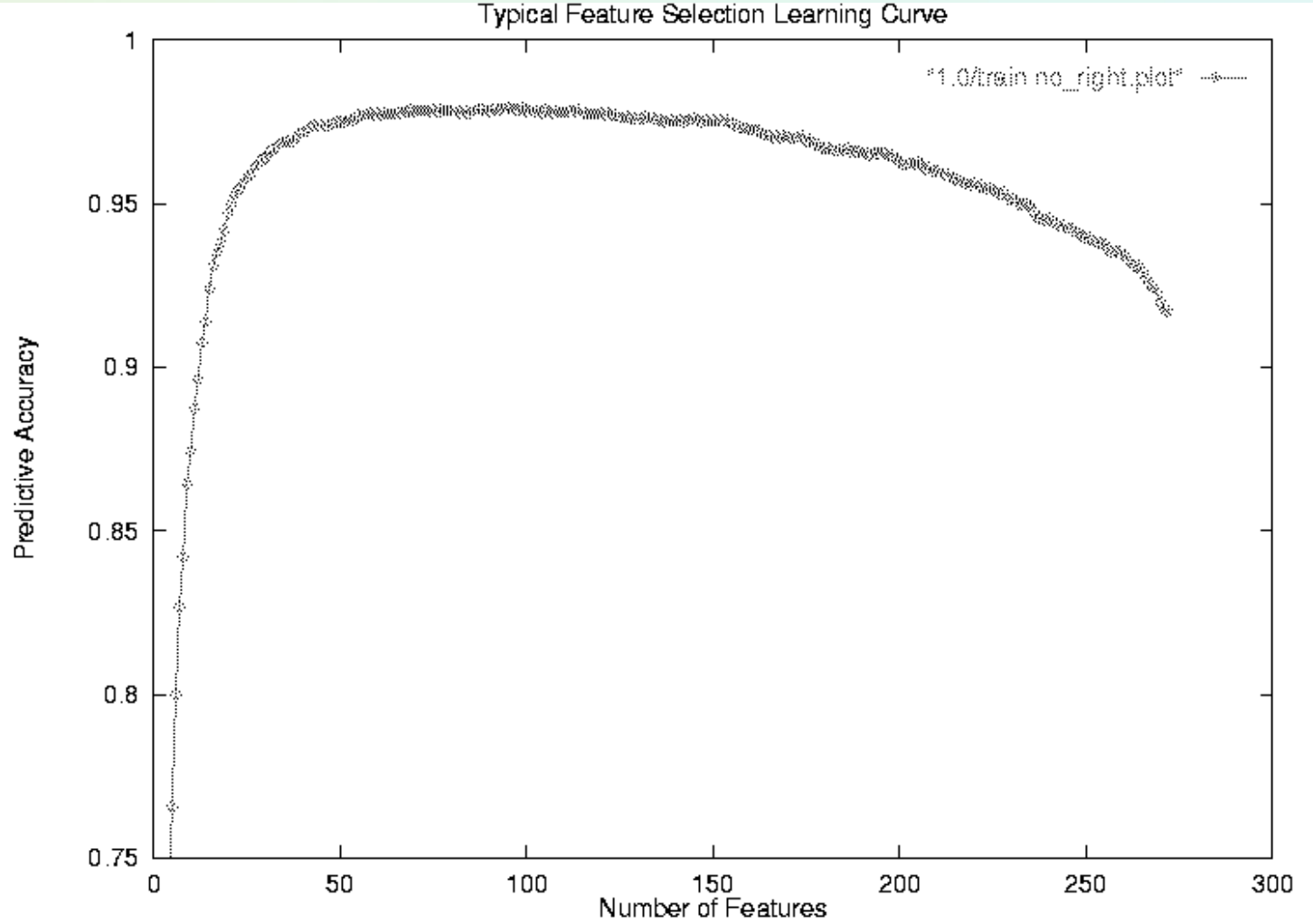
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_m} \end{bmatrix}$$

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

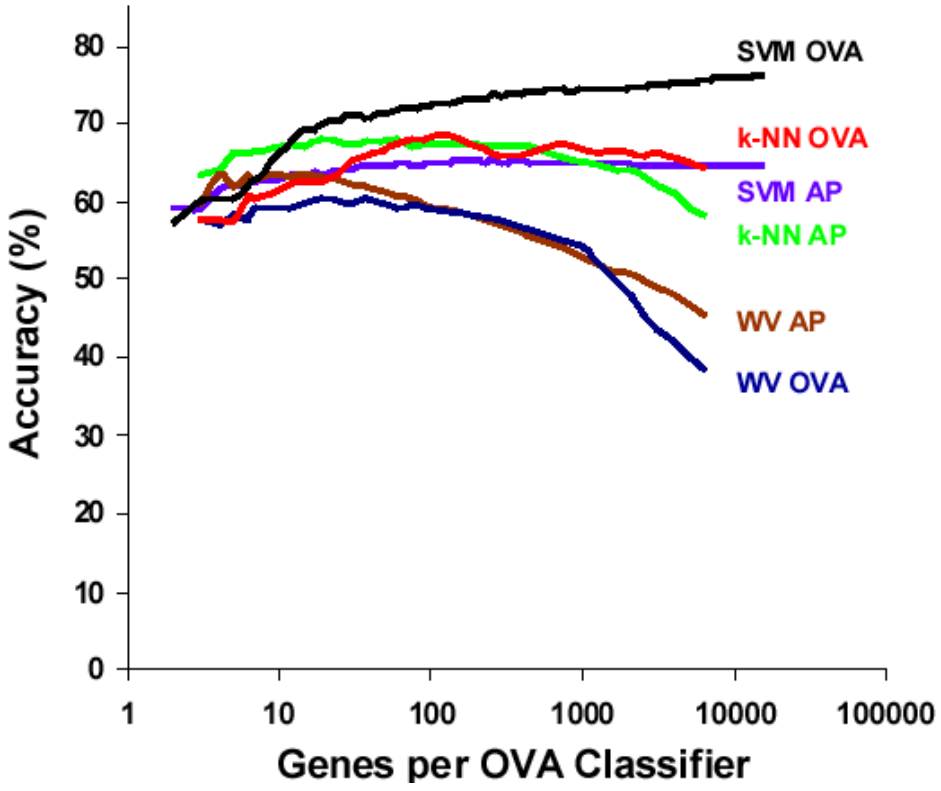
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix}\right)$$

za wykładem A.Berge nt redukcji danych

Przykład działania K-NN nad wysoce wielowymiarowymi danymi

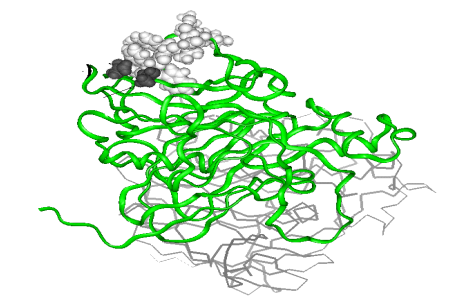


RFE SVM for cancer diagnosis



Differenciation of 14 tumors. *Ramaswamy et al, PNAS, 2001*

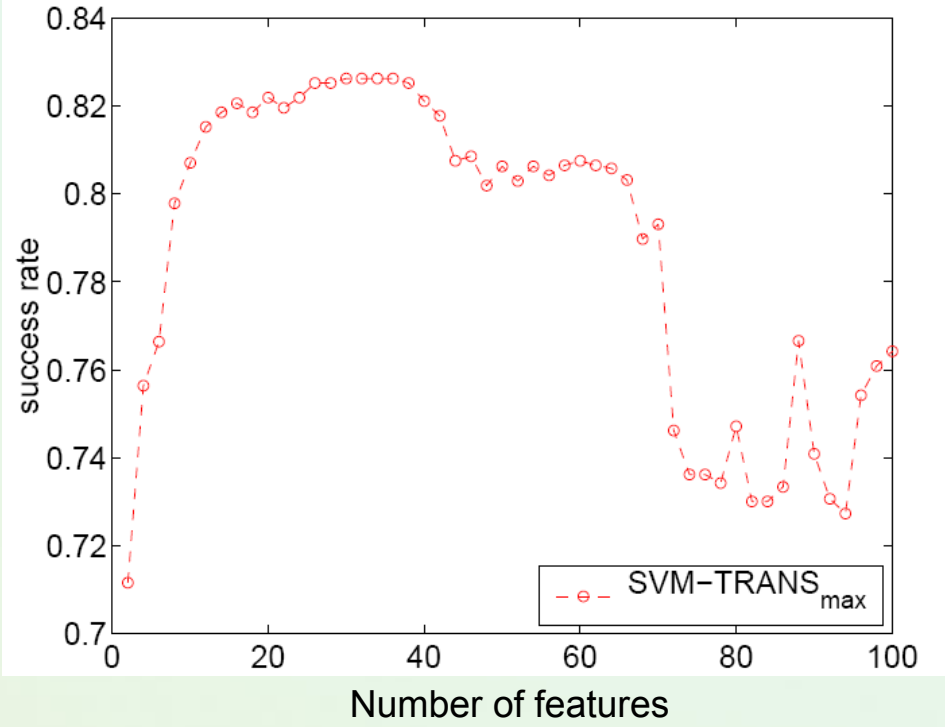
QSAR: Drug Screening



Binding to Thrombin (DuPont Pharmaceuticals)

- 2543 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting; 192 “active” (bind well); the rest “inactive”. Training set (1909 compounds) more depleted in active compounds.

- 139,351 binary features, which describe three-dimensional properties of the molecule.



Weston et al, Bioinformatics, 2002

Rodzaje atrybutów (perspektywa selekcji)

- **Relevant** - istotne – mają wpływ na wynik wyjściowy, i nie mogą być zastąpione przez inne atrybuty
- **Irrelevant** - nieistotne - nie mają wpływu na wynik wyjściowy, nie niosą użytecznej informacji (rodzaj szumu)
- **Redundant** / nadmiarowe – nie wnoszą więcej informacji niż inne / wybrane atrybuty (mogą być zastąpione przez inne atrybuty – pełnia wymienną rolę) [Dash, Hall]

Przeglądy:

Kumar V. Minz. S: Feature selection: A literature review. Smart Computing Review 2014

Li J. et al.: Feature selection. A data perspective. ACM Comoputing Surveys. 2010.

