# „Data Preprocessing"

## Plan wykładu

- Motywacje dla integracji i wstępnego przetwarzania danych:
  - Mieisce w procesie odkrywania wiedzy.
  - Perspektywa projektowania hurtowni danych.
- Ekstrakcja oraz integracja danych pochodzących z różnych źródeł.
- Czyszczenie danych:
  - Uwzględnianie nieznanych wartości.
  - Identyfikacja obserwacji nietypowych.
  - Wpływ szumu informacyjnego.
  - Duplikacja informacji.
- Transformacja danych:
  - Skalowanie i normalizacja dziedzin atrybutów.
  - Agregacja obserwacji.
  - Dyskretyzacja.
- Redukcja (selekcja) danych.

## Miejsce etapów integracji i przetwarzania wstępnego danych w procesie odkrywania wiedzy

- **Definicja**: „Odkrywanie wiedzy w bazach danych to nietrywialny proces poszukiwania wiarygodnych, nowych, potencjalnie użytecznych i zrozumiałych wzorców z danych".
- ***Data mining*** (eksploracja danych) to jeden z etapów procesu odkrywania wiedzy.

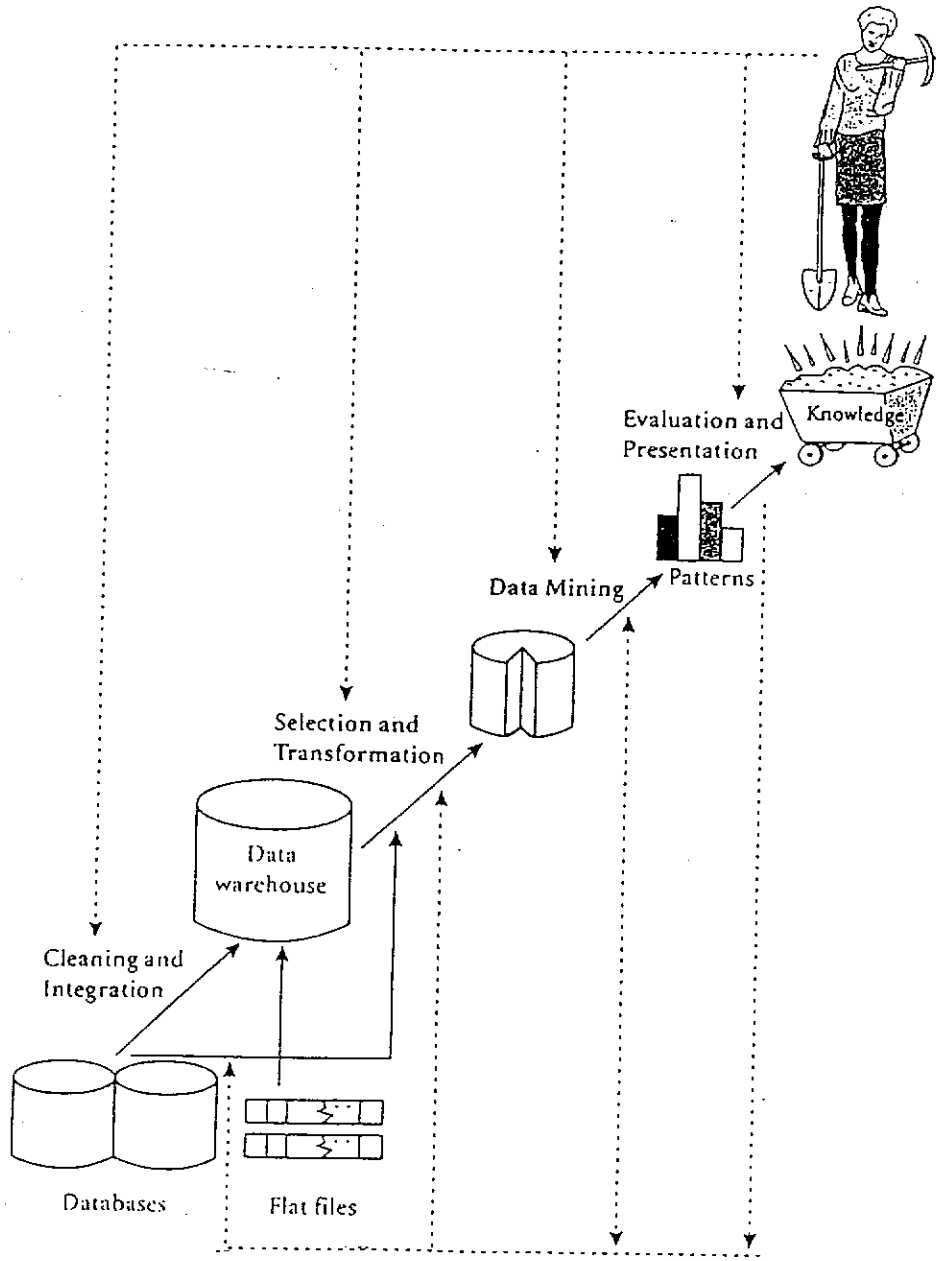
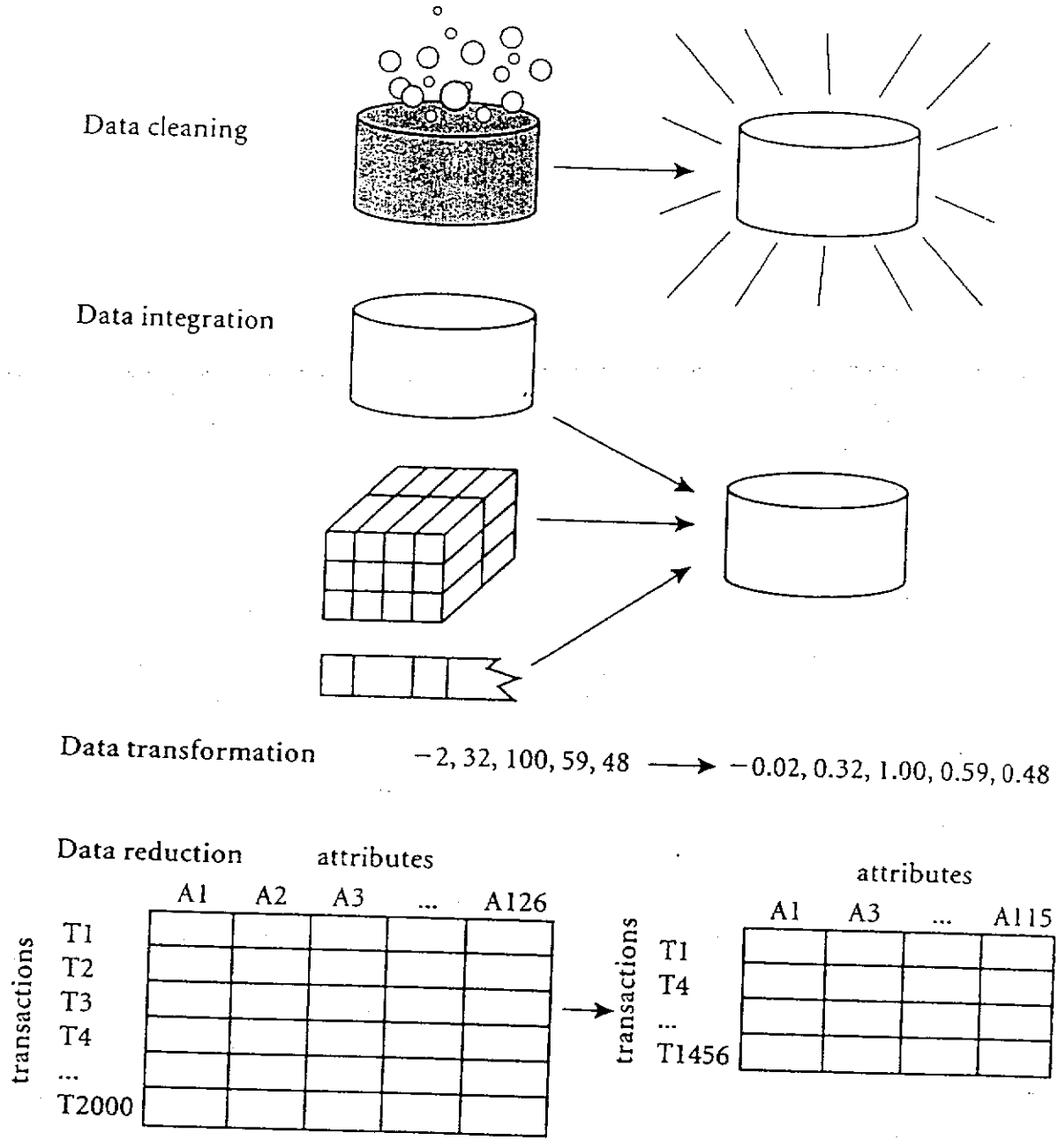


