

---

# Przetwarzanie wstępne danych

## Projekt Eksploracji Danych



JERZY STEFANOWSKI  
Instytut Informatyki  
Politechnika Poznańska

Wykład 3 – transformacje i dyskretyzacja  
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
  - Wybór obiektów
5. Dyskretyzacja atrybutów liczbowych

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

# 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.

# 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                         | primary                               | unsupervised | parametric     | non-hierarchical | univariate   | disjoint     | global | eager | time-insensitive | nominal | non-fuzzy |
| Equal-frequency                     |                                       |              |                |                  |              |              |        |       |                  |         |           |
| Fixed-frequency                     |                                       |              |                |                  |              |              |        |       |                  |         |           |
| Multi-interval-entropy-minimization | primary                               | supervised   | non-parametric | hierarchical     | univariate   | disjoint     | global | eager | time-insensitive | nominal | non-fuzzy |
| ChiMerge                            | primary                               | supervised   | non-parametric | hierarchical     | univariate   | disjoint     | global | eager | time-insensitive | nominal | non-fuzzy |
| StatDisc                            |                                       |              |                |                  |              |              |        |       |                  |         |           |
| InfoMerge                           |                                       |              |                |                  |              |              |        |       |                  |         |           |
| 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.

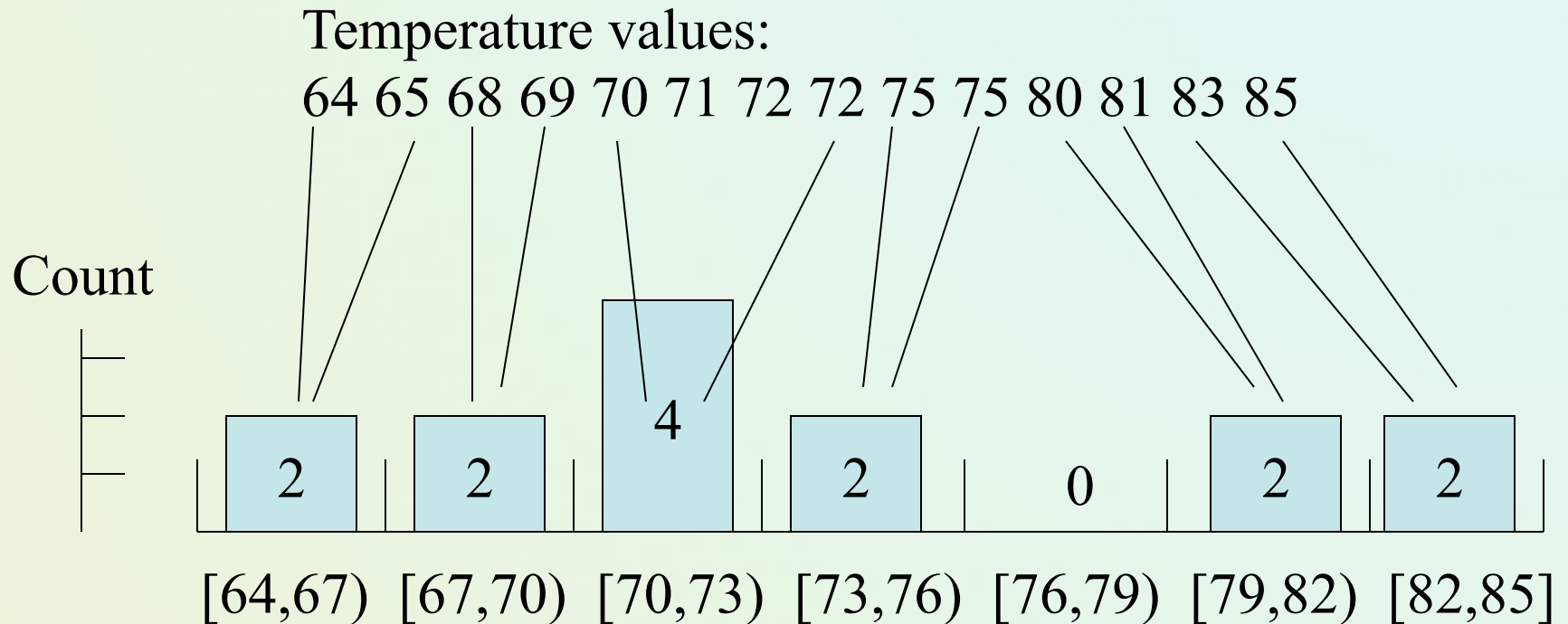
# Przykładowe popularne metody dyskretyzacji

---

- 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)