Proste podejście do selekcji atrybutów

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 ≤ 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
  - Sklearn + dodatkowe biblioteki

Metody doboru zmiennych do modelu regresji

- Zmienne wybiera się na podstawie wiedzy dziedzinowej.
- Wymagania nt. własności zmiennych niezależnych:
  - Są silnie skorelowanych ze zmienną, którą objaśniają.
  - Są nieskorelowane lub co najwyżej słabo skorelowane ze sobą.
  - Charakteryzują się dużą zmiennością: = odchylenie/średnia.
- Jak wykorzystać współczynniki korelacji?

$$r^* = \sqrt{\frac{t^2_{\alpha,n-2}}{n-2+t^2_{\alpha,n-2}}}$$

Ocena zmiennych objaśniających

- Przykład doboru zmiennych do modelu opisującego miesięczne spożycie ryb (w kg na osobę) w zależności od: spożycia mięsa x_1, warzyw x_2, owoców x_3, tłuszczów x_4 oraz wydatków na lekarstwa x_5.

nr	y	X1	X2	X3	X4	x5
1	3	3	0,63	0,63	0,12	14,1
2	3	3	1,07	1,07	0,14	12,77
3	3	3	0,44	0,44	0,1	11
4	3	2	0,26	0,26	0,04	44
5	0	0	0,01	0,0	0,0	60
6	0	0	0,02	0,01	0,0	66
7	0	0	0,02	0,01	0,01	53
8	5	4	0,09	0,09	0,03	60
9	4	2	0,56	0,56	0,19	3
10	3	2	0,11	0,11	0,05	3
11	7	7	1,46	1,46	0,34	23
12	5	5	1,22	1,22	0.24	30
13	5	5	1,22	1,22	0,26	30
14	2	1	0,31	0,13	0,05	39
15	3	2	0,4	0,19	0,05	56

Dobór zmiennych do modelu

- Współczynniki zmienności

y	x1	x2	x3	x4	X5
0,635	0,754	0,917	1,0	0,944	0,632

- Macierz współczynników korelacji

	y	x1	x2	x3	x4	X5
y	1					
x1	0,950	1				
x2	0,750	0,843	1			
x3	0,748	0,851	0,991	1		
x4	0,813	0,860	0,946	0,951	1	
x5	-0,442	-0,395	-0,477	-0,503	-0,539	1

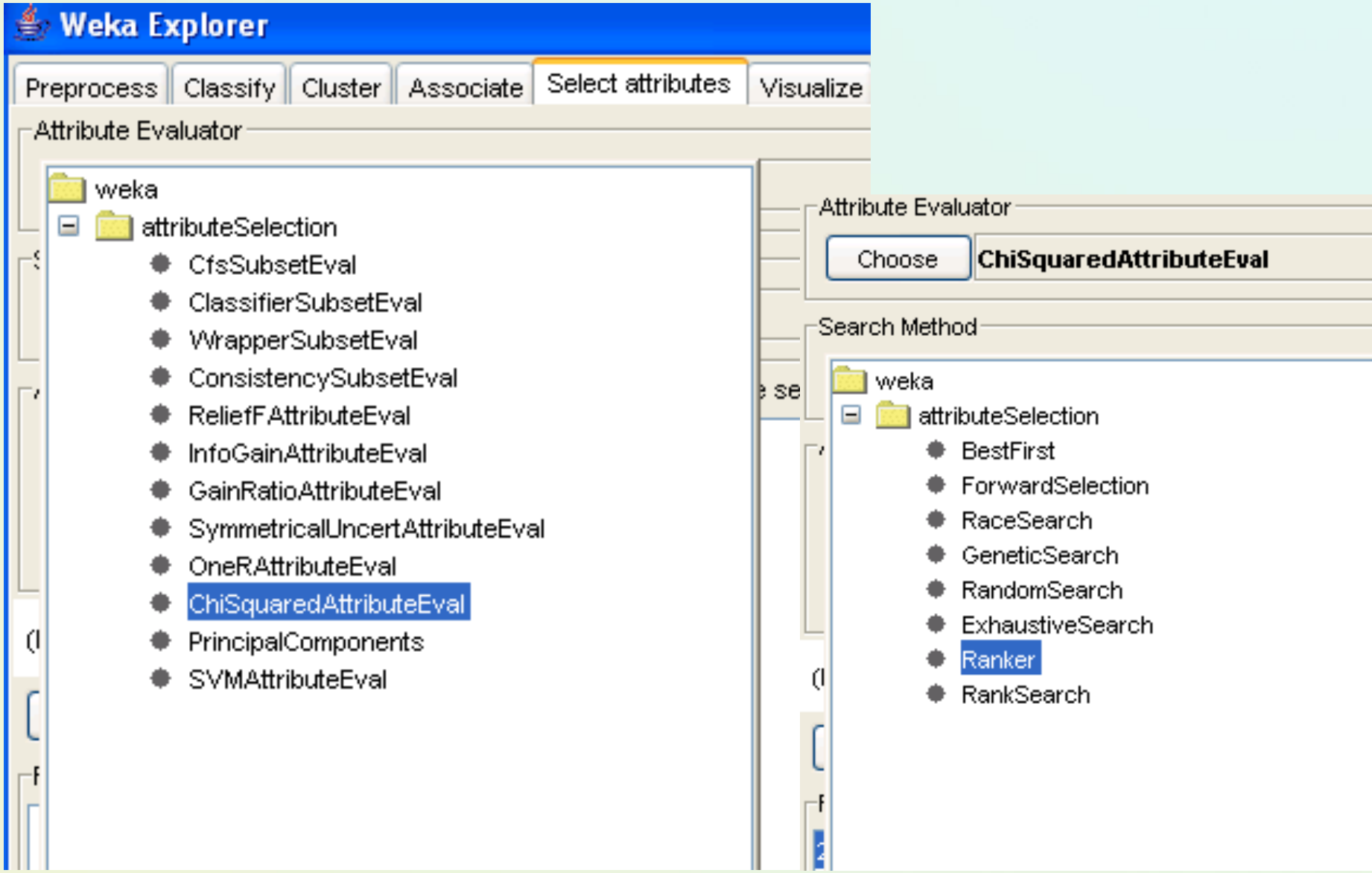
Redukcja rozmiarów danych – Selekcja atrybutów

- 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.*

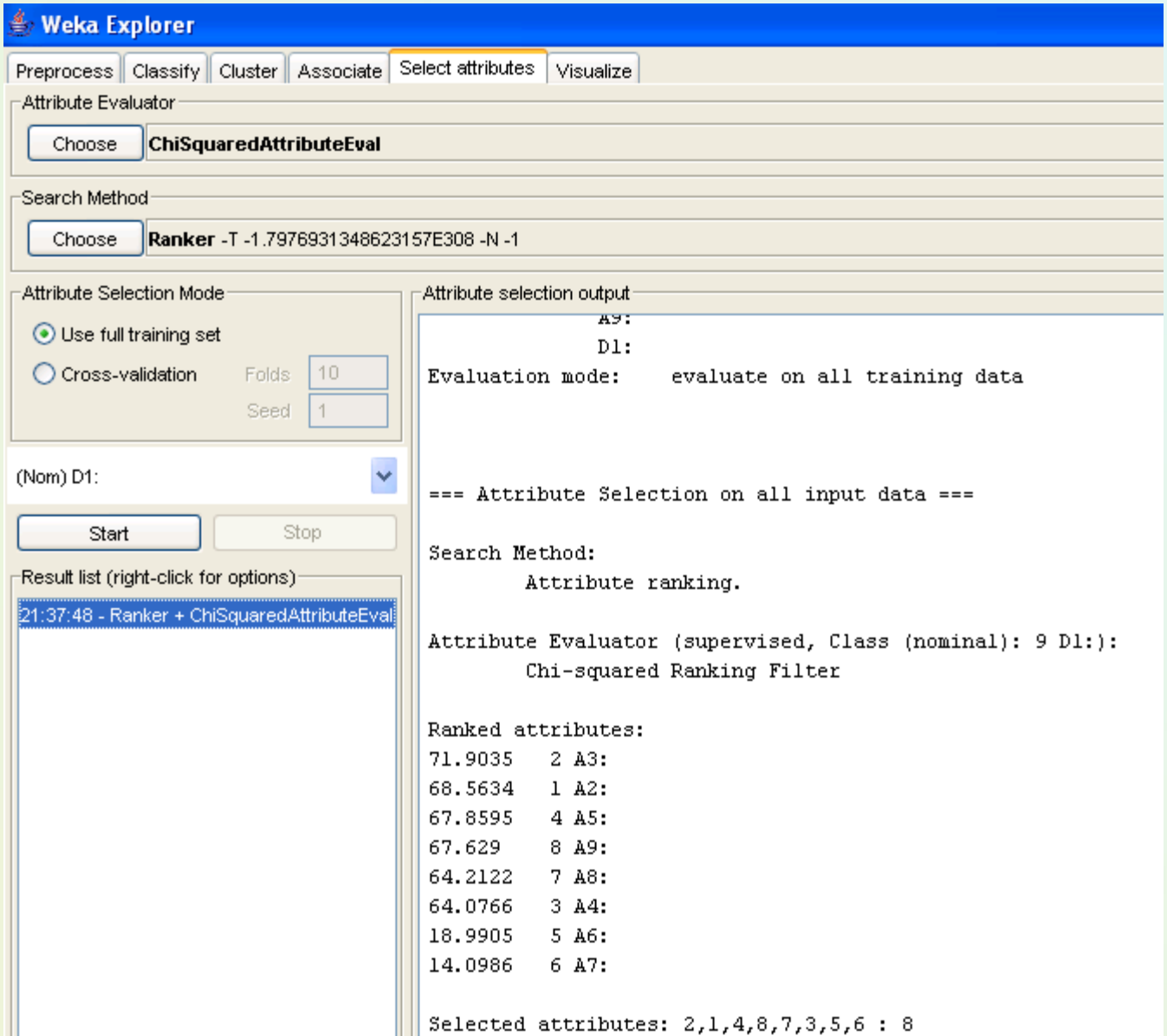
Terminologia i podział metod

- **Univariate method**: considers one variable (feature) at a time.
- **Multivariate method:** considers subsets of variables (features) together.
- **Filter method:** ranks features or feature subsets independently of the predictor (classifier).
- **Wrapper method:** uses a classifier to assess features or feature subsets.

WEKA – attribute selection



Ranking with …? WEKA



Inne biblioteki − filter z feature selection

- `SelectKBest` removes all but the k highest scoring features
- `SelectPercentile` removes all but a user-specified highest scoring percentage of features
- using common univariate statistical tests for each feature: false positive rate `SelectFpr`, false discovery rate `SelectFdr`, or family wise error `SelectFwe`.
- `GenericUnivariateSelect` allows to perform univariate feature selection with a configurable strategy. This allows to select the best univariate selection strategy with hyper-parameter search estimator.

For instance, we can perform a χ^2 test to the samples to retrieve only the two best features as follows:

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.feature_selection import SelectKBest
>>> from sklearn.feature_selection import chi2
>>> X, y = load_iris(return_X_y=True)
>>> X.shape
(150, 4)
>>> X_new = SelectKBest(chi2, k=2).fit_transform(X, y)
>>> X_new.shape
(150, 2)
```

These objects take as input a scoring function that returns univariate scores and p-values (or only scores for `SelectKBest` and `SelectPercentile`):

- For regression: `f_regression`, `mutual_info_regression`
- For classification: `chi2`, `f_classif`, `mutual_info_classif`

The methods based on F-test estimate the degree of linear dependency between two random variables. On the other hand, mutual information methods can capture any kind of statistical dependency, but being nonparametric, they require more samples for accurate estimation.

Feature selection with sparse data

If you use sparse data (i.e. data represented as sparse matrices), `chi2`, `mutual_info_regression`, `mutual_info_classif` will

Scikit-learn (sklearn) library:

Univariate selection
Recursive Feature Elimination (RFE)
Principle Component Analysis (PCA)
Choosing important features (feature importance)

Więcej w dokumentacji sklearn

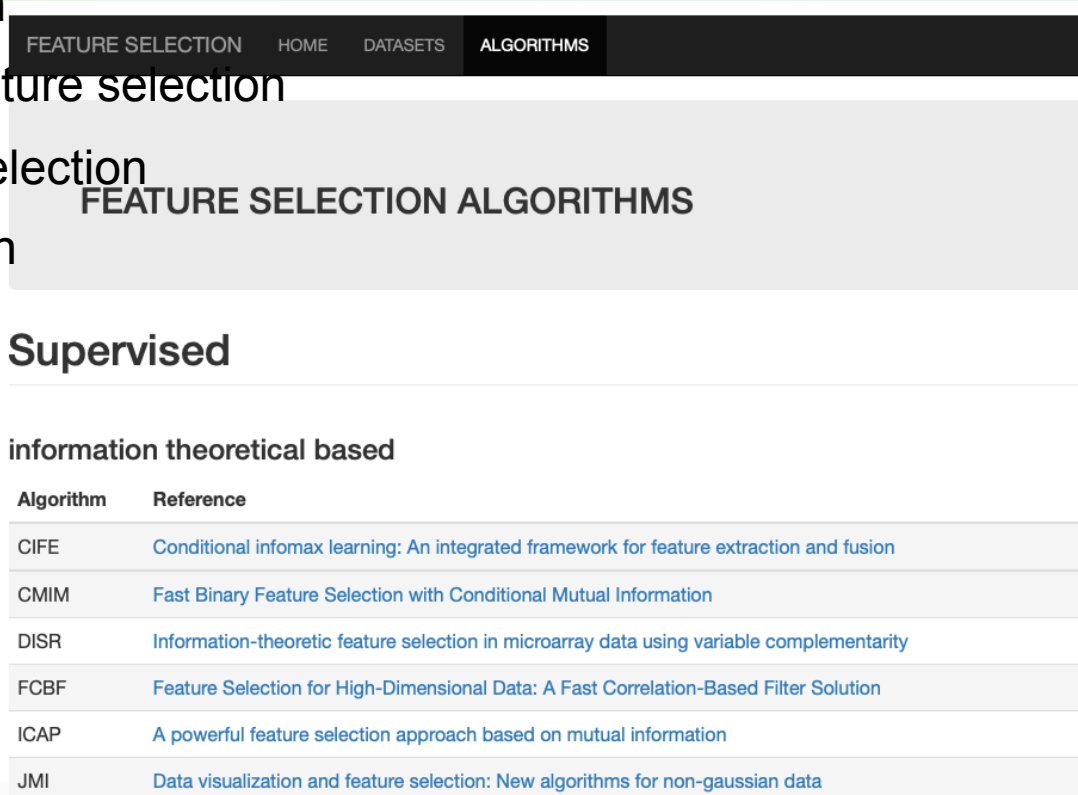
Dodatkowe bliblioteki

scikit-feature is an open-source feature selection repository in Python developed at Arizona State University

http://featureselection.asu.edu/algorithms.php

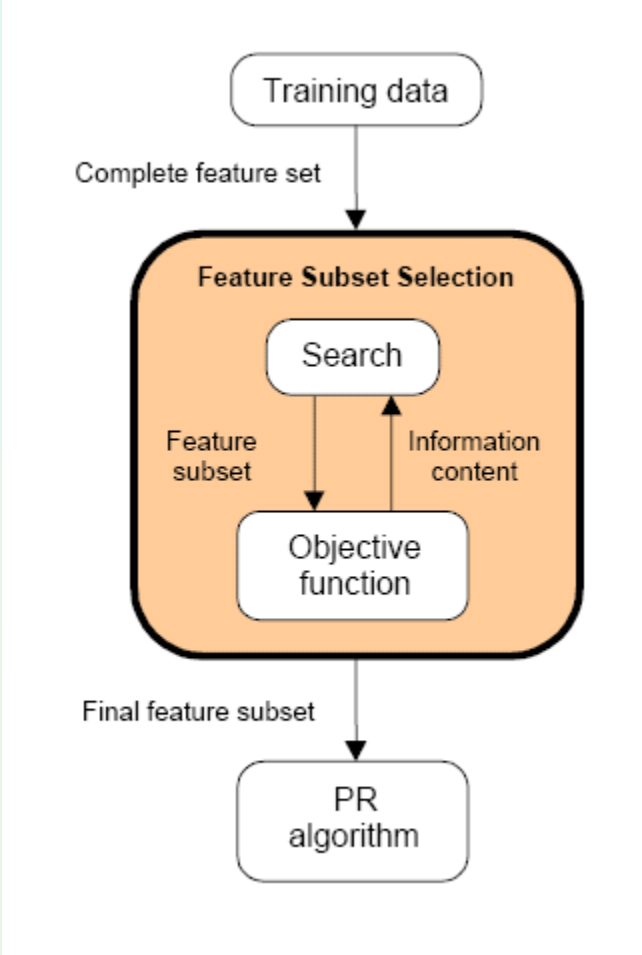
Oryginalnie 40 różnych metod:

- Similarity based feature selection
- Information theoretical based feature selection
- Sparse learning based feature selection
- Statistical based feature selection
- Wrapper based feature selection
- Structural feature selection
- Streaming feature selection



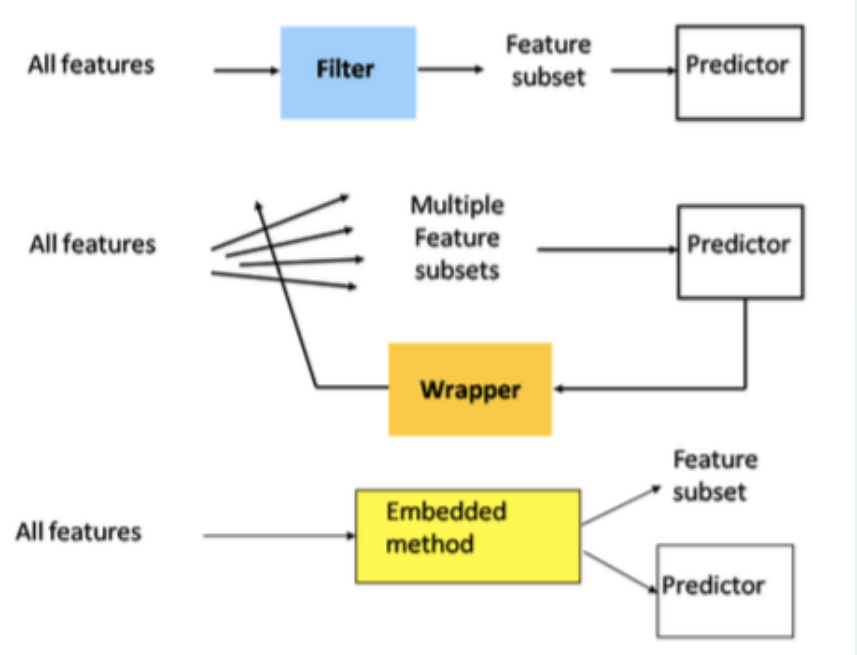
Problem selekcji atrybutów (Feature Selection)

- Optymalne rozwiązanie – NP.
- *n* liczba atrybutów
  → przegląd przestrzeni z 2^n stanami.
- Przestrzeń podzbiorów częściowo uporządkowana (różne struktury grafowe)
- Strategia przeszukiwania oraz
- Miara oceny *J*



Kategoryzacja podejść do selekcji

- Filtrujące, opakowujące, wbudowane
- Strategie przeszukiwania przestrzeni:
  - Kompletne
  - Heurystyczne
  - Losowe
- Kryteria oceny atrybutów oraz ich podzbiorów:
  - miary podobieństwa (odległościowe)
  - miary teorio-informacyjne
  - miary zależności (dependency / statistical)
  - miary spójnościowe
  - Inne (w tym bezpośrednia ocena klasyfikatora)



Podejścia rankingowe

Wejście zbiór cech / atrybutów $F=f_1,f_2,...,f_m$

- Dla każdej cechy $f \in F$ wylicz wartość miary oceny $J(f)$
- Uporządkuj cechy według wartości $J(f)$.
- Weź k cech z góry rankingu albo cechy, dla których zachodzi $J(f) > \delta$, gdzie δ jest ustalonym progiem.

Wyjście: Podzbiór najlepiej ocenionych cech.

Pojedyncze cechy + podobieństwo

Fisher Score

Istotne cechy − podobne wartości dla przykładów z tej samej klasy oraz różne dla przykładów z innych klas

$$S_i = \frac{\sum_{k=1}^{K} n_j (\mu_{ij} - \mu_i)^2}{\sum_{k=1}^{K} n_j \rho_{ij}^2}$$

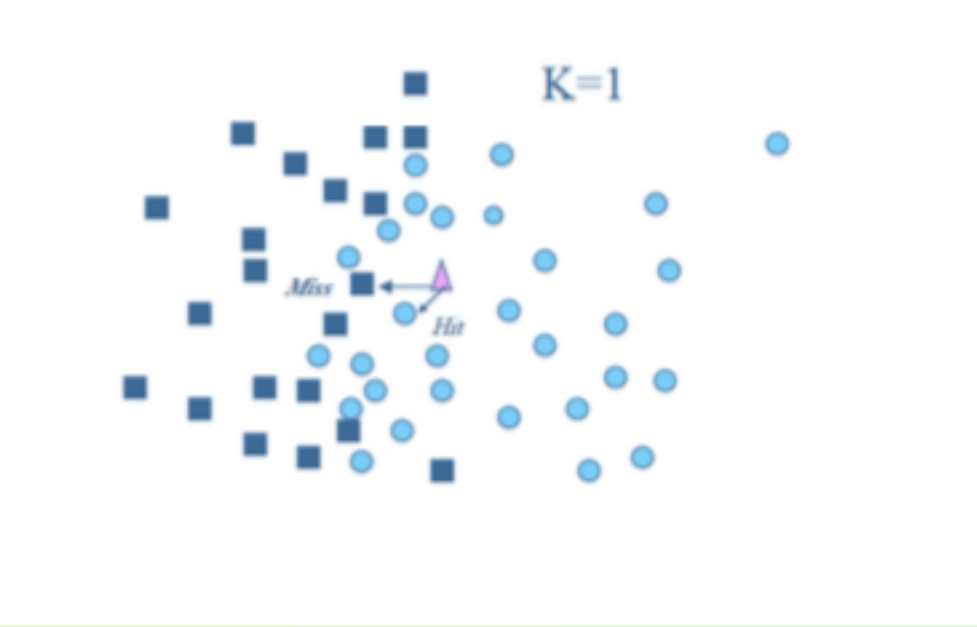
gdzie μij ρij to średnia i wariancja i-tej cechy w j-tej klasie; nj liczba przykładów w j-tej klasie a μi wartość średnia i-tej cechy

Poszukujemy cech maksymalizującymi wskaźnik