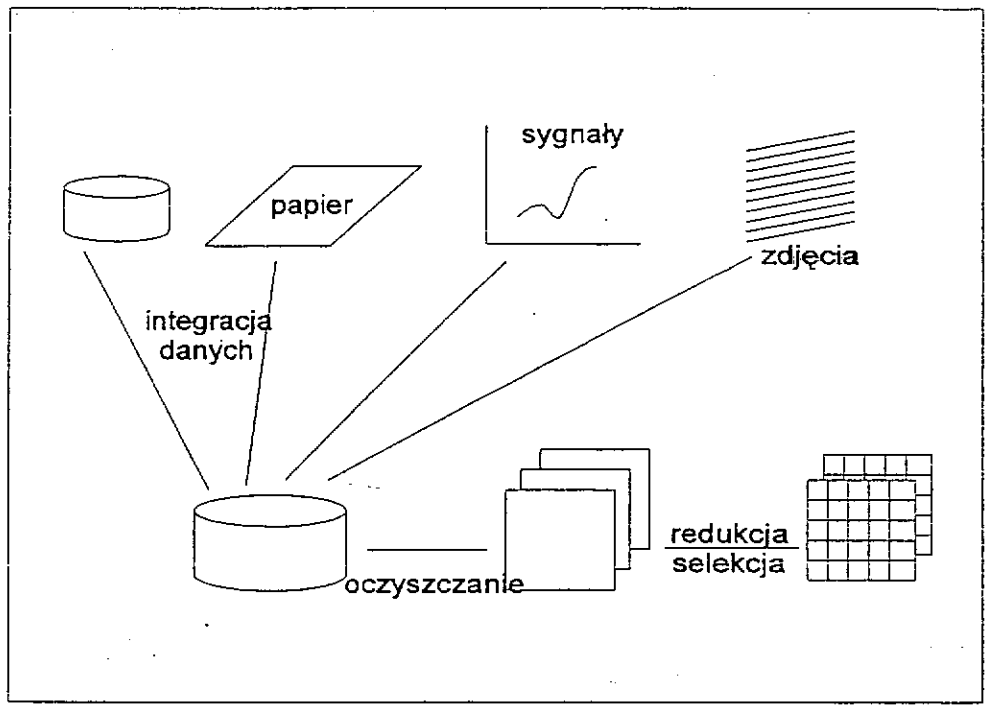

Figure 1.4 Data mining as a step in the process of knowledge discovery.




igure 3.1 Forms of data preprocessing.

# *Motywacje*

- Rzeczywiste dane mogą być niespójne, niekompletne i obdarzone różnego rodzaju zaburzeniami. Ponadto mogą występować różne trudności związane z ekstrakcją oraz integracją danych pochodzących z różnych źródeł.
- Różne rodzaje danych, które podlegają integracji oraz wstępnemu przetwarzaniu:




  - obrazy,
  - sygnały (przebiegi czasowe, sygnały audio, mowa ludzka, sygnały z urządzeń pomiarowych,…)
  - dane rozproszone (ang. spatial/temporal data)
  - surowe obserwacje (pliku w różnych formatach, bazy danych,…)
  - dane tekstowe lub w pewnych językach (HTML, XML,…)
  - także dokumentacja papierowa….

# Motywacje

- Proces przygotowania danych do eksploracji obejmuje ekstrakcje danych z wielu (różnorodnych) źródeł, „czyszczenie", transformacje danych do wspólnych formatów i zapisanie ich do odpowiednich struktur danych (hurtowni, plików, ...)
- Rzeczywiste dane są bardzo „brudne" (ang. dirty).
- Istnieje konieczność:
  - ujednolicenia danych,
  - weryfikacji danych,
  - filtrowania i redukcji 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

## *Integracja – metodyki i narzędzia*

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.**
- **Konflikty strukturalne.**

Więcej o projektach metodyk i narzędziach w: Jarke i in., Hurtownie danych. Podstawy organizacji i funkcjonowania (pol. tłumacz 2003, WSIP)

# *Integracja danych*

- *Integracja danych* – łączenie danych z różnych źródeł w jednolitą, spójną strukturę danych.

  Podstawowe zagadnienia:

  - Integracja schematów,
  - Integracja danych wirtualnych,
  - Integracja danych zmaterializowanych.

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

  Zwróć uwagę na dopasowanie obiektów:

  - Metoda dopasowania za pomocą klucza.
  - Tablice wyszukiwania i funkcje identyczności.

- Potrzeba specjalistycznego oprogramowania do automatyzacji integracji.

