DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi

PREPROCESSING

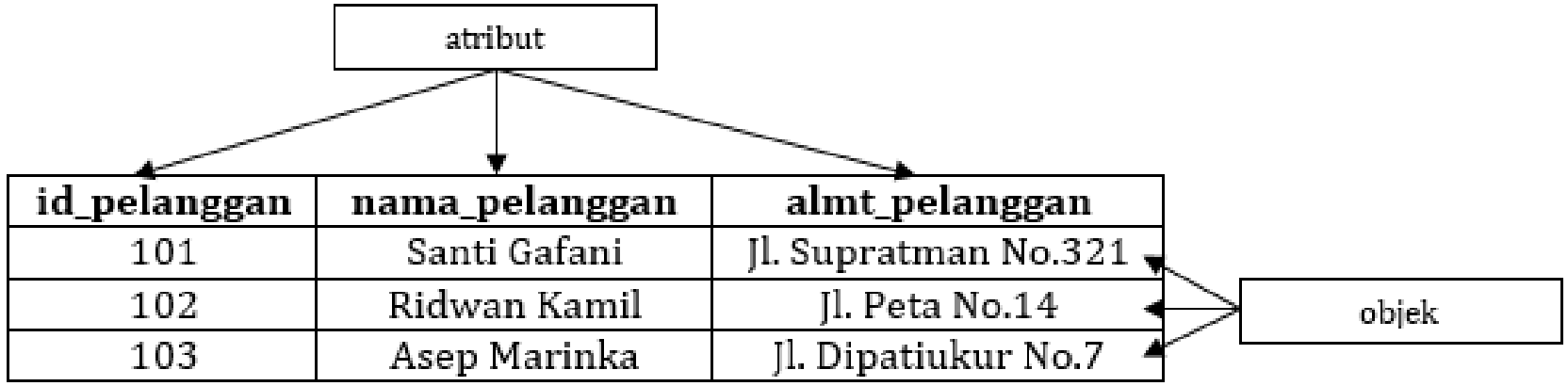
Nizar Rabbi Radliya
nizar@email.unikom.ac.id

DEFINISI SET DATA

Set Data / *Data Set* / Himpunan Data → Kumpulan objek dan atributnya.

Objek = record, point, vector, pattern, event, observation, case, sample, instance, entitas.

Atribut = variabel, field, fitur, atau dimensi.



TIPE DATA

Empat sifat yang dimiliki atribut secara umum, yaitu:

1. Pembeda (distinctness): = dan ≠
2. Urutan (order): <, >, ≤, ≥
3. Penjumlahan, Pengurangan (addition): + dan –
4. Perkalian, Pembagian (multiplication): * dan /

TIPE DATA

Tipe Atribut		Penjelasan	Contoh
Kategoris (Kualitatif)	Nominal	Nilai atribut berupa nominal memberikan nilai berupa nama. Dengan nama inilah sebuah atribut membedakan dirinya pada data yang satu dengan yang lain (=, ≠).	Kode Pos, NIM, Jenis Kelamin.
	Ordinal	Nilai atribut bertipe ordinal mempunyai nilai berupa nama yang mempunyai arti informasi terurut (<, >, ≤, ≥).	Indek Nilai (A, B, C, D, E)
Numerik (Kuantitatif)	Interval	Nilai atribut dimana perbedaan diantara dua nilai mempunyai makna yang berarti (+, -).	Tanggal
	Rasio	Nilai atribut dimana perbedaan diantara dua nilai dan rasio dua nilai mempunyai makna yang berarti (*, /)	Panjang, berat, tinggi

TIPE DATA

Sementara berdasarkan jumlah nilainya, atribut dapat dibedakan menjadi:

1. Diskret

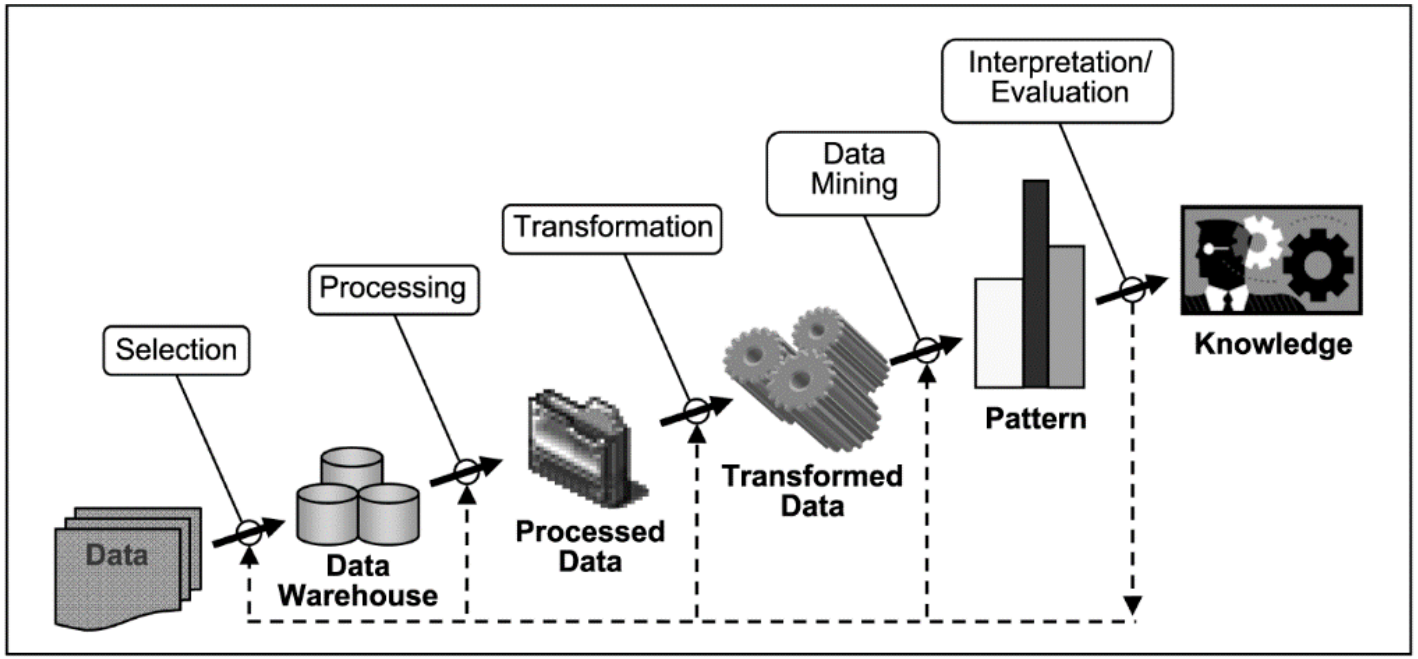
Mempunyai nilai dalam himpunan jumlah yang terbatas atau domainnya terbatas. Contoh: indek nilai (A, B, C, D, E), jenis kelamin (pria, wanita), benar/salah, ya/tidak, 0/1.

2. Kontinu

Mempunyai jangkauan nilai real. Biasanya menggunakan representasi floating point (desimal).

Contoh: panjang, tinggi, berat.

PREPROCESSING



- Agregasi (aggregation)
- Penarikan contoh (sampling)
- Diskretisasi dan binerisasi (discretization and binarization)
- Pemilihan fitur (feature subset selection)
- Transformasi atribut (attribute transformation)

AGREGASI (AGGREGATION)

- ✓ Proses mengkombinasikan dua atau lebih objek ke dalam sebuah objek tunggal;
- ✓ Sangat berguna ketika pada set data ada sejumlah nilai dalam satu fitur yang sebenarnya satu kelompok;
- ✓ Tidak akan menyimpang dari deskripsi fitur tersebut jika nilainya digabungkan.

Agregasi yang dapat dilakukan adalah sum (jumlah), average (rata-rata), min (terkecil), max (terbesar).