Przy ocenie indywidualnej cech − nie rozwiązuje redundancji oraz cech wymiennych

Relief

- Wykorzystuje ocenę odległości między przykładami z różnych klas w przestrzeni cech
  - Podstawą algorytm k-NN
  - Idea – przykłady z tej samej klasy powinny mięć podobne wartości atrybutów, a różnych powinny mieć różne
- Zaproponowany przez Kira, Rendell



Relief (oryginalny)

Input: for each training instance a vector of attribute values and the class value
Output: the vector W of estimations of the qualities of attributes

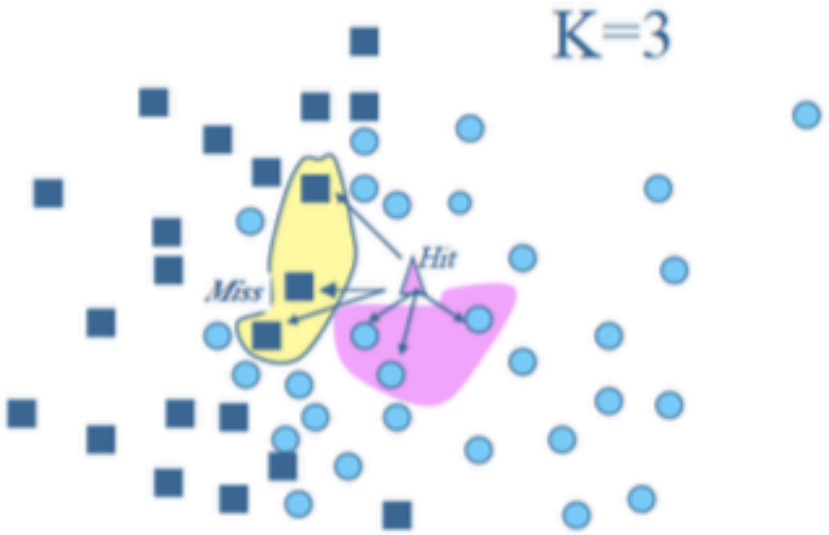
1. set all weights $W[A] := 0.0$;
2. **for** i := 1 **to** m **do begin**
3. randomly select an instance R_i;
4. find nearest hit H and nearest miss M;
5. **for** A := 1 **to** a **do**
6. $W[A] := W[A] - \mathrm{diff}(A, R_i, H)/m + \mathrm{diff}(A, R_i, M)/m$;
7. **end**;

oznaczenia:
R - instancja danych, zawierająca wartości a cech A i wartość klasy
m - parametr definiowany przez użytkownika

$diff(A, I_1, I_2)$ - różnica pomiędzy wartościami atrybutu A w instancjach I_1 i I_2

ReliefF (oryginalny)

- modyfikacja algorytmu Relief
- zaproponowany przez Kononenko et al. w 1997 roku
- zamiast odległości Euklidesa używa odległości Manhatan do poszukiwania sąsiadów
- zamiast jednego *H* i jednego *M* poszukuje *k* najbliższych *H* i *k* najbliższych *M*
- może być stosowany w klasyfikacji wieloklasowej



ReliefF - pseudokod

Input: for each training instance a vector of attribute values and the class value

Output: the vector W of estimations of the qualities of attributes

1. set all weights $W[A] := 0.0$;
2. **for** i := 1 **to** m **do begin**
3. randomly select an instance R_i;
4. find k nearest hits H_j;
5. **for** each class $C \neq class(R_i)$ **do**
6. from class C find k nearest misses $M_j(C)$;
7. **for** A := 1 **to** a **do**
8. $W[A] := W[A] - \sum_{j=1}^{k} \mathrm{diff}(A, R_i, H_j)/(m \cdot k) +$
9. $\sum_{C \neq class(R_i)} [\frac{P(C)}{1 - P(class(R_i))} \sum_{j=1}^{k} \mathrm{diff}(A, R_i, M_j(C))]/(m \cdot k)$;
10. **end**;

Contextual merit measure [Hong J.]

The main assumption of the CM measure, which was developed in [3] is that features important for classification should differ significantly in their values to predict instances from different classes. The CM measure assigns a merit value to a given feature taking into account the degree to which the other features are capable to discriminate between the same instances as the given feature. We calculate the value of CM measure CM_{f_i} of a feature f_i as it has been presented in [3] according the formula (2):

$$CM_{f_i} = \sum_{r=1}^{M} \sum_{\mathbf{x}_k \in \overline{C(\mathbf{x}_i)}} w_{\mathbf{x}_i \mathbf{x}_k}^{(f_i)} d_{\mathbf{x}_i \mathbf{x}_j}^{(f_i)}. \quad (2)$$

where M is the number of instances, $\overline{C(\mathbf{x}_i)}$ is the set of vectors which are not from the same class as the vector $\mathbf{x}_i$, and $w_{\mathbf{x}_i \mathbf{x}_k}^{(f_i)}$ is a weight chosen so that vectors that are close to each other, i.e., they differ only in a few features, have greater influence in determining each feature's CM measure. In [3] weights $w_{\mathbf{x}_i \mathbf{x}_k}^{(f_i)} = \frac{1}{D^2_{\mathbf{x}_i, \mathbf{x}_k}}$ were used when $\mathbf{x}_k$ is one of the K nearest neighbors of $\mathbf{x}_i$ in terms of the distance $D_{\mathbf{x}_i, \mathbf{x}_k}$ between the vectors $\mathbf{x}_i$ and $\mathbf{x}_k$, in the set $\overline{C(\mathbf{x}_i)}$, and $w_{\mathbf{x}_i \mathbf{x}_k}^{(f_i)} = 0$ otherwise. K is the binary logarithm of the number of vectors in the set $\overline{C(\mathbf{x}_i)}$ as in [3].