# *Integracja formatów*

- Problem przywrócenia integralności dziedzin atrybutów.

  Przykłady:

  - Numery kont bankowych lub numery telefonów mogą mieć w jednym systemie typ „String", a w innych typ „Numeric".
  - Płeć zapisywana na różne sposoby (pełna nazwa, skróty, kody,...)
  - Daty reprezentowane w różnych formatach („ddmmyy", „yymmdd", „yyyy-mm-dd",...)
  - Pola przechowujące walutę.
  - Różne systemy używają różnych rozmiarów pól wartości tekstowych.
  - Pola tekstowe ukrywają ważne dodatkowe informacje.

# Nieznane wartości atrybutów

Różna semantyka nieznanych wartości atrybutów (ang. *unknown attribute values*):

- brakujące wartości atrybutów (ang. *missing values*),

- niedostępne wartości atrybutów (ang. *absent values*).

***Sposoby uwzględniania brakujących wartości:***

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

Przykład niekompletnej tablicy decyzyjnej.

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $D$ |
|---|---|---|---|---|---|
| $x_1$ | 3 | 2 | 1 | 0 | Φ |
| $x_2$ | 2 | 3 | 2 | 0 | Φ |
| $x_3$ | * | 2 | * | 1 | Ψ |
| $x_4$ | 2 | 3 | 2 | 1 | Ψ |
| $x_5$ | 3 | * | * | 3 | Φ |
| $x_6$ | * | 0 | 0 | * | Ψ |
| $x_7$ | 3 | 2 | 1 | 3 | Ψ |
| $x_8$ | 1 | * | * | * | Φ |
| $x_9$ | * | 2 | * | * | Ψ |
| $x_{10}$ | 3 | 2 | 1 | * | Φ |

* ? nieznana wartość atrybutu $a_i$

## Uwzględnianie brakujących wartości atrybutów

*Podstawowe sposoby dla przetwarzania wstępnego*:

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).

# Noisy data

- ***Szum informacyjny*** (ang. *noise*) – błąd przypadkowy lub zmienność mierzonej wielkości (atrybutu).
- Przyczyny (?)
  - błędy urządzeń pobierających dane,
  - błędy transmisji,
  - błędy ludzkie,
  - ograniczenia technologiczne,
  - niespójności i niekonsekwencje nazewnictwa.

  Konsekwencje danych niekompletnych i sprzecznych.
- Różne metody:
  1. Wygładzenie danych (ang. *smoothing techniques*).
  2. Tworzenie przedziałów (ang. *binning*).
  3. Algorytmy skupień i obiekty reprezentacji skupień
  4. Wykorzystywanie modeli predykcji (regresja)
  5. Konsultacja z użytkownikiem / ekspertem

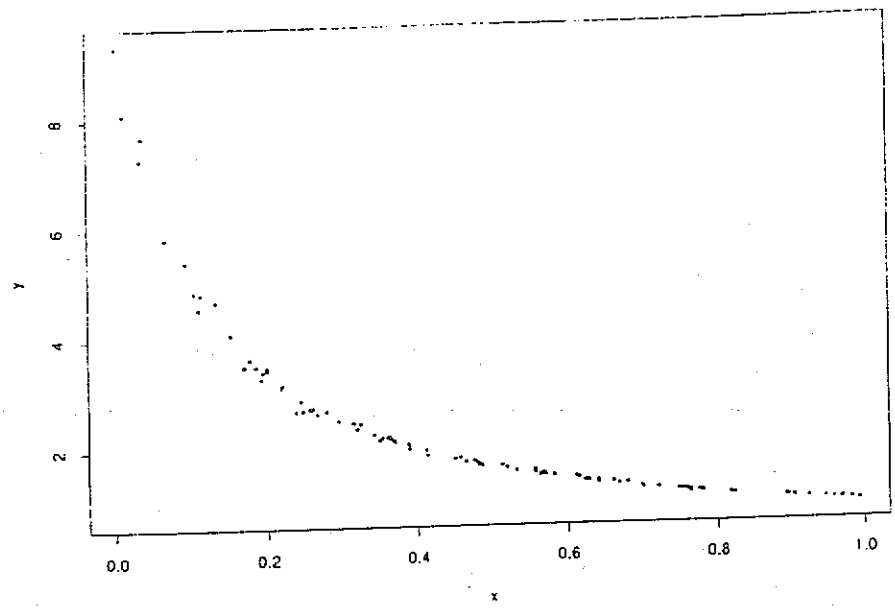



Figure 2.2 A simple nonlinear relationship between variable $V_1$ and $V_2$. (In these and subsequent figures $V_1$ and $V_2$ are on the $X$ and $Y$ axes respectively).

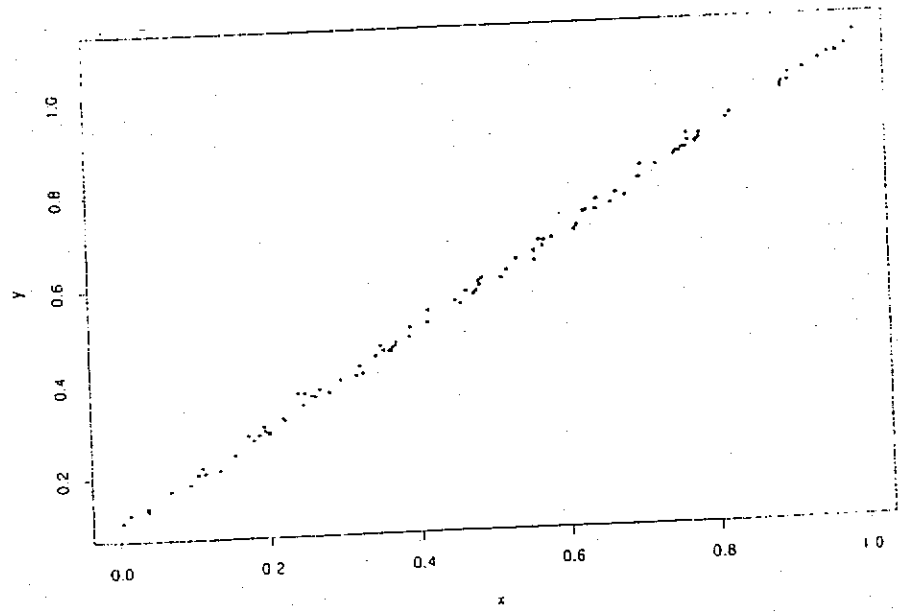


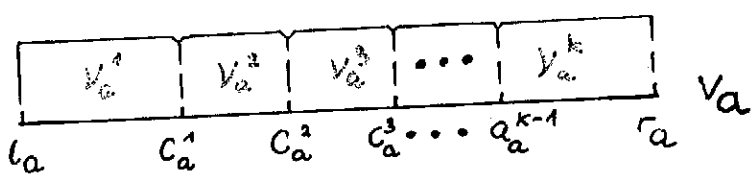

Figure 2.3 The data of figure 2.2 after the simple transformation of $V_2$ to $1/V_2$.

