Pengantar Data Mining #4: Praproses Data [2]

Isnan Mulia, S.Komp, M.Kom

Data Reduction

Reduksi jumlah data

- Mengganti data asli dengan representasi data yang berukuran lebih kecil
- Bisa berupa teknik parametrik atau nonparametrik

Kompresi data

- Data dipadatkan/dikompres
- Bisa bersifat *lossless* atau *lossy*

Data Reduction – Reduksi Jumlah Data

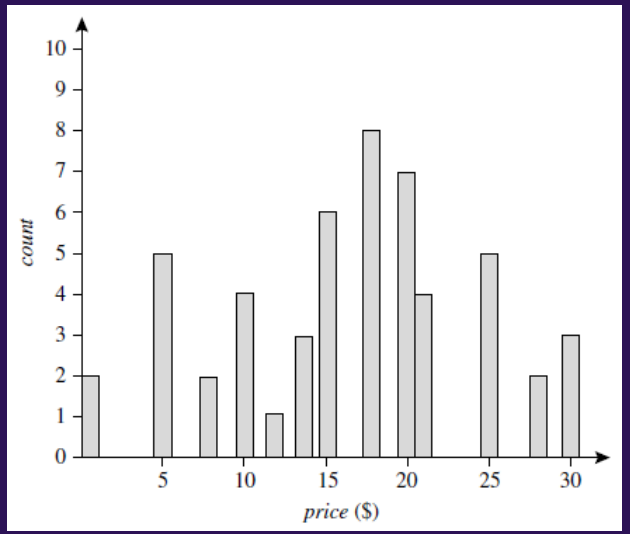
Histogram:

- Menggunakan *binning* untuk memperkirakan distribusi data
- Data dibagi menjadi *n* buah ember/*bin* menggunakan ketentuan:
  - ✓ *Equal-width*: lebar/rentang setiap ember/*bin* seragam
  - ✓ *Equal-frequency*/ *equal-depth*: frekuensi data dari setiap ember/*bin* sama
- Efektif dalam memperkirakan data yang jarang maupun padat, juga data yang sangat miring/*skewed* dan seragam
- Histogram multidimensional dapat menangkap dependensi di antara atribut

Histogram:

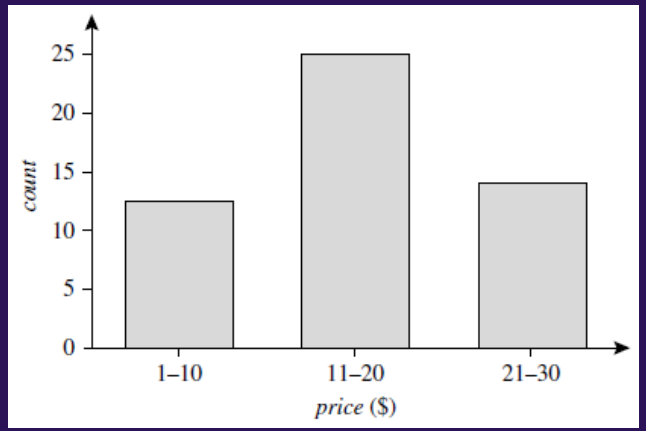
Contoh data:

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.



← Histogram menggunakan *bucket* tunggal

Histogram *equal-width* →



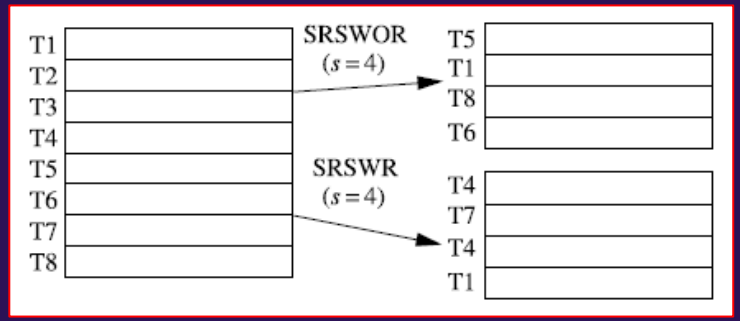
Data Reduction – Reduksi Jumlah Data

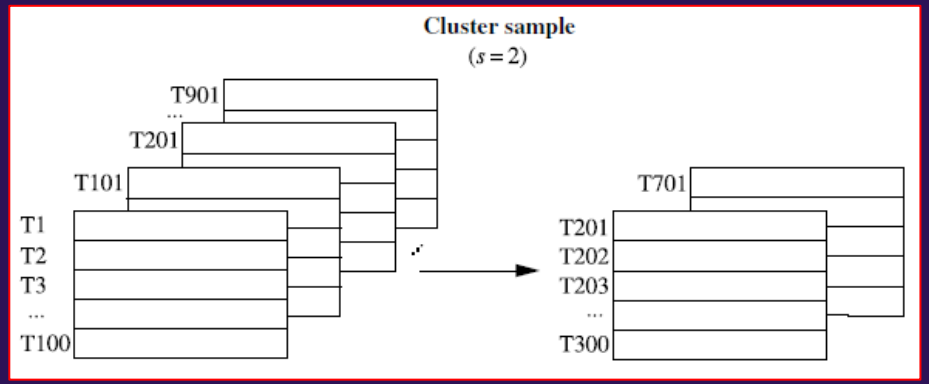
Sampling:

- Merepresentasikan sebuah dataset besar berukuran N dalam bentuk data sampel acak yang berukuran s, dengan $s < N$
- Umum digunakan untuk reduksi data
- Strategi:
  - *Simple random sample without replacement* (SRSWOR): setiap *tuple* dapat terambil sebagai sampel dengan peluang yang sama
  - *Simple random sample with replacement* (SRSWR): setiap *tuple* yang terambil sebagai sampel dapat terpilih kembali
  - *Cluster sample*: data dikelompokkan menjadi beberapa *cluster*
  - *Stratified sample*: data dikelompokkan berdasarkan atribut tertentu menjadi *strata*

Data Reduction – Reduksi Jumlah Data

Sampling:





Startified sample
(according to *age*)

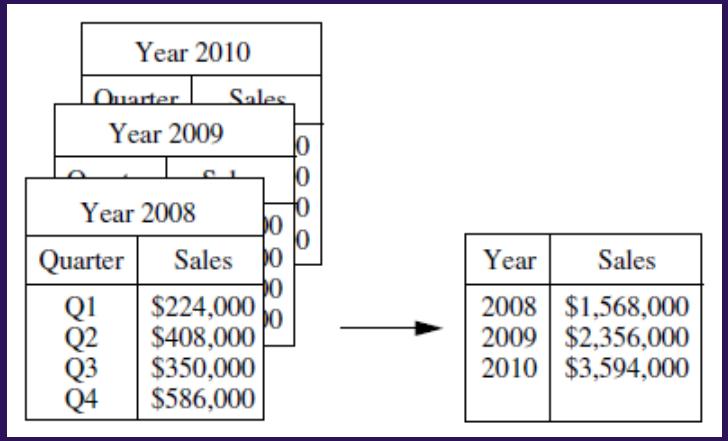
T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

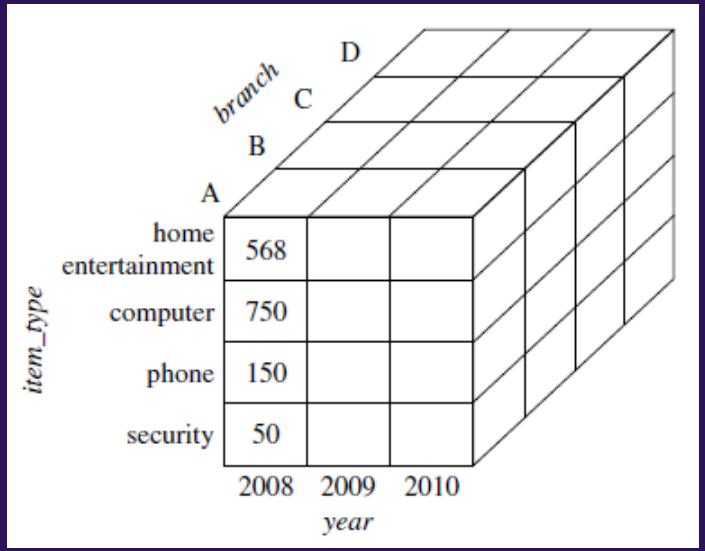
T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Data Reduction – Reduksi Jumlah Data

Data Cube Aggregation:

- Mengumpulkan data & menyatakannya dalam satuan yang lebih besar
- Umum digunakan pada *data warehouse*





Data Transformation

- Mengubah data menjadi bentuk tertentu, sehingga proses *mining* dapat berjalan lebih efisien
- Strategi:
  - *Smoothing*
  - *Attribute construction*
  - Agregasi
  - Normalisasi
  - Diskretisasi
  - Generalisasi

Attribute/Feature Construction:

- Memunculkan sebuah atribut baru sebagai hasil kombinasi dari beberapa atribut lain yang sudah tersedia
- Dapat menemukan informasi mengenai hubungan antara beberapa atribut
- Contoh:
  - Atribut panjang & lebar → atribut luas = panjang × lebar

Normalisasi/standarisasi:

- Mengubah nilai atribut menjadi nilai yang lain pada rentang tertentu
- Tujuan: memberikan bobot yang seimbang terhadap semua atribut
- Variasi:
  - Normalisasi *min-max*
    - $v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$

      min_A = nilai terkecil atribut semula
      max_A = nilai terbesar atribut semula
      new_min_A = nilai terkecil atribut setelah normalisasi
      new_max_A = nilai terbesar atribut setelah normalisasi