Idea kontekstu innych cech + odległości przykładów z tej samej klasy

Metody wykorzystujące information gain

- Dobra intepretacja i łatwość obliczeń

$$IG(A_i,C) = H(A_i) - H(A_i|C)$$

Max Info gain lub gain ratio

Lecz nie uwzględnia obecności cech redundantnych o wymiennej roli

Inne propozycje − mutual information oraz simetrical uncertainty

$$SU(X,Y) = 2\frac{IG(X,Y)}{H(X)+H(Y)}$$

Cecha X jest “predominant” dla C, jeśli SU(X,C)>ρ i nie ma innej cechy Y takiej że S(X,Y) > S(X,C)

Cecha X jest nadmiarowa wobec cechy Y, jeśli S(X,Y) > S(X,C)

Informacja wzajemna (Mutual Information)

- Miara zależności zmiennych losowych X i Y, wywodząca się z teorii informacji
- Ocenia ile informacji o X można poznać, znając Y / o ile poznanie jednej z tych zmiennych zmniejsza niepewność o drugiej
  - Jeśli zmienne X i Y są niezależne, to ich wzajemna informacja jest zerowa (znajomość jednej nie mówi niczego o drugiej)
  - Jeśli X i Y są identyczne, to każda zawiera pełną wiedzę o drugiej

Definicja

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x) \cdot (y)}$$

gdzie p(x,y) oznacza wspólny rozkład prawdopodobieństwa (ang. joint probability distribution) X i Y, a p(x) i p(y) oznaczają prawdopodobieństwa w rozkładach zmiennych X i Y

Entropia − interpretacja redukcji niepewności I(X,Y) = H(X) − H(X|Y) = H(Y) − H(X|Y)

W selekcji - max. związaek cechy/ cech X i wyjścia y

Porównanie niektórych pojedynczych

Summary important methods for evaluation in filter method

Name	Key Concept	Mathematical formula	Merits	Demerits
Information gain (IG)	For higher IG, feature is more relevant	IG(D|X) = I(D)- I(D|X)	Robust to the effect of outlier and act as white box classifier	Splits of features are very sensitive to training data.
Correlation Coefficient (CC)	For higher value of CC, Features are more redundant	$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{n\sigma_a\sigma_b}$	It is easy to work out and easy to interpret	It measures only linear relationships between features.
Symmetric uncertainty (SU)	Feature having higher value with target class is better	$SU(X_k, w) = 2\left[\frac{I(X_k, w)}{H(X_k) + H(w)}\right]$	Appropriate for high dimensional data set	It fails to take into consideration the interaction between features

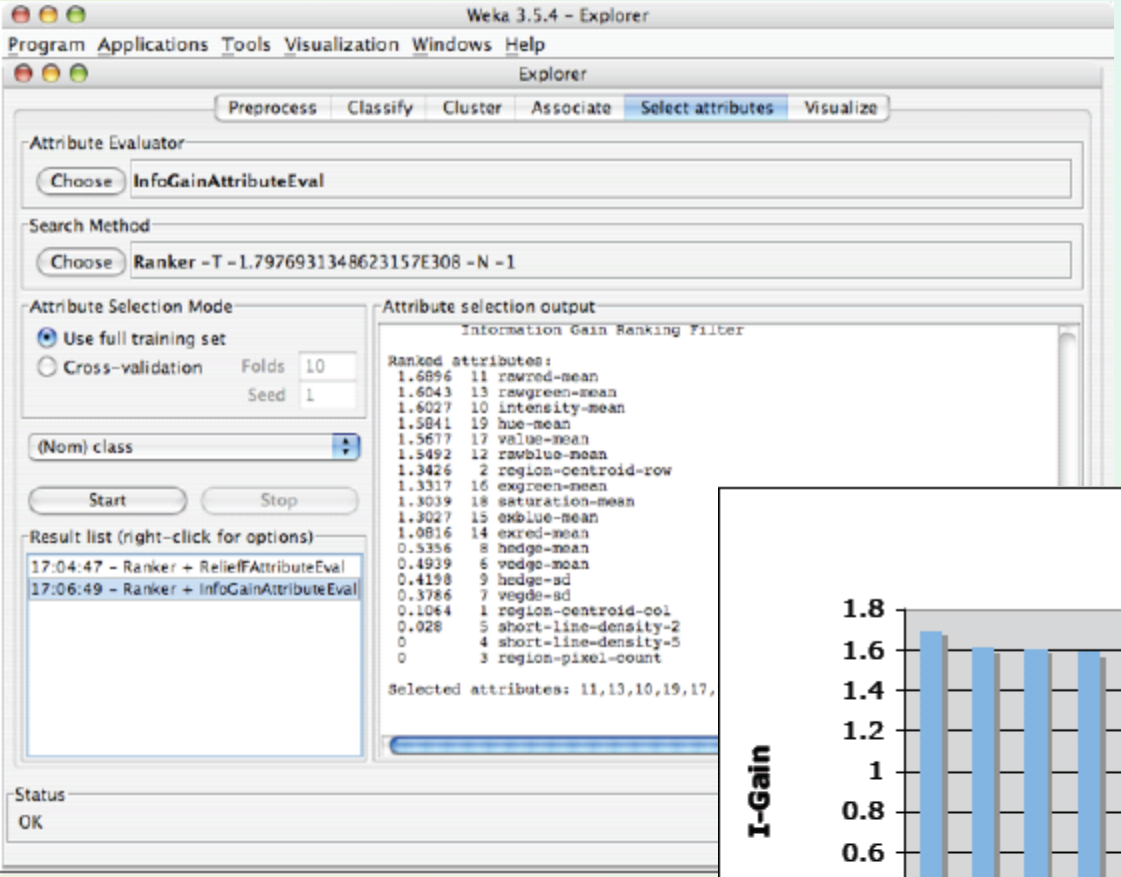
Podejścia rankigowe

Na ogół proste i efektywne obliczeniowo − rzędu m (liczby obiektów) i n (liczby atr).

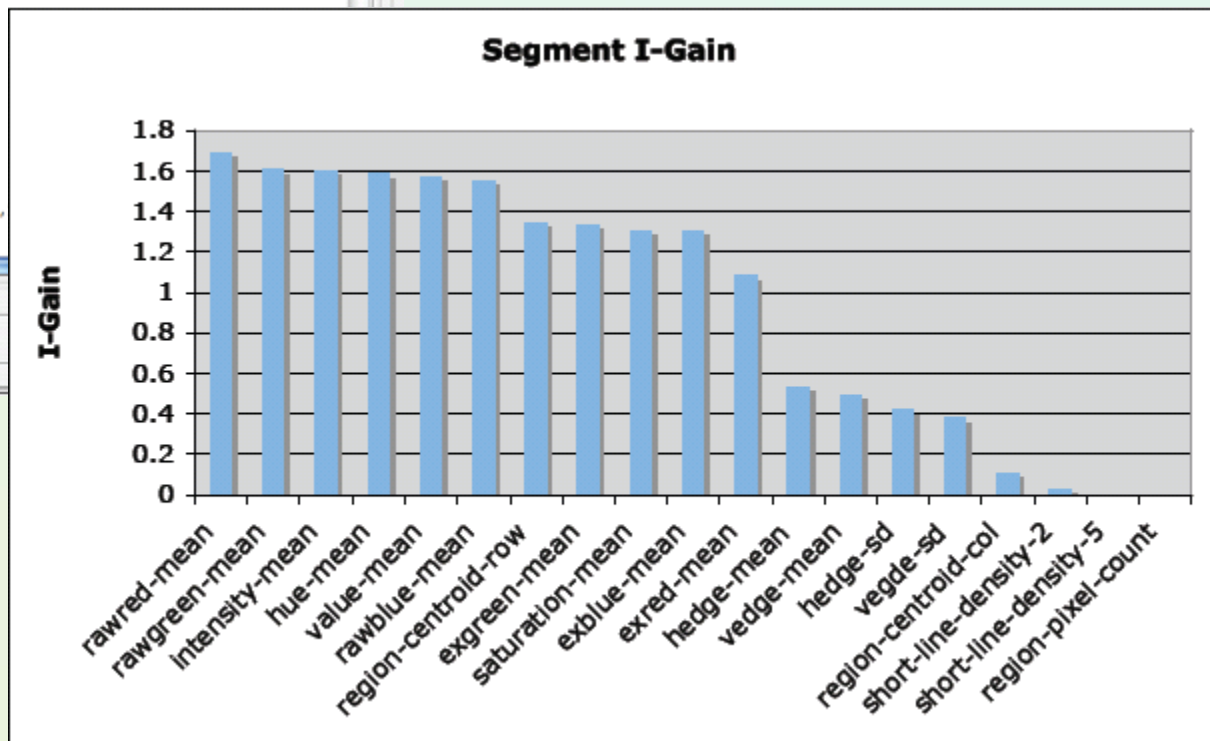
Lecz

- Nie eliminuje atrybutów redundantnych (pełniących wymienną rolę)
- Nie uwzględnia wzajemnych zależności oraz synergii podzbiorów cech
- Trudność wyboru konkretnej liczny cech

Jak wykorzystać ranking atrybutów

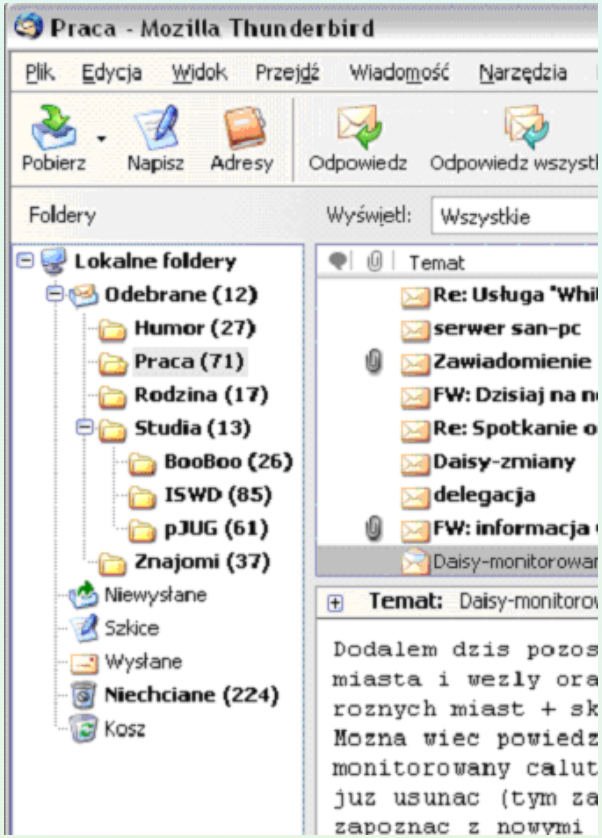


- Wybierz powyżej progu τ
- Mediana czy raczej percetyle?



Przykład selekcji cech: Kategoryzacja e-majli

- Motywacje:
  - Strumień danych bardzo wielu wiadomości.
  - Wspomagać użytkownika w organizacji tematycznej zbyt wielu wiadomości.