## Dyskretyzacja

- Różne typy atrybutów (jakościowe vs. liczbowe)

**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.

- Niektóre algorytmy uczące przetwarzają atrybuty jakościowe; Dyskretyzacja stosowana w etapie przetwarzania wstępnego przed użyciem algorytmów.
- Redukcja danych w rezultacie dyskretyzacji.




Ogólny podział algorytmów dyskretyzacji:

1. Nadzorowane i nienadzorowane.
2. Globalne i lokalne.
3. Dynamiczne i statyczne.

## Klasyfikacja metod dyskretyzacji

Możliwości podziału wielu metod:

- nienadzorowane i nadzorowane,
- lokalne i globalne,
- statyczne i dynamiczne.

Dyskretyzacja stosowana w przetwarzaniu wstępnym lub lokalnie w algorytmie indukcji.

Metody **nienadzorowane** (ang. *unsupervised*) nie wykorzystują informacji o przydziale obiektów do klas decyzyjnych; metody, w których ta informacja jest wykorzystywana, nazywane są **nadzorowanymi** (ang. *supervised*).

Metody **dynamiczne** przetwarzają informację o wszystkich atrybutach liczbowych równocześnie i w trakcie obliczeń starają się znaleźć punkty graniczne równocześnie dla wielu atrybutów. Metody **statyczne** – pojedyncze atrybuty i zadana liczba punktów granicznych.

Dotychczas zaproponowano wiele metod dyskretyzacji.

## Podstawowe algorytmy dyskretyzacji

- Podział równymi przedziałami (*equal-width interval*)

  Podziel zakres przedziału atrybutu na $N$ podprzedziałów równej długości.

- Podział przedziałami o równej czestości (*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).

- Inne, np. oparte na funkcjach rozróżnialności.

## Dyskretyzacja wykorzystująca entropię warunkową

Dla podzbioru przykładów $x \in S$ ich wartości atrybutu $a$ są sortowane według wzrastającego porządku.

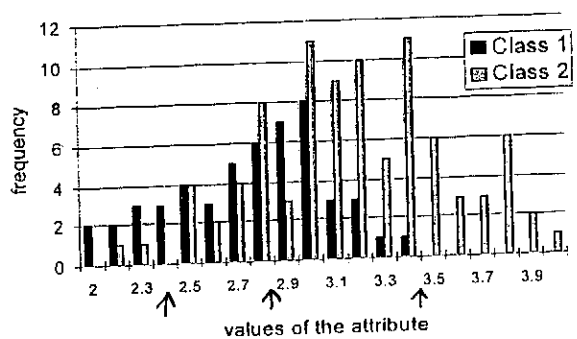
Dla przykładów $S$ wybierz punkt graniczny (ang. *cut point*) $c^a$ dokonując podziału $S = S_1 \cup S_2$.

Dla zbioru $S$ entropia jest zdefiniowana jako:

$Ent(S) = -\sum_{i=1}^{r} p_i \cdot \lg_2 p_i$ (gdzie $r$ liczba klas decyzyjnych).

Entropia warunkowa $Ent(S,c^a)$ dla podziału $S_1 \cup S_2$ wynikającego z punktu granicznego $c^a$:

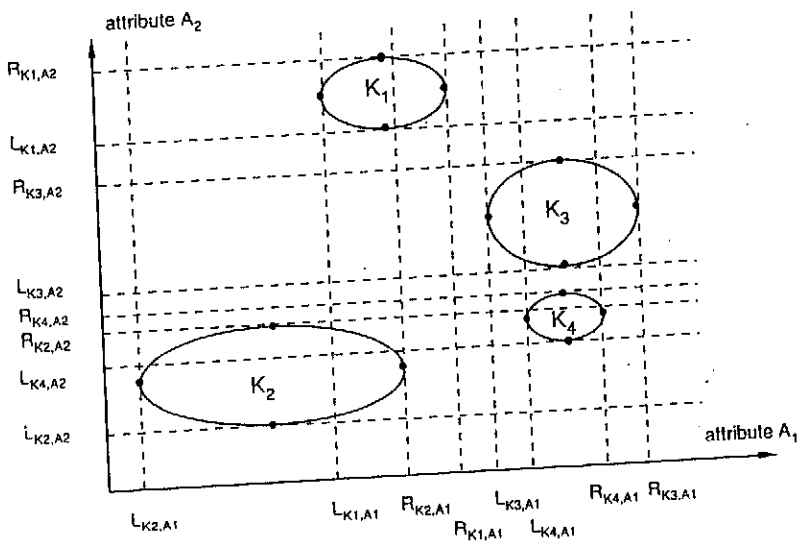
$$\frac{|S_1|}{|S|} \cdot Ent(S_1) + \frac{|S_2|}{|S|} \cdot Ent(S_2)$$




Poszukuje się takiego punktu granicznego $c^a$, dla którego wartość entropii warunkowej jest minimalna.

Właściwość Fayyad'a i Iraniego redukuje liczbę punktów kandydujących.

Proces binarnej dyskretyzacji można kontynuować do warunku zatrzymania, np. $Ent(S) - Ent(S,c^a) > \delta$.




**Cluster definability in terms of attributes $A_1$ and $A_2$**
**Single Local Covering Option (LEM2) of LERS**

Table C3.4.1. Data with numerical attributes

| IQ | Weight | Height | Class |
|---|---|---|---|
| 109 | 63 | 175 | no |
| 105 | 90 | 170 | yes |
| 115 | 61 | 178 | yes |
| 107 | 85 | 182 | no |
| 107 | 62 | 179 | no |
| 113 | 92 | 172 | yes |

(IQ, 105) -> (Class, yes)
(IQ, 107) -> (Class, no)
(IQ, 109) -> (Class, no)
(IQ, 113) -> (Class, yes)
(IQ, 115) -> (Class, yes)

zbiór reguł wygenerowany algorytmem LEM 2 bezpośrednio ze zbioru przykładów uczących.