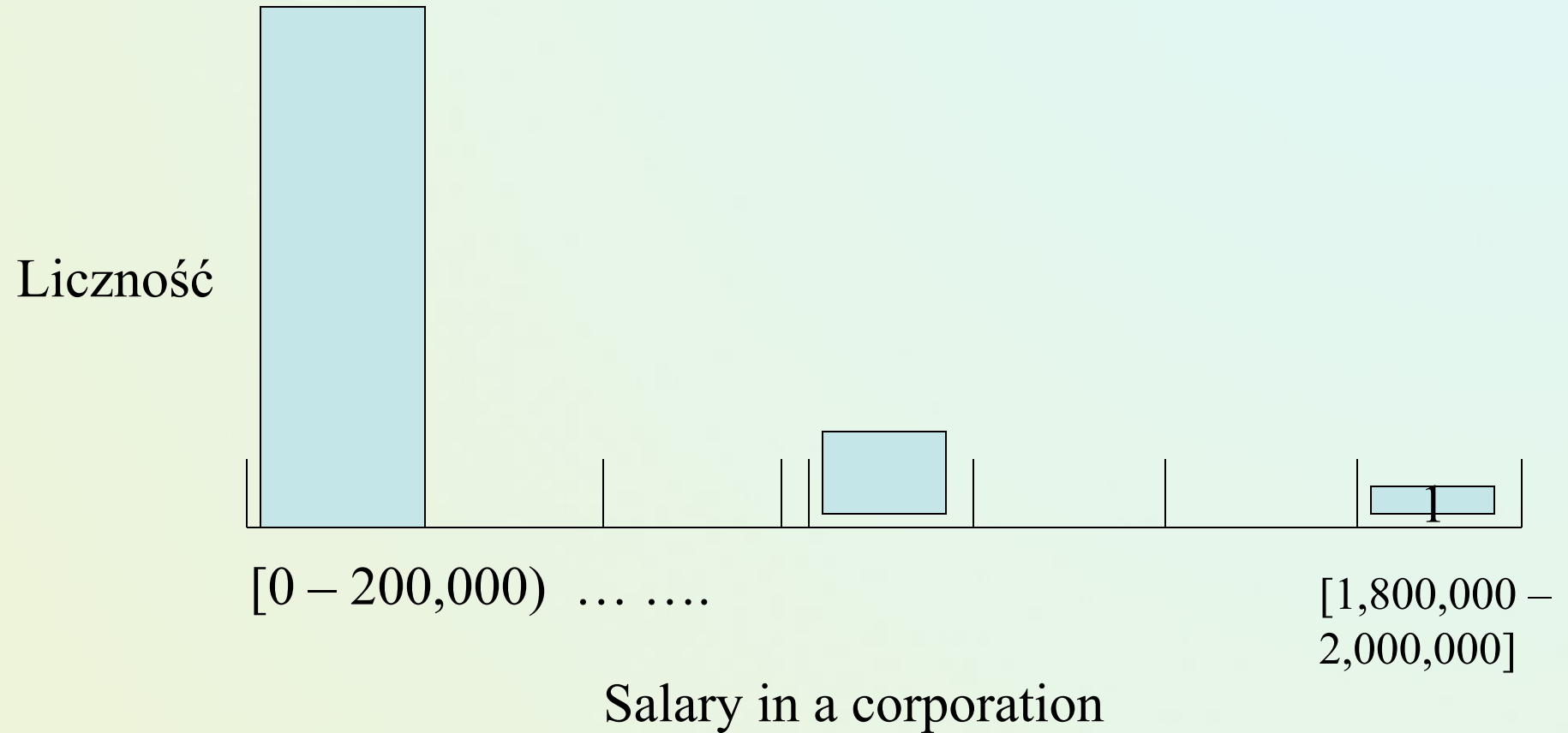
---



Equal Width, bins  $\text{Low} \leq \text{value} < \text{High}$

# Discretization: Equal-Width / niewłaściwe bardzo skośne rozkłady

---



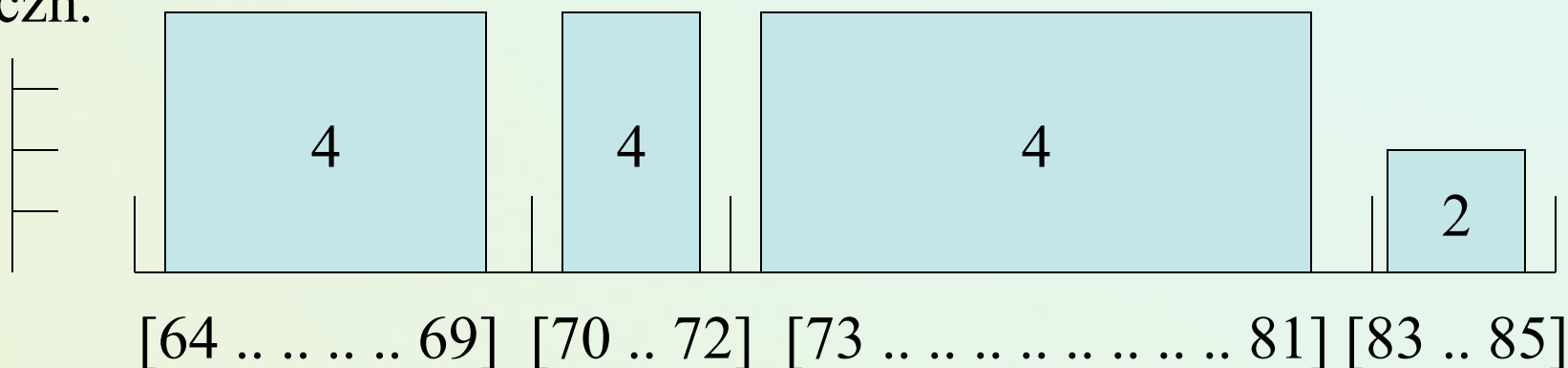
# Discretization: Equal-Frequency

---

Temperatura (F):