  - Wiele kategorii (folders) → coś więcej niż wyłącznie identyfikacja spam-u.
  - Problem mniej zbadany niż inne zadania klasyfikacji typowych tekstów.
- Studium przypadku - J.Stefanowski, M.Zienkowicz: Classification of Polish Email Messages: Experiments with Various Data Representations 2007



Zbyt wiele cech …

- Pozyskano kilka tysięcy cech!
  - Pomimo usuwania zbyt częstych słów nadal duży rozmiar danych (kilka tysięcy).
  - Trudności dla dalszej fazy uczenia się klasyfikatorów.
- Selekcja atrybutów → metody filtrowania oparte na funkcjach informacyjnych i statystykach $\chi 2$ → redukcja do rzędu kilkuset atrybutów oraz wzrost trafności o ok. 10%
- Więcej praca mgr M. Zienkowicz

Dane	Header	Servers	Subject	Body	Attach
MarCas	99	54	15	160	75
DW	42	28	11	137	283
MNews	24	0	12	198	0

Liczności zbiorów atrybutów po selekcji

Analiza strumienia danych tekstowych

Wspólna praca B.Szopka i J.Stefanowski 2007 / Enron emails

Liczba atrybutów			
brak selekcji	selekcja AS1	selekcja AS2	selekcja AS3
7297	191	119	40
3924	799	406	183
5381	650	287	102
9214	361	216	49
5744	1160	660	298
2655	580	295	160
3052	678	340	151
2542	589	328	208
2644	729	512	401
3032	495	335	141

algorytm	*Średnia trafność klasyfikacji*	
	brak selekcji	selekcja AS2
NB	39.98 ± 5.89	51.33 ± 10.14
kNN	24.57 ± 11.71	50.28 ± 7.13
C4.5	47.11 ± 6.47	49.55 ± 7.57
SVM	34.71 ± 7.81	51.31 ± 9.19
NB	71.17 ± 6.91	68.38 ± 8.64
kNN	59.40 ± 10.63	67.76 ± 8.45
C4.5	65.27 ± 9.12	66.53 ± 9.05
SVM	68.32 ± 9.66	70.23 ± 9.36
NB	45.27 ± 3.86	45.46 ± 3.74
kNN	23.38 ± 4.86	37.79 ± 4.18
C4.5	34.87 ± 5.40	40.50 ± 5.74
SVM	41.71 ± 4.15	41.54 ± 5.03
NB	24.53 ± 2.21	15.78 ± 5.98
kNN	22.12 ± 4.20	26.65 ± 2.21
C4.5	33.08 ± 3.03	33.07 ± 3.06
SVM	33.07 ± 3.06	33.06 ± 3.05
NB	40.38 ± 13.21	37.63 ± 10.53
kNN	34.83 ± 18.19	37.23 ± 17.15
C4.5	41.64 ± 12.94	41.19 ± 16.57
SVM	44.88 ± 19.82	43.47 ± 18.50
NB	70.44 ± 4.43	68.52 ± 6.02
kNN	65.30 ± 7.78	70.61 ± 5.42
C4.5	74.24 ± 8.19	73.69 ± 7.42
SVM	72.44 ± 8.22	74.74 ± 6.18
NB	64.83 ± 15.89	66.00 ± 13.50
kNN	50.57 ± 15.54	59.90 ± 15.96
C4.5	54.25 ± 12.40	57.94 ± 12.42
SVM	63.84 ± 18.92	67.02 ± 17.48

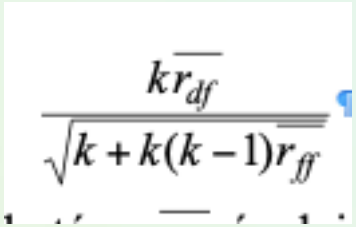
Inne ożliwe miary oceny

- Ocena pojedynczych atrybutów:
  - testy $\chi 2$ i miary siły związku,
  - miary specjalizowane odległości,
  - Contextual merit index
  - Consistency measures
  - ...
- 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*)
  - …

Ocena pozbiorów

Correlation-based merit measure

- Oblicza się korelację między wszystkimi parami dwoma zmiennych (atrybutów) *r*
- „Dobroć” zbiory podzbioru atrybutów *F* w stosunku do atrybutu decyzyjnego *d*



- gdzie *F* zawiera *k* atrybutów, r(df) średnia korelacja atrybutu *f* z klasyfikacją (atrybut *d*); r(ff) średnia wzajemna korelacja atrybutów (obliczona z *r*)

Dogodne dla atrybutów liczbowych − atrybuty jakościowe przybliżenie poprzez binaryzacje (transformacje do wielu atr.)

Selekcja stat. podzbioru zmiennych

- W analizie dyskryminacyjnej uwzględniaj zmienne o dobrych właściwościach dyskryminujących
- Przykład kryterium jakości dyskryminacji Wilksa:

$$\lambda = \frac{|S_w|}{|S_W + S_B|}$$

gdzie macierz zmienności wewnątrzklasowej

$$S_W = \frac{1}{n-k} \sum_{j=1}^{k} \sum_{i \in C_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T$$