Table C3.4.3. Data discretized by Minimal Entropy Method

| IQ | Weight | Height | Class |
|---|---|---|---|
| 105..113 | 61..90 | 175..182 | no |
| 105..113 | 90..92 | 170..175 | yes |
| 113..115 | 61..90 | 175..182 | yes |
| 105..113 | 61..90 | 175..182 | no |
| 105..113 | 61..90 | 175..182 | no |
| 113..115 | 90..92 | 170..175 | yes |

An example of the rule set induced from Table C3.4.3 is:

(IQ, 105..113) & (Weight, 61..90) -> (Class, no)
(Weight, 90..92) -> (Class, yes)
(IQ, 113..115) -> (Class, no)

## Dyskretyzacja wykorzystująca entropię warunkową

Przykład obliczeń

$$Ent(S) = -\frac{3}{6}\cdot \lg\frac{3}{6} - \frac{3}{6}\cdot \lg\frac{3}{6} = 1$$

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

Podobnie dla *Weight* najlepszy T=90 i dla *Height* T=175.

## Przykład zastosowania różnych algorytmów dyskretyzacji

***Equal Interval Frequency***

Atrybuty *IQ*, *Weight*, *Height* są zdyskretyzowane binarnie wykorzystując odpowiednio punkty graniczne: 109, 85 i 178.

| IQ | Weight | Height | Class |
|---|---|---|---|
| 109..115 | 61..85 | 170..178 | no |
| 105..109 | 85..92 | 170..178 | yes |
| 109..115 | 61..85 | 178..182 | yes |
| 105..109 | 85..92 | 178..182 | no |
| 105..109 | 61..85 | 178..182 | no |
| 109..115 | 85..92 | 170..178 | yes |

Przykład wygenerowanych reguł (algorytm LEM2):

(Weight, 61..85) & (Height,170..178) → (Class,no)

(IQ, 105..109) & (Height,178..182) → (Class,no)

(Weight, 85..92) & (Height,170..178) → (Class,yes)

(IQ, 109..115) & (Height,178..182) → (Class,yes)

# Redukcja liczności pomiarów:

**Metody parametryczne**:

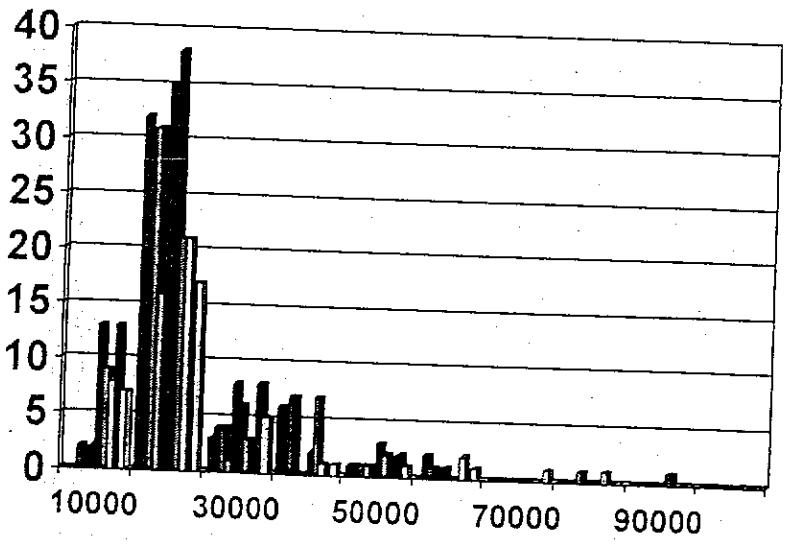
Załóż, że dane pasują do pewnego modelu formalnego; estymuj parametry tego modelu; zapamiętaj parametry tego modelu i pomiń obserwacje elementarne (za wyjątkiem obserwacji „odstających"). Przykład: regresja wielowymiarowa i modele logistyczno-liniowe.

**Metody nieparametryczne**:

Zapamiętują zredukowaną reprezentację bez zakładania parametryzacji modelu. Przykłady:

1. Agregacja histogramów,
2. Grupowanie – analiza skupień,
3. Próbkowanie.

**Agregacja histogramów**:




Podziel uporządkowany szereg obserwacji na przedziały i zapamiętaj miarę reprezentującą obserwacje wewnątrz przedziału.

Sposoby tworzenia przedziałów: *równej długości, równej częstości, V-optimal, Max-Diff.*

# Redukcja liczby przykładów

**Pytania** [*Weiss, Indurkhia, Predictive data mining*]:

1. How many cases?
2. Are all cases residing in a database needed for effective mining?

**Różne punkty widzenia**:

- Jeśli dane są silnie „zaszumione" (ang. noise), to wiarygodne odkrycie pojęć nie zależy od większej liczby przypadków.
- Uczenie się z większej liczby przypadków może być zależne od specyfiki zastosowania oraz charakteru i złożoności pojęć, które zamierza się odkryć (nauczyć się):
- Proste pojęcia nie wymagają dużej ilości dodatkowych przykładów,
- Złożone pojęcia – więcej przykładów:
    1. Wieloklasowe problemy uczenia się.
    2. Regresja.
    3. Klasyfikacja z niezrównoważonymi klasami decyzyjnymi.
- Przetarg pomiędzy spodziewanym oszacowaniem błędu (dokładności) a złożonością procesu uczenia się.

**Co robić ?**

Właściwe losowanie prób (ang. *random sampling*).

# Tworzenie prób losowych

Właściwe losowanie reprezentatywnej prób (ang. *random sampling*).

*Pojęcia statystyczne*:

- **Zbiorowość generalna (populacja)** - zbiór elementów obejmujący wszystkie obiekty.
- **Próba** z populacji - podzbiór zbiorowości generalnej, obejmujący część jej elementów wybranych w określony sposób.
- Znaczenia terminu ***reprezentatywności***:
  - próba jest reprezentatywna, kiedy występują w niej wszystkie wartości zmiennej czy zmiennych nas interesujących,
  - próba jest reprezentatywna, kiedy rozkłady interesujących zmiennych odpowiadają rozkładom tych zmiennych w populacji,
  - próba jest reprezentatywna. kiedy zależności między zmiennymi odpowiadają analogicznym zależnościom tych zmiennych w populacji.