AGREGASI (AGGREGATION)

Tabel 1. Set Data Transaksi Pembelian Oleh Pelanggan

Cabang	IDT	Tanggal	Total
Bandung	B01001	30-01-2015	250.000
Bandung	B01002	30-01-2015	300.000
Tasikmalaya	T01001	30-01-2015	500.000
Tasikmalaya	T01002	30-01-2015	450.000
Tasikmalaya	T01003	31-01-2015	350.000

Tabel 2. Set Data Transaksi Pembelian Oleh Pelanggan Setelah Agregasi

Cabang	Tanggal	Total
Bandung	30-01-2015	550.000
Tasikmalaya	30-01-2015	950.000
Tasikmalaya	31-01-2015	350.000

AGREGASI (AGGREGATION)

Beberapa alasan melakukan agregasi:

- ✓ Set data yang lebih kecil akan membutuhkan memori penyimpanan yang lebih sedikit (pengurangan data atau perubahan skala).
- ✓ Waktu pemrosesan dalam algoritma data mining menjadi lebih cepat.
- ✓ Agregasi bertindak untuk mengubah cara pandang terhadap data dari level rendah menjadi level tinggi.
- ✓ Perilaku pengelompokan objek atau atribut sering kali lebih stabil dari pada objek individu itu sendiri (lebih sedikit variasinya).

DISKRETISASI DAN BINERISASI (DISCRETIZATION AND BINARIZATION)

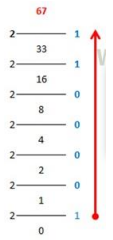
- ✓ Transformasi data dari tipe kontinu dan diskret ke atribut biner disebut binerisasi (binarization).
- ✓ Transformasi data dari atribut kontinu ke atribut kategoris disebut diskretisasi (discretization).

BINERISASI (BINARIZATION)

- ✓ M macam nilai kategoris, masing-masing diberikan nilai yang unik dengan nilai integer dalam jangkauan [0,M-1]
- ✓ Jumlah bit yang dibutuhkan untuk binerisasi adalah N=$[log_2(M)]$

Tabel 3. Konversi Atribut Kategoris ke Tiga Atribut Biner

Nilai Kategoris	Nilai Integer	Nilai Biner		
		X1	X2	X3
Rusak	0	0	0	0
Jelek	1	0	0	1
Sedang	2	0	1	0
Bagus	3	0	1	1
Sempurna	4	1	0	0



DISKRETISASI (DISCRETIZATION)

- ✓ Pertama, memutuskan berapa jumlah kategori yang harus digunakan.
- ✓ Kedua, menentukan bagaimana memetakan nilai-nilai dari atribut kontinyu ke nilai kategoris.

Contoh nilai yang ada pada tabel 4 diubah menjadi atribut katarogikal dengan nilai: rendah, sedang, tinggi.

Tabel 4. Contoh Atribut Kontinu Yang Akan Didiskretisasi

Atribut Kontinu
125
100
70
120
95
60
220
85
75
90

Pendekatan equal width:

Range data [60 - 220]

Rendah: range [60-113]

Sedang: range [114-167]

Tinggi: range [168-220]

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

- ✓ Sebagain fungsi dari transformasi atribut adalah standarisasi dan normalisasi.
- ✓ Tujuan dari standarisasi dan normalisasi adalah untuk membuat keseluruhan nilai mempunyai suatu sifat khusus.

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Salah satu contoh transformasi standarisasi adalah dengan cara:

1. Hitung nilai tengah dengan median;
2. Hidung absolute standard deviation dengan persamaan.

Rumus persamaan yang akan digunakan:

$$\sigma_A = \sum_{i=1}^{m} |x_i - \mu|$$

$$x' = \frac{(x - \mu)}{\sigma_A}$$

Median untuk **jumlah data (*n*) ganjil**

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

Median untuk **jumlah data (*n*) genap**

$$Me = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Sebagai contoh lakukan standarisasi dari data set berikut x = {2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2, 1, 1.5, 1.1}. Dari data tersebut dihitung median = μ=(1.9+2)/2=1.95.

Tabel 5. Contoh Standarisasi

x	x-μ	| x-μ |	x′
0.5	-1.45	1.45	-0.24
1.0	-0.95	0.95	-0.16
1.1	-0.85	0.85	-0.14
1.5	-0.45	0.45	-0.08
1.9	-0.05	0.05	-0.01
2.0	0.05	0.05	0.01
2.2	0.25	0.25	0.05
2.3	0.35	0.35	0.06
2.5	0.55	0.55	0.1
3.1	1.15	1.15	0.19
		$\sigma_A = 6.1$	

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Transformasi atribut menggunakan normalisasi menggunakan pendekatan linear, yang pertama kita terlebih dahulu menghitung rata-rata (persamaan 1) dan varian (persamaan 2) dengan rumus:

$$x_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \qquad \text{(persamaan 1)}$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - x_k) \qquad \text{(persamaan 2)}$$

Data hasil normalisasi dapat dihitung menggunakan cara pertama dengan persamaan berikut:

$$x_{ik} = \frac{x_{ik} - x_k}{\sigma_k} \qquad \text{(persamaan 3)}$$

Hasil normalisasi dengan cara persamaan 3 didapatkan fitur yang mempunyai sifat ***zero-mean dan unit variance***.