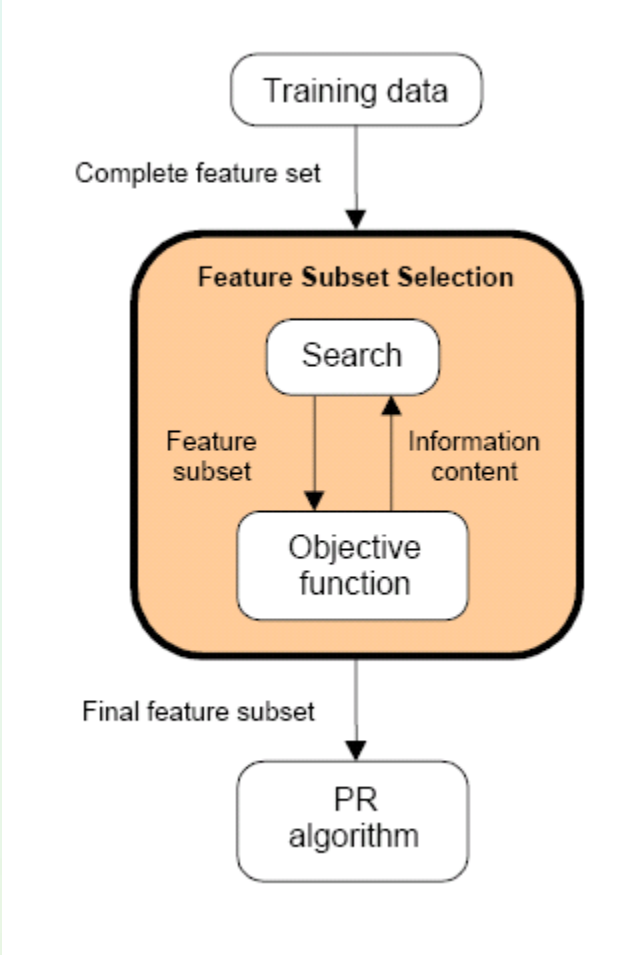
a macierz zmienności międzyklasowej

$$S_B = \frac{1}{k-1} \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$$

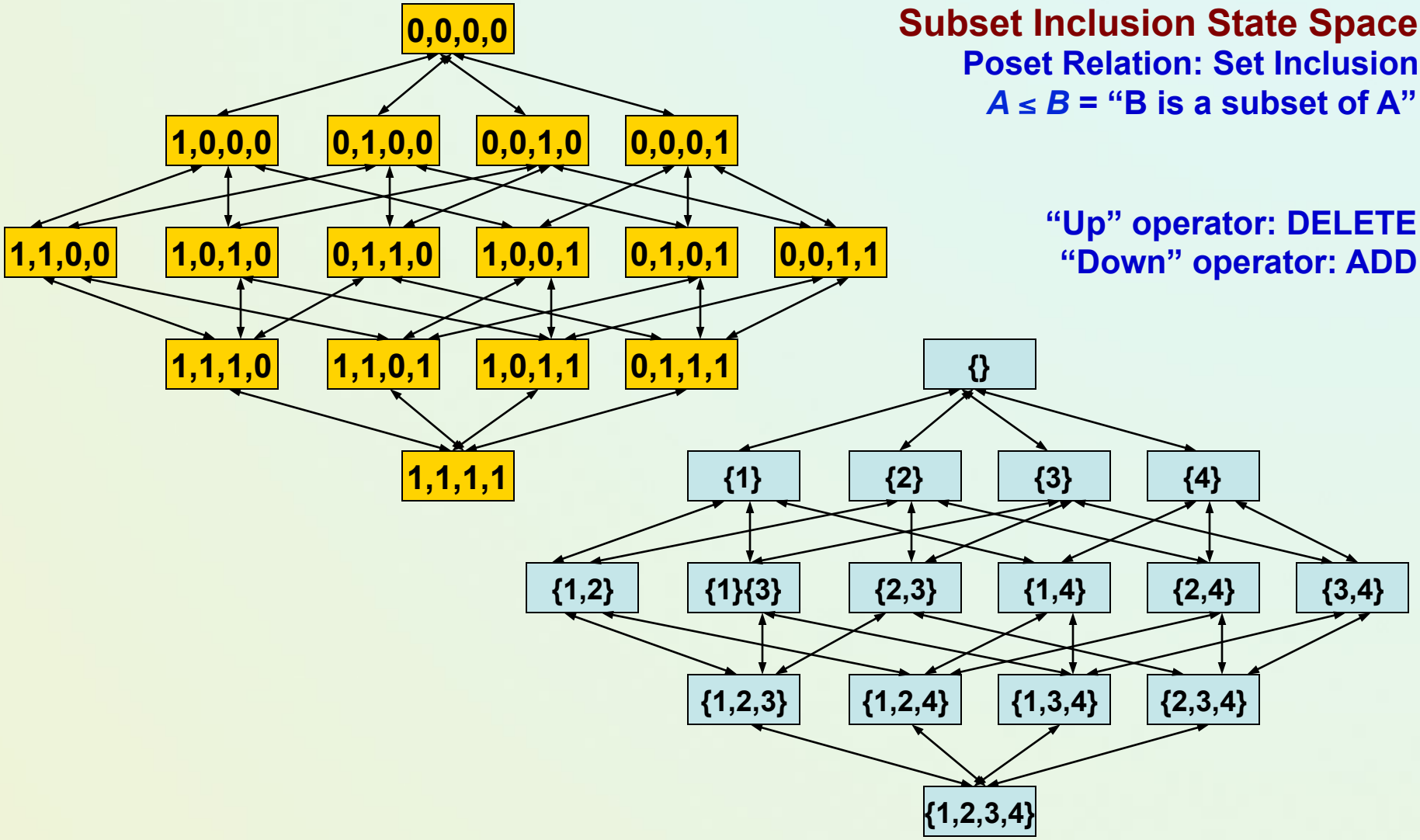
im mniejsza wartość λ tym silniejsze zróżnicowanie między klasami.

Problem selekcji atrybutów (Feature Selection)

- Optymalne rozwiązanie – NP.
- *n* liczba atrybutów
  → 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

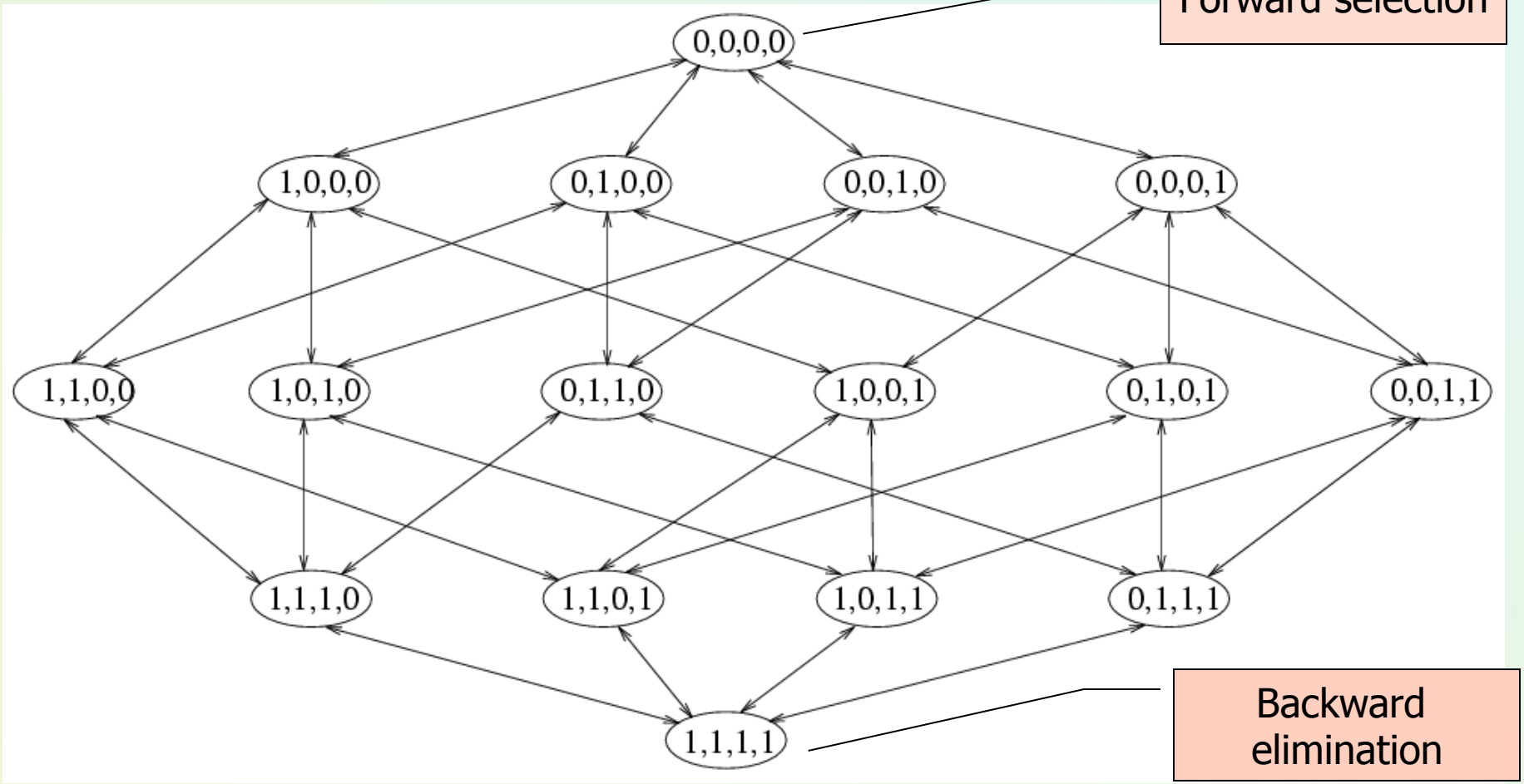
***A* ≤ *B* = "B is a subset of A"**

"Up" operator: DELETE

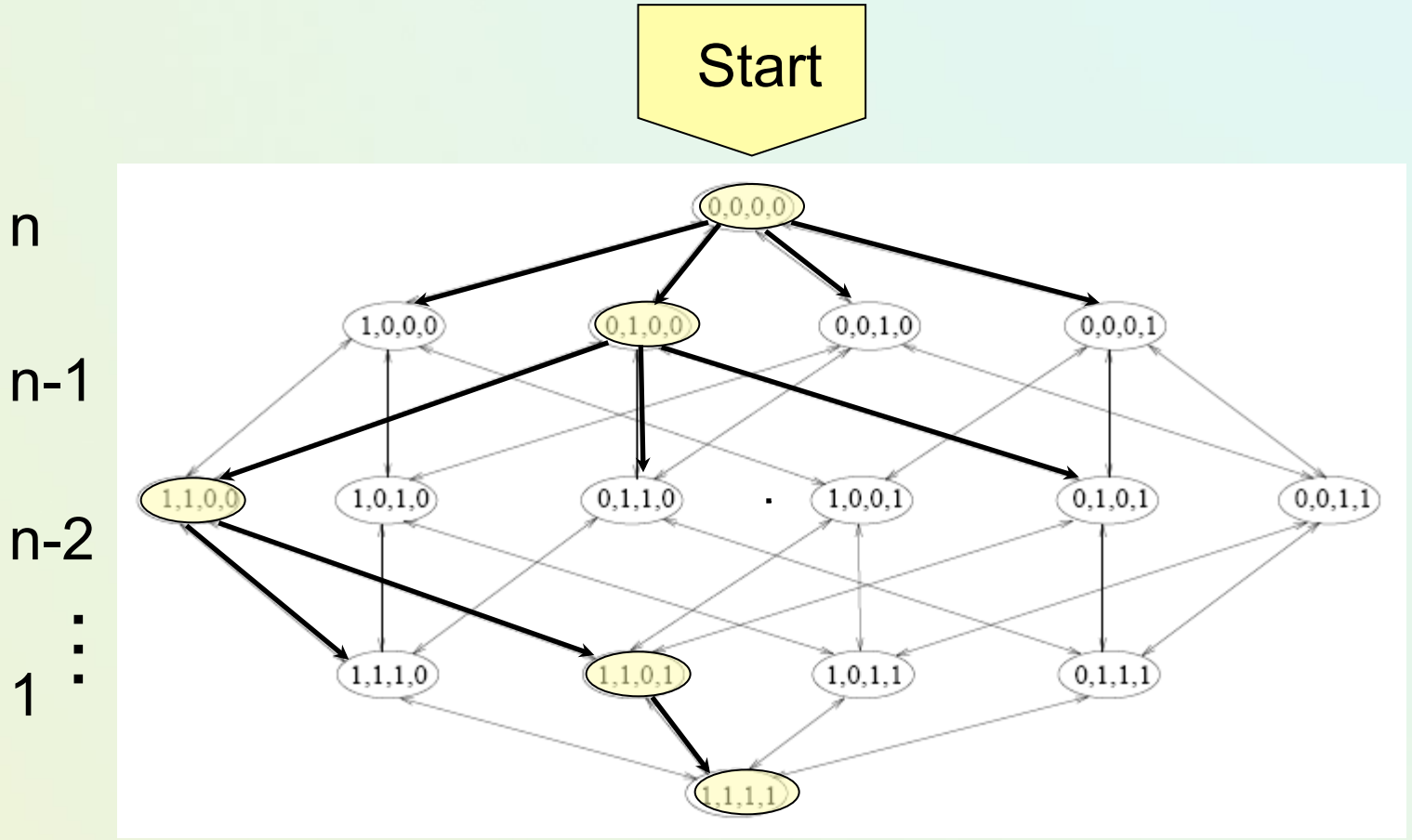
"Down" operator: ADD

Przeszukiwanie przestrzeni stanów

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



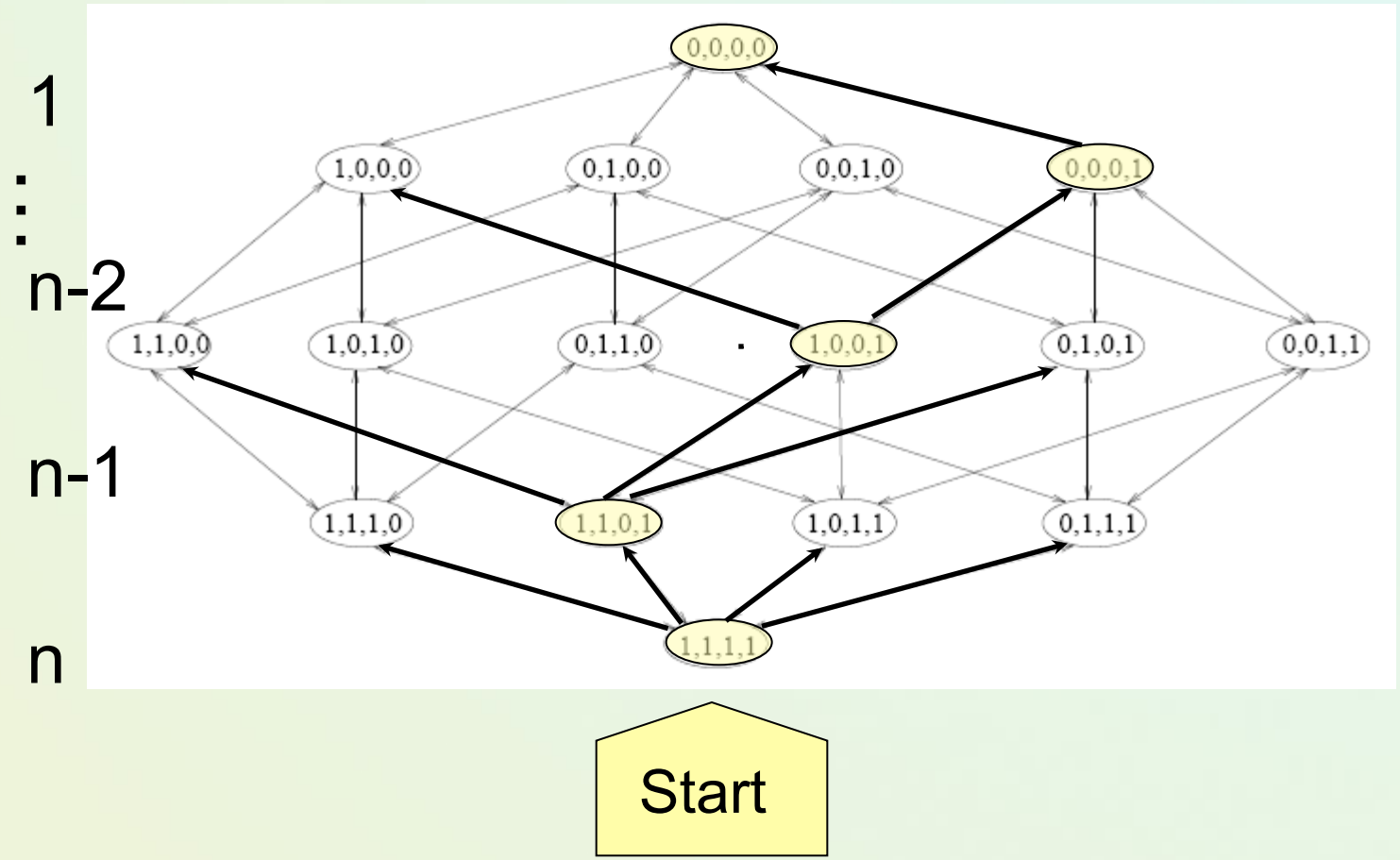
Forward Selection (wrapper)



Also referred to as SFS: Sequential Forward Selection

Backward Elimination (wrapper)

Also referred to as SBS: Sequential Backward Selection



Strategie poszukiwania podzbiorów cech

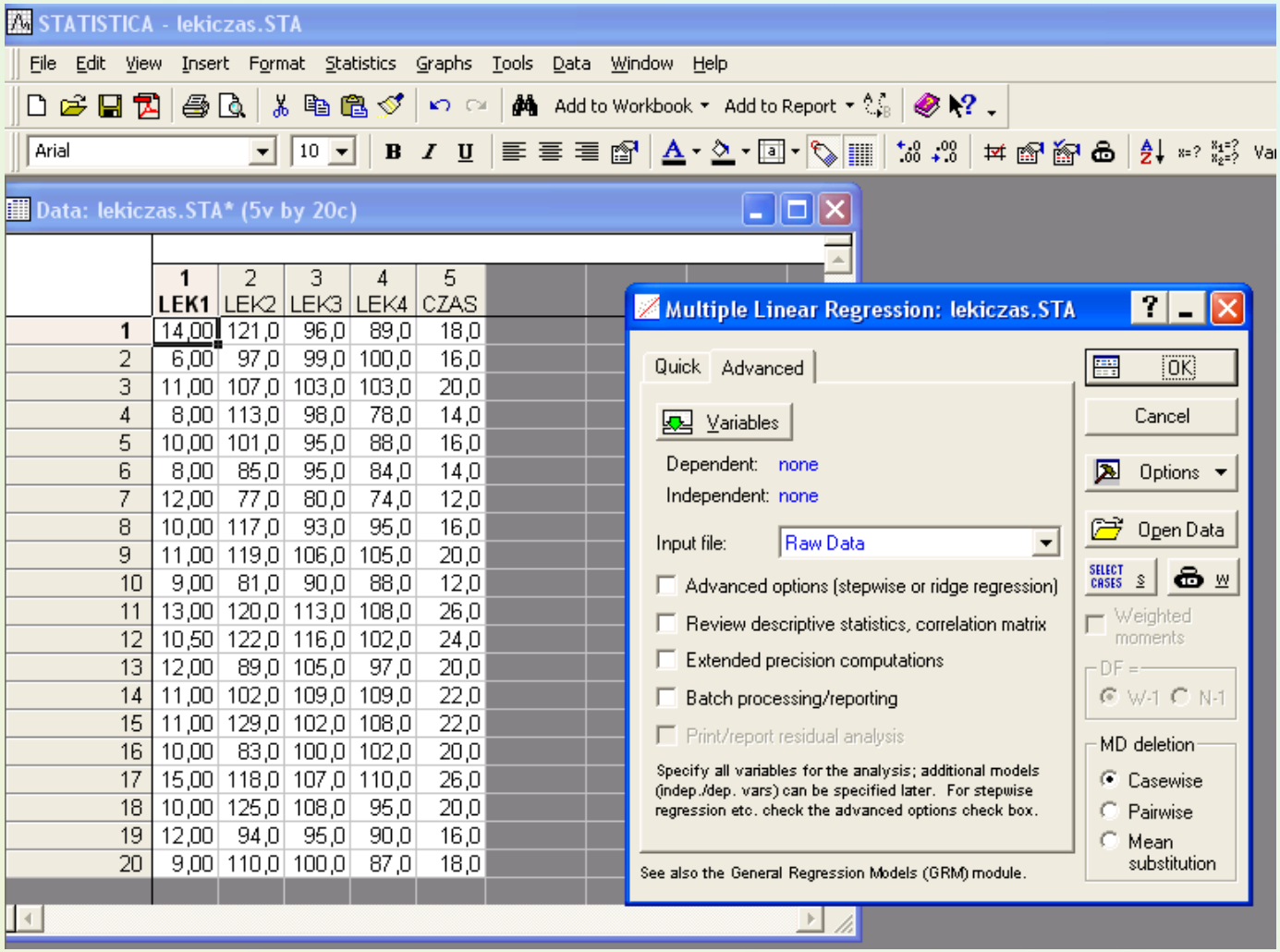
- commonly used search strategies:
  - **forward selection**
    - FSubset={}; greedily add features one at a time
  - **forward stepwise selection**
    - FSubset={}; greedily add or remove features one at a time
  - **backward elimination**
    - FSubset=AllFeatures; greedily remove features one at 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

Regresja wielowymiarowa - krokowa

- Postępująca (*forward*)
  - Zakłada kolejne dołączanie do listy zmiennych objaśniających tych zmiennych, które mają najistotniejszy wpływ na zmienną zależną.
- Wsteczna (*backward*)
  - Usuwamy ze zbioru zmiennych, ta które mają najmniejszy wpływ na zmienną zależną.
  - Stosując R2 lub testy istotności współczynników modelu (*F*).

Przykład forward selection – medycyna

Predict the length of patient's stay in a hospital depending on prescribed medicines.



Consistency measure

Analiza przypadków “inconsistency” w opisach przykładów [Dash, Liu, Motoda]

- two patterns are considered inconsistent if they match all but their class labels for example an inconsistency is caused by two instances (0;1; a) and (0;1; neg a) with different classes
- the inconsistency count for a pattern is the number of times it appears in the data minus the number of the largest class among them
- the inconsistency rate is the sum of all the inconsistency counts for all possible patterns of a feature subset divided by the total number of patterns

Wykonaj przeszukiwanie przestrzeni cech w celu znalezienia podzbioru o najlepszej wartości tej miary [B&B]

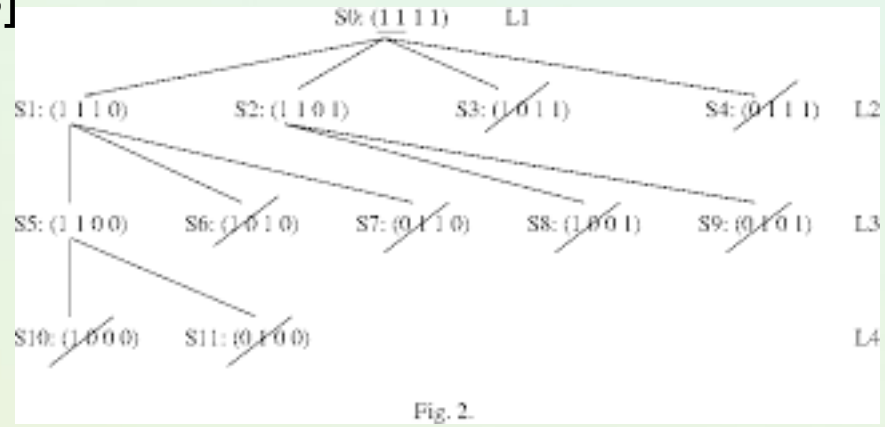
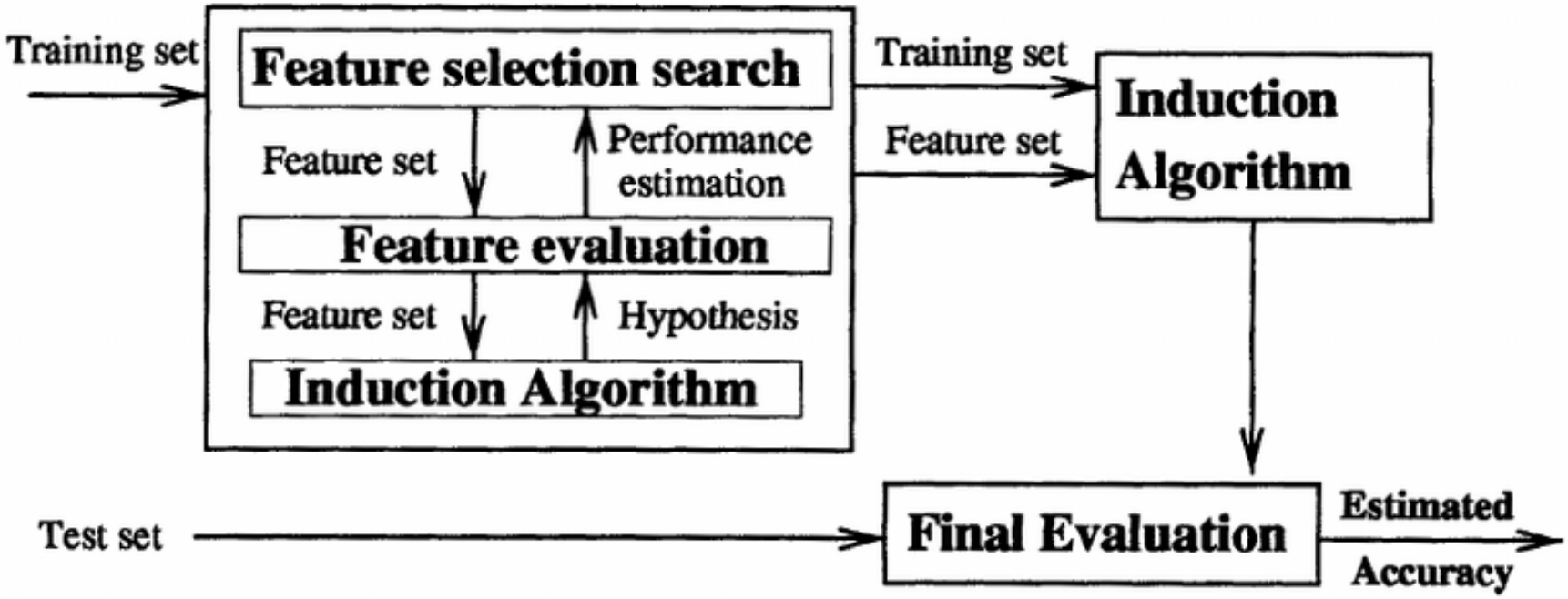


Fig. 2.

Wrapper [Kohavi et al.]



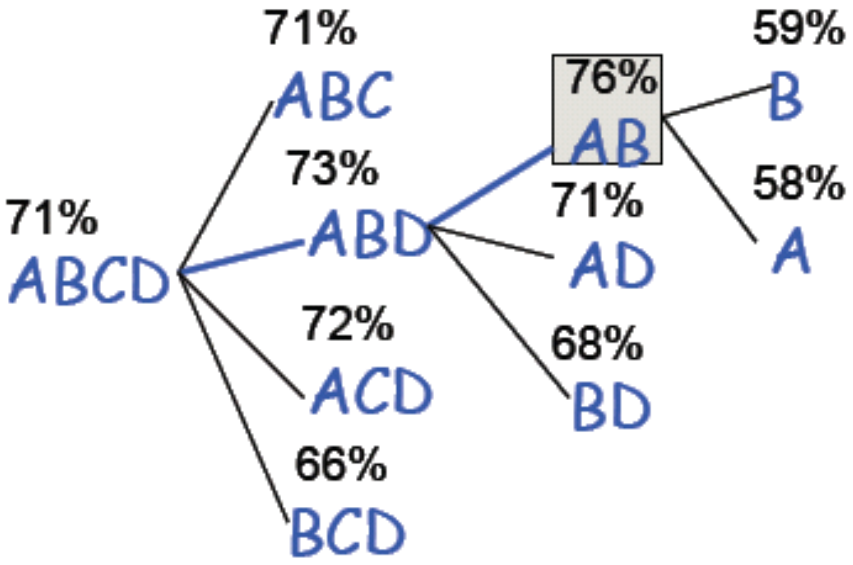
Za Holger Fröhlich et al. Feature Selection for Support Vector Machines by Means of Genetic Algorithms.

Ukierunkowany na optymalizacje wybranego klasyfikatora

Kosztowny obliczeniowo

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”

Motywacje dla selekcji cech

- Pamiętaj, że niektóre metody mają możliwości wbudowana selekcji:
  - decision trees: użycie kryteriów info gain, gain ratio, Gini indeks skupia uwagę na najbardziej „informatywnych” cechach.
  - Random forests: ocena ważności atrbutów
  - neural nets: backprop wzmacnia określone połączenia.
  - kNN, IBL: strojenie wag w rozszerzonej definicji miary odległości