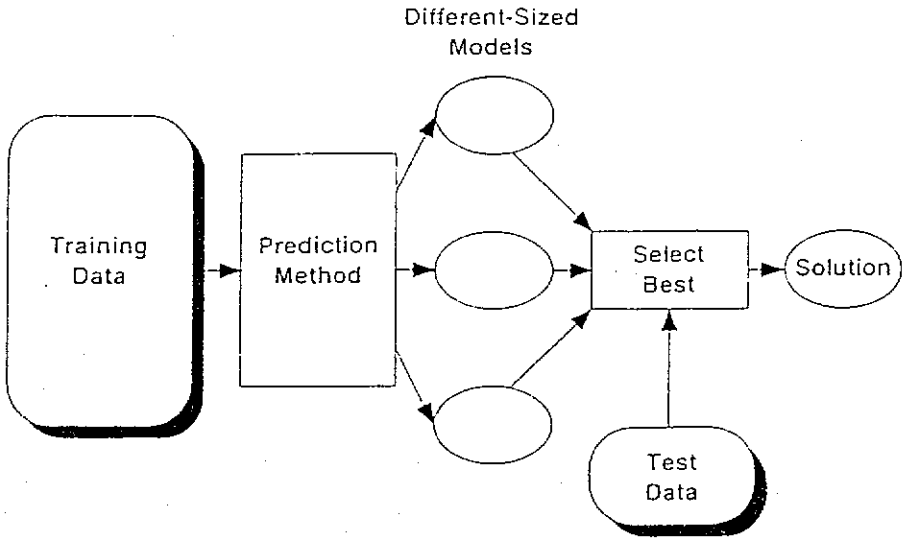



Figure 2.4: Fitting the Right-Size Model to Data

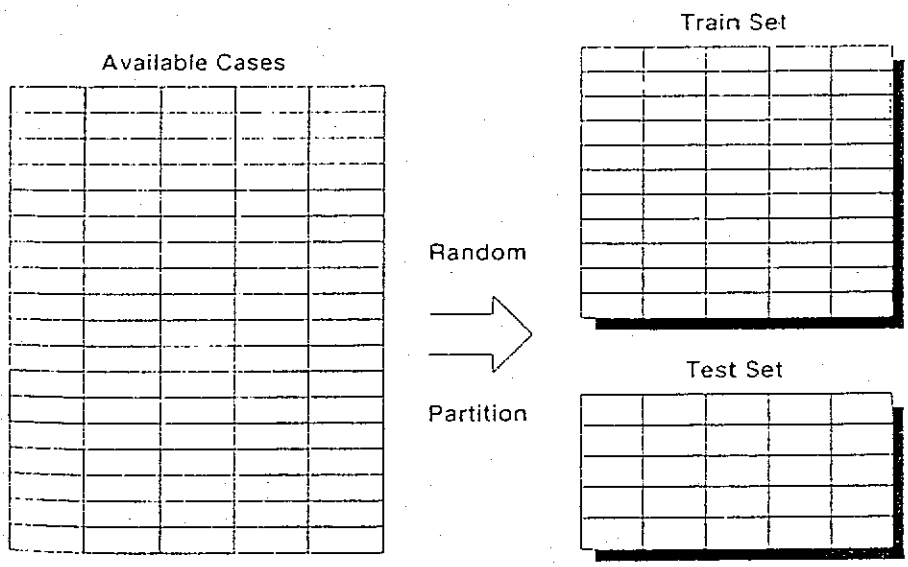



Figure 2.5: Random Train/Test Model

# Techniki losowania

$D$ – oryginalne dane ($N$ przykładów)

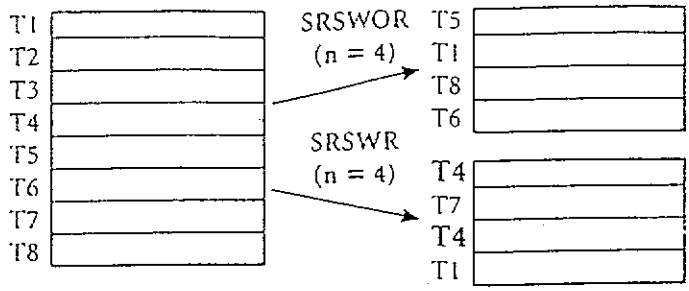
Próba – $n << N$ wylosowanych przykładów

***Podstawowe sposoby losowania***:

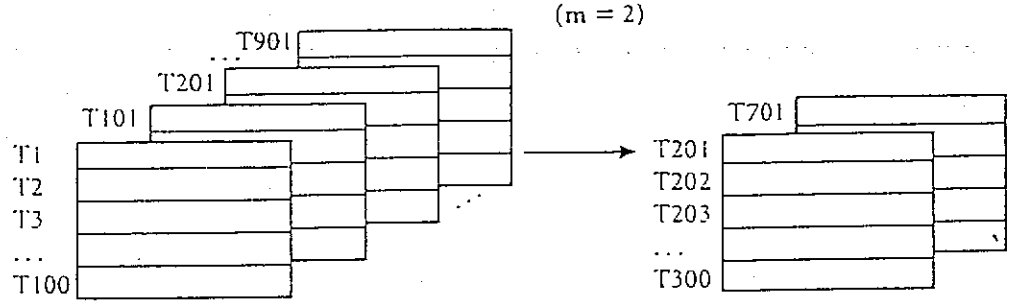
- Proste (jednokrotne) losowanie bez powtórzeń (ang. SRSWOR - *Simple Random Sampling WithOut Replacement*).
- Proste losowanie z powtórzeniami (SRSWR) – przykład po wylosowaniu powraca do populacji.
- Losowanie wielu podzbiorów (prób) – Dane $D$ podzielone na $M$ podzbiorów (rozłączne vs. nierozłącznie).
- Losowanie warstwowe (ang. *stratified*).

***Ponadto,***

- Techniki podziału (***train and test / holdout***):
  1. Losowy podział na zbiór przykładów uczących i zbiór przykładów testowych.
  2. Specyficzne techniki - „*k-cross- validation*", *leaving-one-out*, *bootstraping*.
- Uczenie przyrostowe (ang. ***incremental learning***)




Cluster sample
(m = 2)




Stratified sample
(according to *age*)

| | |
|---|---|
| T38 | young |
| T256 | young |
| T307 | young |
| T391 | young |
| T96 | middle-aged |
| T117 | middle-aged |
| T138 | middle-aged |
| T263 | middle-aged |
| T290 | middle-aged |
| T308 | middle-aged |
| T326 | middle-aged |
| T387 | middle-aged |
| T69 | senior |
| T284 | senior |

| | |
|---|---|
| T38 | young |
| T391 | young |
| T117 | middle-aged |
| T138 | middle-aged |
| T290 | middle-aged |
| T326 | middle-aged |
| T69 | senior |