    - *Min* = \$12.000, *max* = \$98.000, *v* = \$73.600

      ➔ $v' = \frac{73.600 - 12.000}{98.000 - 12.000}(1 - 0) + 0 = \frac{61.600}{86.000} = 0{,}716$

Normalisasi/standarisasi:

- Normalisasi *z-score* → nilai standar; untuk melihat seberapa jauh penyimpangan nilai atribut terhadap rataan
  - $v_i' = \frac{v_i - \bar{A}}{\sigma_A}$ $\bar{A}$ = rata-rata atribut, σ_A = simpangan baku atribut
  - Rata-rata = \$54.000, simpangan baku = \$16.000, *v* = 73.600

    → $v = \frac{73.600 - 54.000}{16.000} = \frac{19.600}{16.000} = 1{,}225$

- Normalisasi penskalaan desimal
  - $v_i' = \frac{v_i}{10^j}$ j = bilangan bulat terkecil sedemikian sehingga $\max(|v_i'|) < 1$
  - Sebuah atribut dengan rentang nilai -562 s.d. 875 → dibagi 1000

    -562/1000 = -0,562

    875/1000 = 0,875

Diskretisasi:

- Mengubah nilai variabel yang cukup besar menjadi nilai yang lebih kecil
- Mengubah nilai variabel yang kontinu menjadi interval data yang terbatas
- Jenis:
  - *Supervised*: diskretisasi dilakukan berdasarkan suatu kelas data yang telah diketahui
  - *Unsupervised*: diskretisasi dilakukan tanpa menggunakan informasi yang telah diketahui
- Teknik:
  - *Binning*
  - Analisis histogram
  - Analisis *cluster*

Diskretisasi:

Data usia: 1, 5, 9, 4, 7, 11, 14, 17, 13, 18, 19, 31, 33, 36, 42, 44, 46, 70, 74, 78, 77

Nilai Data	1, 5, 4, 9, 7	11, 14, 17, 13, 18, 19	31, 33, 36, 42, 44, 46	70, 74, 78, 77
Label	Anak-anak	Remaja	Dewasa	Lansia

Hasil diskretisasi:

Anak-anak, Anak-anak, Anak-anak, Anak-anak, Anak-anak, Remaja, Remaja, Remaja, Remaja, Remaja, Remaja, Dewasa, Dewasa, Dewasa, Dewasa, Dewasa, Dewasa, Lansia, Lansia, Lansia, Lansia

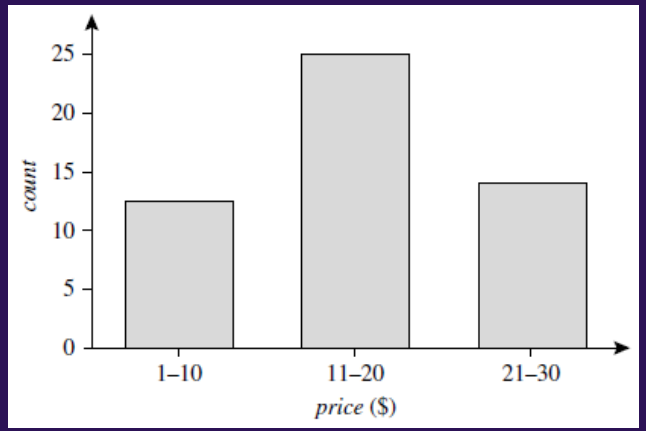
Diskretisasi:

Contoh data:

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Hasil diskretisasi:

A, A, A, A, A, A, A, A, A, A, A, A, A, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, C, C, C, C, C, C, C, C, C, C, C, C, C, C



Generalisasi/pembangkitan hirarki konsep:

- Mendefinisikan nilai suatu atribut pada konsep dengan hirarki yang lebih tinggi
- Dapat diterapkan pada data yang mengandung informasi:
  - Kewilayahan (geografis): data alamat → data kota
  - Waktu: data per hari → data per bulan

No	Tanggal_Daftar	Bulan
1	03-09-2021	September 2021
2	01-10-2021	Oktober 2021
3	20-09-2021	September 2021
4	31-10-2021	Oktober 2021
5	15-10-2021	Oktober 2021
6	09-09-2021	September 2021
7	30-09-2021	September 2021

Soal Latihan

No	Panjang (cm)	Lebar (cm)
1	40	10
2	43	12
3	47	14
4	42	12
5	47	*y*
6	43	12
7	48	15
8	46	14
9	49	16
10	*x*	12

Diketahui data hasil pengukuran panjang & lebar dari 10 objek

1. Isilah variabel *x* & *y* menggunakan rata-rata dari masing-masing atribut
2. Lakukan praproses terhadap data yang sudah lengkap menggunakan metode:
   a) Normalisasi *min-max* dalam range 0-1
   b) *Equal-frequency binning*, dengan frekuensi 5

Recap

- *Data reduction*
- *Data transformation*

Next: apa tujuan melakukan klasifikasi data?

つづく