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Sebagai contoh ada data $X = \{x_1, x_2, x_3, x_4, x_5\}^T$, dimana untuk $x_1 = \{0, 2, 1\}$, $x_2 = \{1, 7, 1\}$, $x_3 = \{2, 6, 3\}$, $x_4 = \{5, 1, 4\}$, $x_5 = \{3, 3, 4\}$.

Jangkauan nilai untuk fitur pertama adalah [0,5], fitur kedua [1,7], fitur ketiga [1,4]. Masing-masing fitur memiliki jangkauan yang tidak sama.

Tabel 6. Contoh Data Belum Normal

Fitur 1	Fitur 2	Fitur 3
0	2	1
1	7	1
2	6	3
5	1	4
3	3	4

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Jika dilakukan normalisasi menggunakan pendekatan linear yang pertama, dihitung terlebih dahulu rata-rata dan standar deviasi. Untuk fitur pertama, didapatkan:

$$x_1 = \frac{1}{5} \times (0 + 1 + 2 + 5 + 3) = 2.2$$

$$\sigma_1^2 = \frac{1}{5-1} \times ((0 - 2.2)^2 + (1 - 2.2)^2 + (2 - 2.2)^2 + (5 - 2.2)^2 + (3 - 2.2)^2) = 3.7$$

$$\sigma_1 = 1.9235$$

$$x_{11} = \frac{0 - 2.2}{1.9235} = -1.1437$$

$$x_{21} = \frac{1 - 2.2}{1.9235} = -0.6239$$

$$x_{31} = \frac{2 - 2.2}{1.9235} = -0.1040$$

$$x_{41} = \frac{5 - 2.2}{1.9235} = -1.4557$$

$$x_{51} = \frac{3 - 2.2}{1.9235} = -0.4159$$

Fitur 1
0
1
2
5
3

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Tabel 7. Hasil Normalisasi *Zero-Mean* dan *Unit Variance*

Fitur 1	Fitur 2	Fitur 3
-1.1437	-0.6954	-1.0550
-0.6239	1.236	-1.0550
-0.1040	0.8499	0.2638
1.4557	-1.0817	0.9231
0.4159	-0.3091	0.9231

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Teknik linear yang lain adalah dengan menskalakan jangkauan setiap fitur dalam jangkauan [0,1]:

$$x_{ik} = \frac{x_{ik} - \min(x_k)}{\max(x_k) - \min(x_k)}$$

Tabel 8. Hasil Normalisasi Linear [0,1]

Fitur 1	Fitur 2	Fitur 3
0	0.1667	0
0.2000	1.0000	0
0.4000	0.8333	0.6667
1.0000	0	1.0000
0.6000	0.3333	1.0000

Tabel 6. Contoh Data Belum Normal

Fitur 1	Fitur 2	Fitur 3
0	2	1
1	7	1
2	6	3
5	1	4
3	3	4

TRANSFORMASI ATRIBUT (ATTRIBUTE TRANSFORMATION)

Teknik linear yang lain adalah dengan menskalakan jangkauan setiap fitur dalam jangkauan [-1,1] :

$$x_{ik} = \frac{2x_{ik} - (\max(x_k) + \min(x_k))}{\max(x_k) - \min(x_k)}$$

Tabel 9. Hasil Normalisasi Linear [-1,1]

Fitur 1	Fitur 2	Fitur 3
-1.0000	-0.6667	-1.0000
-0.6000	1.0000	-1.0000
-0.2000	0.6667	0.3333
1.0000	-1.0000	1.0000
0.2000	-0.3333	1.0000

Tabel 6. Contoh Data Belum Normal

Fitur 1	Fitur 2	Fitur 3
0	2	1
1	7	1
2	6	3
5	1	4
3	3	4

NEXT

SIMILARITY