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

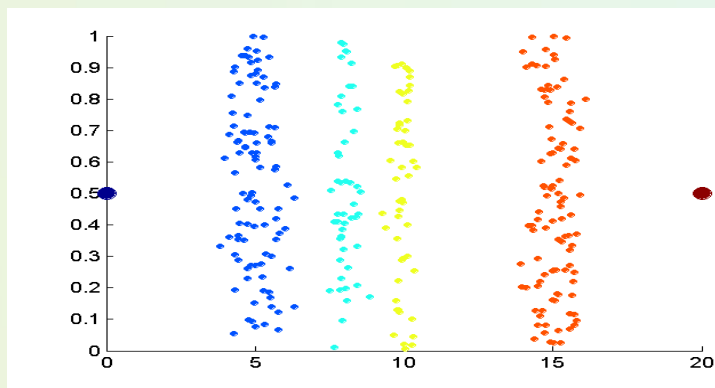
Liczn.



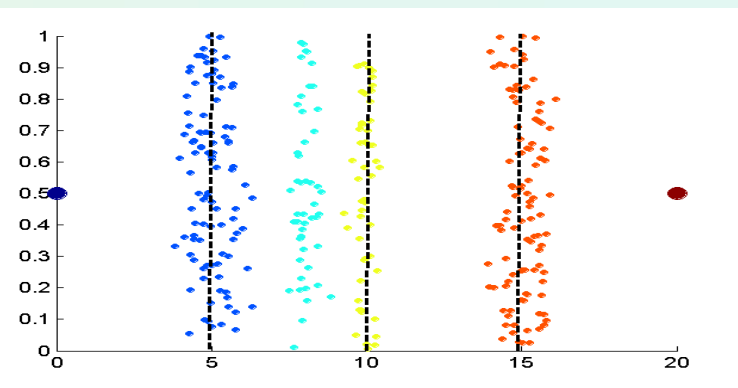
W przybliżeniu 4 na przedział

# Dyskretyzacja nienadzorowana

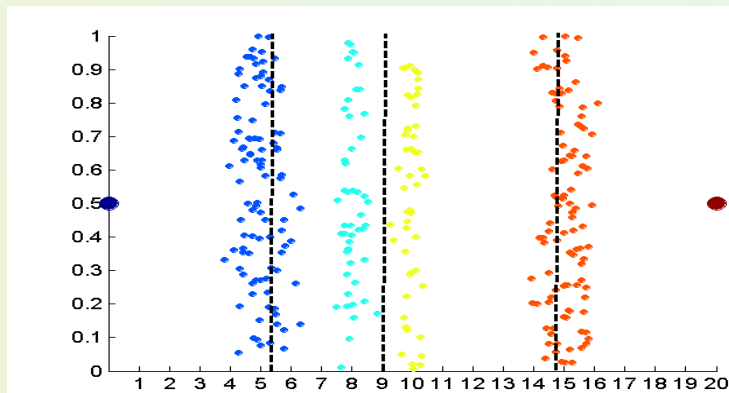
## Różne podejścia do specyficznego rozkładu danych