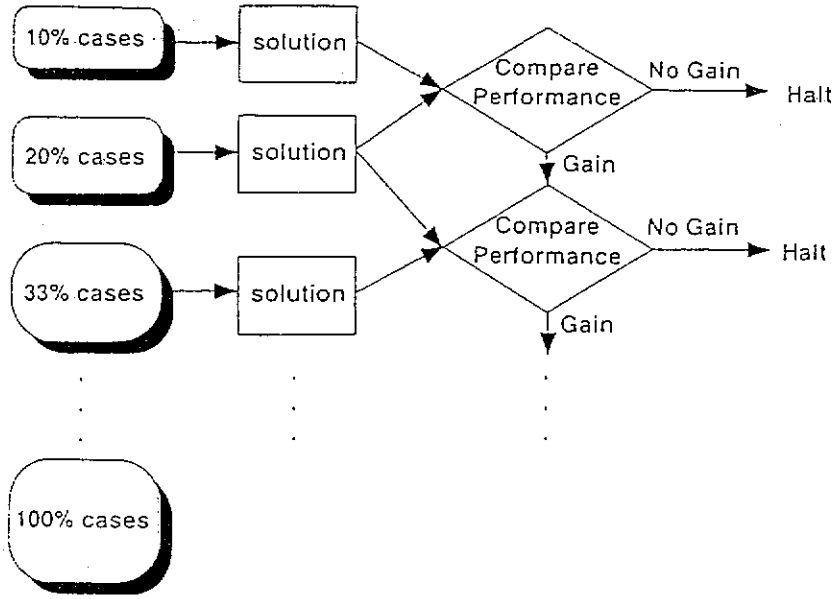



Figure 4.13: Incremental Sampling and Mining

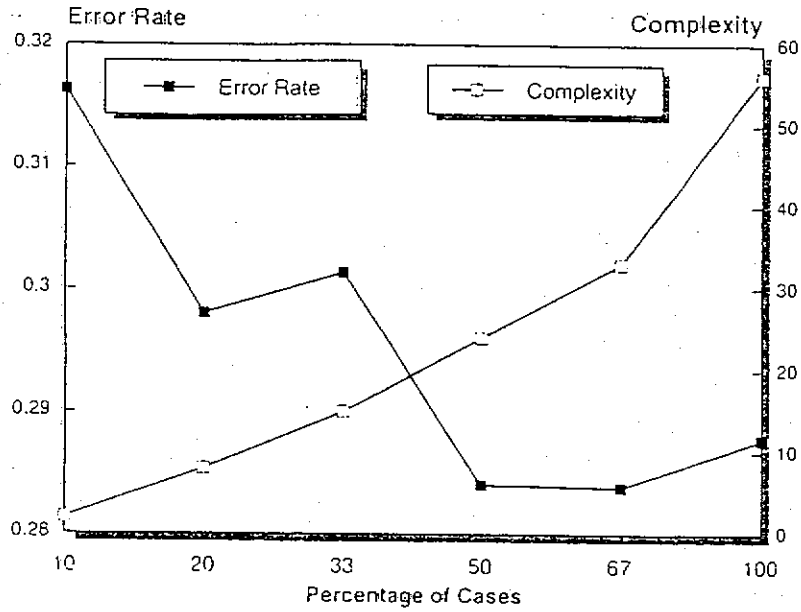



Figure 4.12: A Sample Trend for Error and Complexity

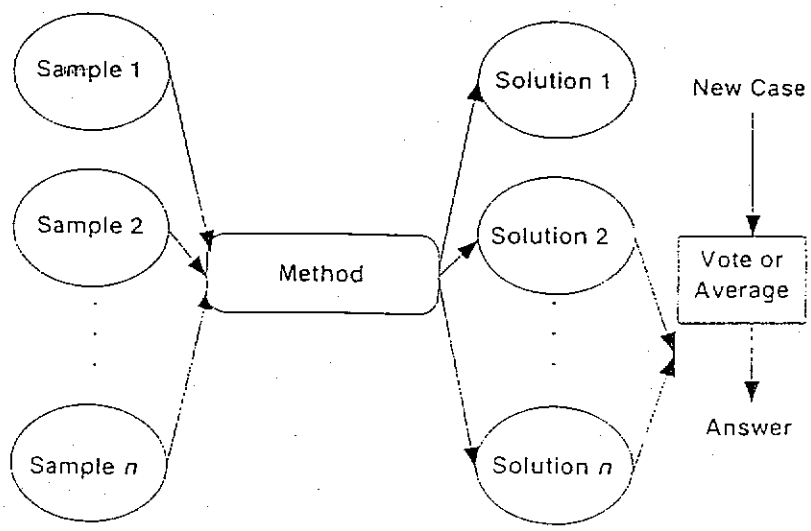



Figure 4.14: Combining Solutions from Different Samples

# Selekcja atrybutów

Dany jest $n$ elementowy zbiór przykładów (obiektów). Każdy przykład $x$ jest zdefiniowany na $V_1 \times V_2 \times \ldots \times V_m$ gdzie $V_i$ jest dziedziną $i$-tego atrybutu. W przypadku uczenia nadzorowanego przykłady są zdefiniowane jako $<x,y>$ gdzie $y$ określa pożądaną odpowiedź, np. klasyfikację przykładu.

Cel selekcji atrybutów:

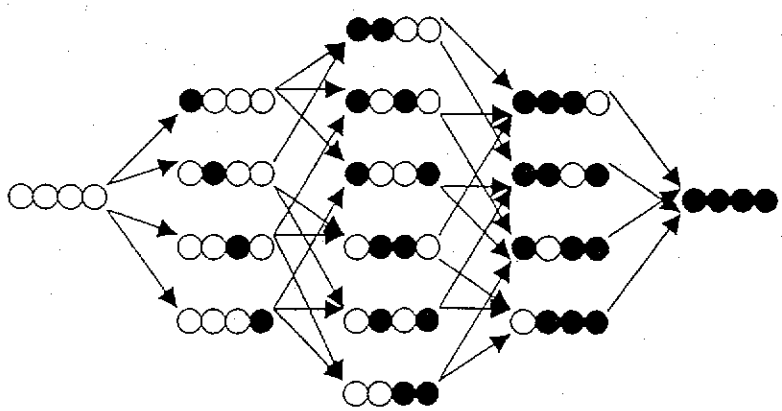
*Wybierz minimalny podzbiór atrybutów, dla którego rozkład prawdopodobieństwa różnych klas obiektów jest jak najbliższy oryginalnemu rozkładowi uzyskanemu z wykorzystaniem wszystkich atrybutów.*

W przypadku uczenia nadzorowanego (klasyfikowania):

*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.*

Selekcja atrybutów to problem przeszukiwania

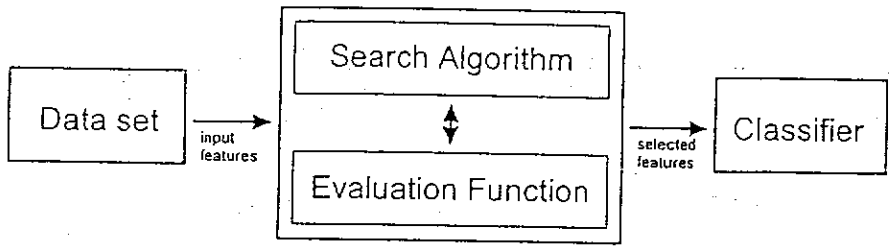
Każdy stan reprezentuje podzbiór atrybutów ($2^m$ możliwych stanów!).