  - Bayes nets: analiza statystyk (tabele rozkładu) wskazuje na ważność cech.

Formalizacja dla niektórych podejść

Część algorytmów minimalizuje złożoną funkcję:

$$\min_{\alpha} \hat{R}(\alpha, \sigma) = \min_{\alpha} \sum_{k=1}^{m} L(f(\alpha, \sigma \circ x_k), y_k) + \Omega(\alpha)$$

$G(\sigma)$

Empirical error

Regularization capacity control

$$L = \sum_i (y_i - \sum_p \beta_p x_{ip})^2 + \lambda \sum_p |\beta_p|$$

Przykład − regresja wielowymiarowa z metodą LASSO gdzie λ współczynnik kary i wiele βp -> 0

Embeded trees− analiza feature importance

Najprostsza heurystyka analizy struktury drzewa – im wyżej i częściej występuję warunek z atrybutem

Oceń dla warunku redukcję miary (impurity) oraz wagę – liczbę przykładów w węźle

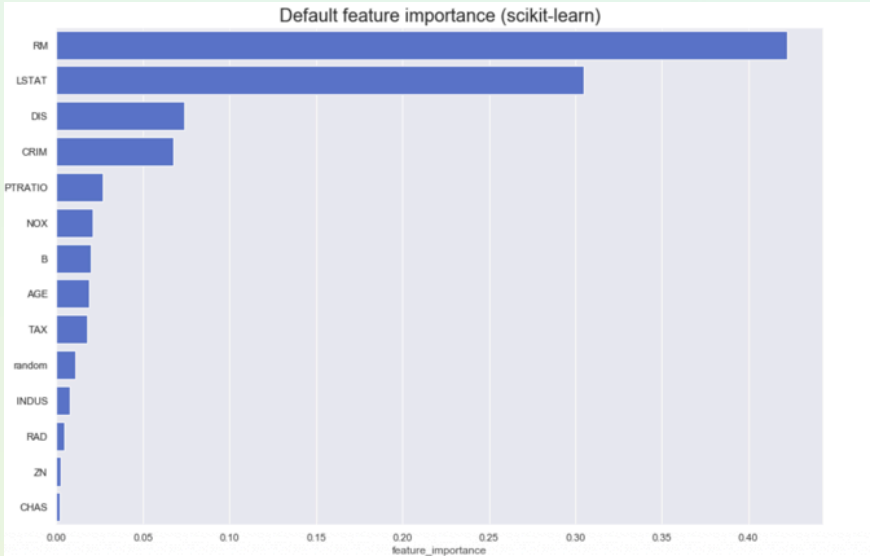
np. w węźle t dla atrybutu A spadek entropi Δ(t)=0.1 oraz N = 60 przykładów, to ocena atr A = 0.1*60=6

Sumuj takie wystąpienia w całym drzewie (lub zespole drzew)

feature_importances_ in Scikit-Learn

Pro: − szybkie obliczenia

Cons: - przybliżone obliczenia i może nadmiernie faworyzować wielo_ wartościowe oraz liczbowe atrybuty



Permutations − “noised-up” method

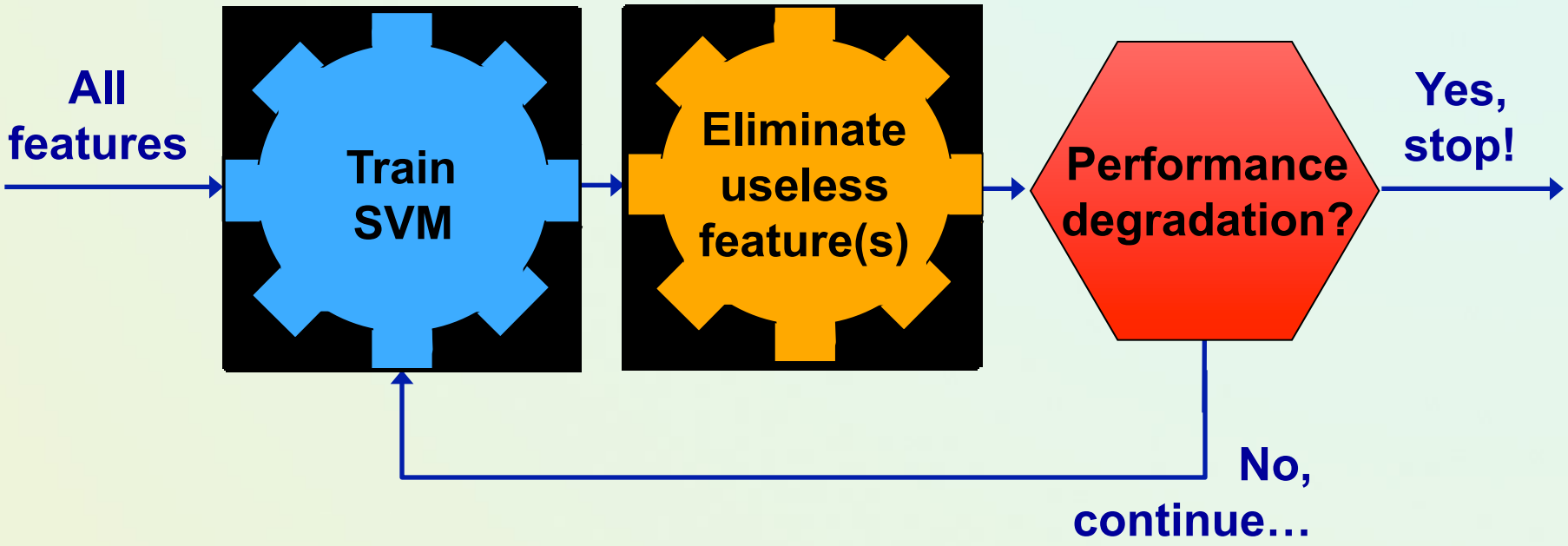
- L. Breiman wprowadził dla innego rankingu ważności cech poprzez analizę klasyfikowania w random forests / bagging
- Hipoteza − cecha jest tym ważniejsza, jeśli losowa zmiana jej wartości (permutacja) w zbiorze testowych mocno pogarsza trafność predykcji
- W bagging / RF − wykorzystuje się zbiór out of bag (OOB)dla każdego z drzew
  - najpierw ocenia się predykcję z oryginalnymi cechami O;
  - nastepnie dla każdej z cech tworzy zastępczy zbiór O’:
    losowo permutuje się wartości cechy i ponownie testuje klasyfikator

  zachowuje się informacje o predykcjach w każdej z tych sytuacji
- Po zbudowaniu całego zespołu:
  - Dla każdego z przykładów uczących zi (1...N) – określ predykcje klasyfikatora używając pamięci testowania OOB jeśli zawierają zi – i oblicz dec. zespołu; następnie łączny bląd klasyfiowania e
  - Analogicznie dla każdej cechy xj oblicz z pamięci OOB – O’ bląd zespołu klasyfikatora
- Ważność cechy Fj = ej − e Im większa, tym ważniejsza xj

Pro: b. wiarygodna; stosowalna także dla innych klasyfikatorów niż drzewa

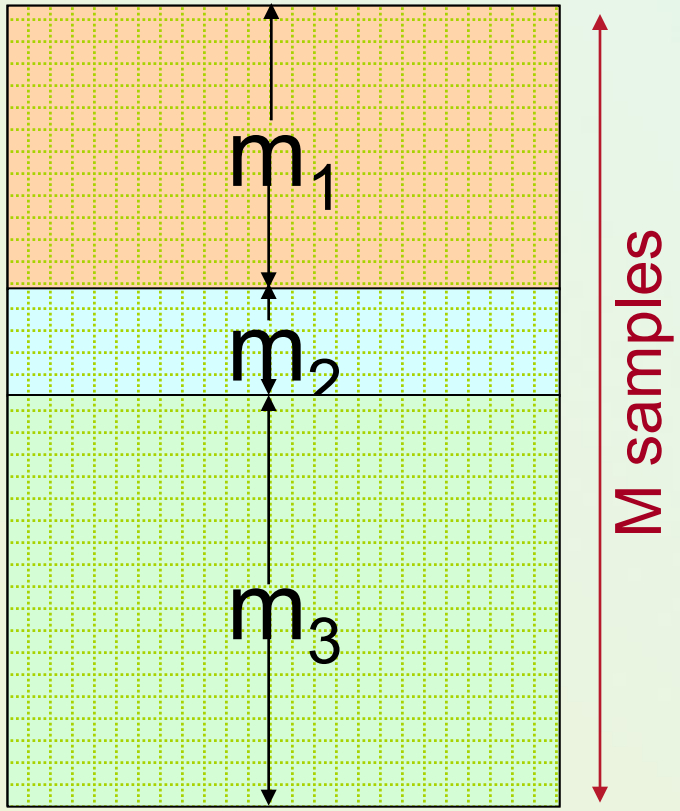
Cons: przeszacowuje skorelowane atrubuty, kosztowniejsza, nie w podst. bibliotekach

Permutation − masking features



Recursive Feature Elimination (RFE) SVM. *Guyon-Weston, 2000. US patent 7,117,188*

Poszukiwanie eksperymentalne



Split data into 3 sets:
training, validation, and test set.

1) For each feature subset, train predictor on training data.

2) Select the feature subset, which performs best on validation data.

   - Repeat and average if you want to reduce variance (cross-validation).

3) Test on test data.

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}

Podsumowanie

- Właściwe przygotowanie danych jest bardzo istotne, gdyż decyduje o jakości dalszych etapów.
- Należy mu poświęcić odpowiednio wystarczająco czasu.
- Niektóre mechanizmy → raczej „sztuka” i eksperckie doświadczenie niż rutynowe sformalizowane działania.

I to by było na tyle …

Nie zadawalaj się tym
co usłyszałeś – poszukuj więcej!
Czytaj książki oraz samodzielnie
eksploruj dane!