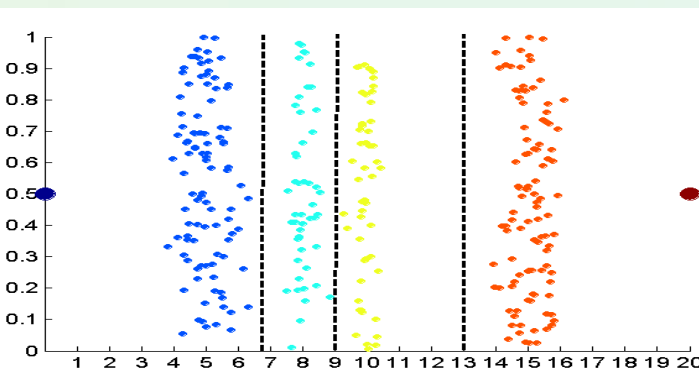
Data



Equal interval width



Equal frequency

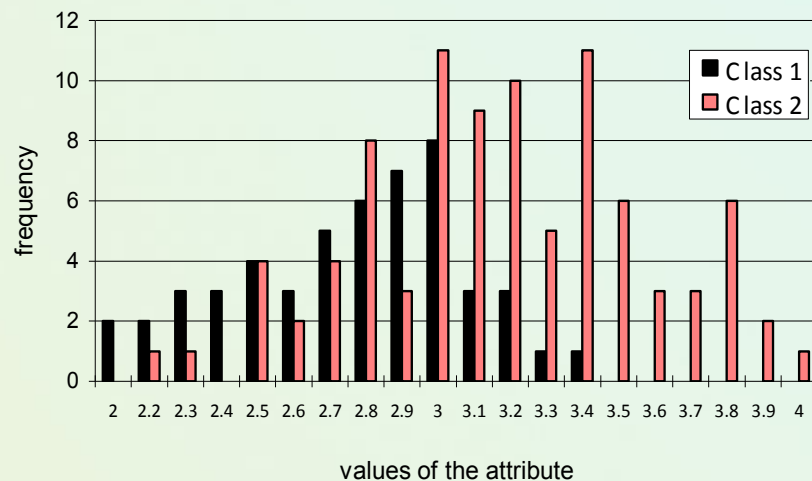


K-means

# Supervised (class) discretization – nadzorowana dyskretyzacja

---

- Użyj dodatkowo informacji o etykietyzacji (class labels) przykładów.



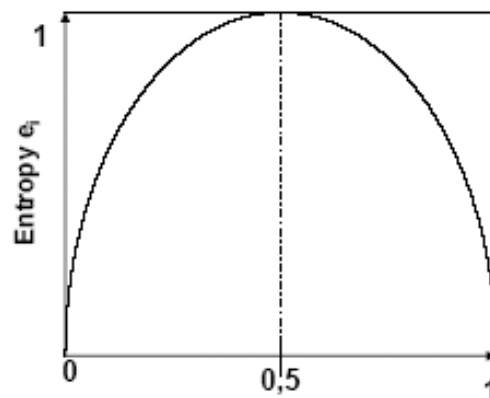
- Minimal entropy based approaches; Chi-Merge, others

# 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?
- Informacja dodatkowa  $\rightarrow$  entropia warunkowa

# Entropy-Based Discretization

---

- Dany jest zbiór przykładów  $S$ ; Jeśli  $S$  podzielimy na 2 podprzedziały  $S_1$  i  $S_2$  używając punktu (cut point)  $T$ , entropia warunkowa:

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

- Przeanalizuj wszystkie punkty cięcia
- Wybierz ten, który minimalizuje entropię.
- Proces można kontynuować aż do warunku zapytania

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

- MDL principle jako podstawa warunku zatrzymania

# 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  $T=107$  (intr. Lewy  $< T$ )

|     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|
| 105 | 107 | 107 | 109 | 113 | 115 |
| yes | no  | no  | no  | yes | yes |

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

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

# Dyskretyzacja lokalna vs. wiele atrybutów

---

- Nadzorowana dyskretyzacja z miarą entropii
  - W zasadzie lokalna, można sterować liczbą przedziałów
- Wersja globalizowana?
  - Przeanalizuj wszystkie atrybuty i oceń je (np. gain entropii).
  - Wybierz jeden z atrybutów.
  - Wykonaj dyskretyzację binarną i podziel tabelę danych na dwie podtabelę
  - Powtórz procedurę w każdej z podtabeli

# Porównanie różnych metod [Grzymała]

- Ocena eksperymentalna z LEM2

330

M. R. Chmielewski and J. W. Grzymała-Busse

**Table 2.** Accuracy Rate after Discretization

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



# 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 zadawajaj się tym  
co usłyszałeś – poszukuj więcej!  
Czytaj książki oraz samodzielnie  
eksploruj dane!