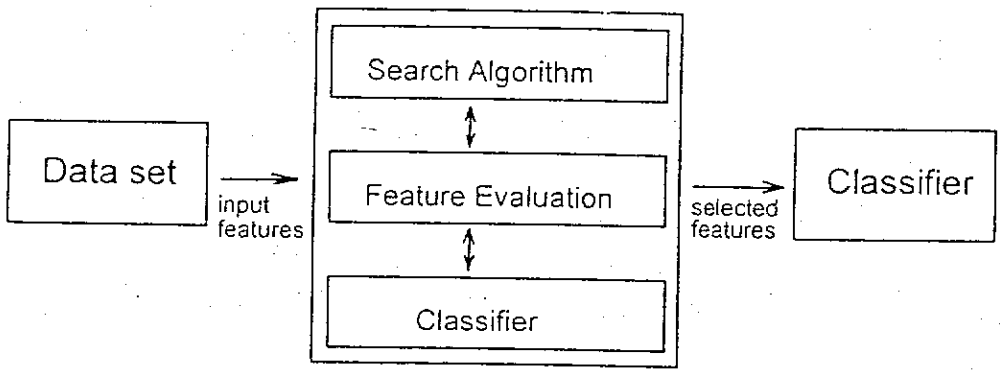
Trzy elementy selekcji atrybutów:

- *algorytm przeszukiwania* – przegląda przestrzeń podzbiorów,
- *miara oceny* – podzbioru atrybutów,
- *klasyfikator* –stworzony algorytmem uczenia się z wybranego podzbioru cech.

Dwa sposoby integracji - *filter* i *wrapper* model.




Filter model – atrybuty wybierane niezależnie od klasyfikatora.




"Wrapper model" - klasyfikator użyty jako element funkcji oceny

Uwaga: - zalecany niezależny zbiór przykładów weryfikujących.

# Selekcja atrybutów

Selekcja atrybutów dla problemów analizy danych ze znaną **klasyfikacją obiektów** (np. wyrażona przy pomocy atrybutu decyzyjnego).

Selekcja w trakcie wstępnego przetwarzania danych.

Model „**filter**" vs. „wrapper"

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):

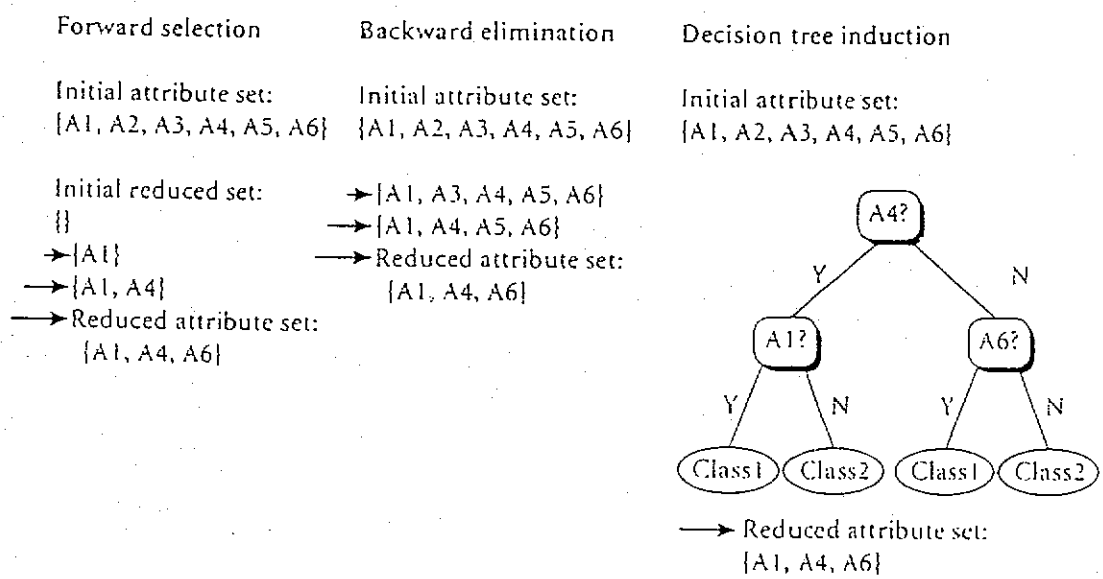
- Miara korelacji wzajemnych,
- Statystyki $\lambda$ Wilksa, $T^2$-Hotellinga, odległości $D^2$ Mahalanobisa,
- Redukty w teorii zbiorów przybliżonych,
- Techniki dekompozycji na podzbiory (ang. *data table templates*)
- ...

# Selekcja atrybutów

Dla $m$ atrybutów liczba podzbiorów $2^m$ –dokładne rozwiązania są kosztowne.

Heurystyczne algorytmy przeszukiwania:

- BSS - *backward stepwise selection/elimination*,
- FSS *forward stepwise selection*,
- specjalizowane, np. Połączenie BSS i FSS. Beam-search SFFS, podejścia "losowania", algorytmy genetyczne...




Inne techniki:

Proste badanie znaczenia pojedynczych atrybutów; dobrze opracowane dla problemów klasyfikacji, np. testy statystyczne jak $\chi^2$, względna entropia, ocena znaczenia atrybutu dla wybranego algorytmu indukcji wiedzy.

Wewnętrzny wybór istotnych (*relevant*) atrybutów dla klasyfikacji w systemach indukcji symbolicznej reprezentacji wiedzy (drzewa decyzyjne lub reguły decyzyjne).

## *Correlation-based merit measure*

Oblicza się korelację między wszystkimi parami dwoma zmiennych (atrybutów) $r$

„Dobroć" zbiory podzbioru atrybutów $F$ w stosunku do atrybutu decyzyjnego $d$:

$$\frac{k\overline{r_{df}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$$

gdzie $F$ zawiera $k$ atrybutów, $\overline{r_{df}}$ średnia korelacja atrybutu $f$ z klasyfikacją (atrybut $d$); $\overline{r_{ff}}$ średnia wzajemna korelacja atrybutów (obliczona z $r$)

**Problem ?**

*Korelacja z atrybutem dyskretnym*

Atrybut dyskretny $y$ ($t$ wartościowy) podlega zamianie na $t$ atrybutów binarnych (1 – wartość atrybutu występuje, 0 w przeciwnym przypadku).

Badamy korelacje atrybutu liczbowego $x$ z $y$ – tj. atrybutu $x$ z $t$ atrybutami binarnymi i obliczamy prawdopodobieństwo, że atrybut $y$ przyjmie wartości przyjmie wartości $y_i$ (gdzie $i$=1,...,$t$):

$$r_{xy} = \sum_{i=1}^{t} p(y = y_i) \cdot r_{xy_{bi}}$$