Materi 3

DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi | UNIKOM | 2016

Nizar Rabbi Radliya | nizar.radliya@yahoo.com

Nama Mahasiswa	
NIM	
Kelas	

Kompetensi Dasar

Memahami set data yang digunakan pada proses data mining dan mampu melakukan pemrosesan awal pada set data.

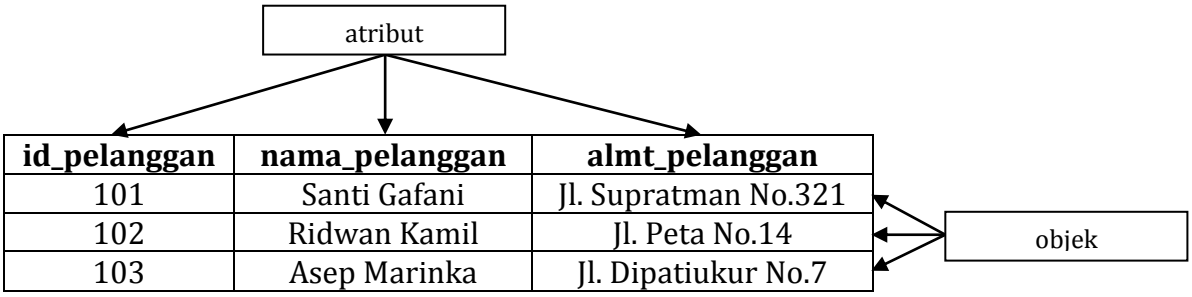
Pokok Bahasan

Set Data dan Pemrosesan Awal:
1. Definisi dan tipe data
2. Kualitas data
3. Pemrosesan Awal Data

I. Definisi Set Data

Set data (*data set*/himpunan data) merupakan kumpulan objek dan atributnya. Nama lain dari objek yang sering digunakan diantaranya *record, point, vector, pattern, event*, *observation*, *case*, *sample*, *instance*, entitas. Objek digambarkan dengan sejumlah atribut yang menerangkan sifat atau karakteristik dari objek tersebut. Atribut juga sering disebut variabel, *field*, fitur, atau dimensi. Atribut adalah sifat/properti/karakteristik objek yang nilainya bisa bermacam-macam dari satu objek dengan objek lainnya, dari satu waktu ke waktu yang lainnya.

Sebagai contoh seorang pelanggan merupakan objek, dimana objek pelanggan tersebut memiliki beberapa atribut seperti id pelanggan, nama, alamat dan lain-lain. Setiap pelanggan memungkinkan memiliki nilai atribut yang berbeda dengan pelanggan lainnya, serta memungkinkan perubahan nilai atribut dari waktu ke waktu.



id_pelanggan	nama_pelanggan	almt_pelanggan
101	Santi Gafani	Jl. Supratman No.321
102	Ridwan Kamil	Jl. Peta No.14
103	Asep Marinka	Jl. Dipatiukur No.7

Gambar 1. Perbedaan Atribut dan Objek

II. Tipe Data

Tipe atribut dapat dibedakan dari nilai beserta sifatnya. Ada 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 /

Umumnya tipe atribut ini ada dua yaitu kategoris (kualitatif) dan numerik (kuantitatif). Dari kedua tipe tersebut dibagi lagi menjadi beberapa sub tipe yang disesuaikan dengan sifat nilai yang dimilikinya.

Tabel 1. Tipe Atribut

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, 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

Atribut nominal dan ordinal merupakan tipe kategoris, nilainya kualitatif; dimana nilai tersebut sebenarnya simbolik; tidak mungkin dilakukan operasi aritmatika. Sedangkan interval dan rasio merupakan tipe numerik, nilainya kuantitatif; dimana nilai tersebut dapat dilakukan operasi aritmatika; bisa direpresentasikan dengan nilai integer atau kontinu.

Sementara berdasarkan jumlah nilainya, atribut dapat dibedakan menjadi dua, yaitu:

1. Diskret

Sebuah atribut dapat bernilai diskret jika mempunyai nilai dalam himpunan jumlah yang terbatas. Jenis ini bisa ditemukan pada atribut kategoris yang hanya mempunyai beberapa variasi nilai (domain), seperti indek nilai yang hanya mempunyai

lima kemungkinan nilai (A, B, C, D, E). Contoh lainnya adalah jenis kelamin (pria, wanita), benar/salah, ya/tidak, 0/1.

2. Kontinu

Sedangkan atribut yang bernilai kontinu akan mempunyai jangkauan nilai real. Seperti variabel panjang, tinggi, berat dimana nilainya biasanya menggunakan representasi *floating point* (desimal). Namun, meskipun menggunakan representasi real, ukuran presisi jumlah angka di belakang koma tetap digunakan.

III. Karakteristik Set Data

Ada tiga karakteristik umum set data yang mempunyai pengaruh besar dalam data mining, yaitu dimensionalitas, sparsitas, resolusi. Berikut adalah penjelasan dari ketiga karakteristik tersebut:

1. Dimensionalitas (*dimensionality*)

a. Dimensionalitas dapat diartikan sebagai jumlah atribut yang dimiliki oleh objek-objek dalam data set.

b. Data dengan jumlah dimensi yang sedikit (rendah) punya kecendrungan berbeda secara kualitatif dengan data dalam kontek yang sama, tetapi dengan jumlah dimensi yang lebih banyak (tinggi).

c. Kesulitan yang berhubungan dengan data dimensi tinggi sering disebut sebagai *curse of dimensionality*.

d. Untuk itu pada tahap *preprocessing* (proses awal) perlu dilakukan pengurangan dimensi (*dimensionality reduction*)

2. Sparsitas (*sparsity*)

a. Untuk set data dengan fitur asimetrik (jumlah fitur yang terisi nilai tidak sama antara satu data dengan data yang lain), banyak atribut data mempunyai nilai 0 di dalmnya; dalam banyak kasus, kurang dari 1% mempunyain nilai bukan 0.

b. Dalam praktiknya, tentu ini menguntungkan karena komputasi menjadi lebih ringan (cepat) dan kapasitas penyimpanan juga lebih sedikit.

3. Resolusi (*resolution*)

a. Untuk data yang digambarkan dalam bentuk grafik yang memerlukan koordinat spasial, karakteristik resolusi yang digunakan juga akan berpengaruh.

b. Pola dalam data bergantung pada level resolusi.

c. Jika resolusi terlalu baik (tidak ada perbedaan/halus), pola mungkin tidak akan kelihatan, jika resolusi terlalu kasar atau sempit, pola juga akan hilang.

IV. Jenis Set Data

Jenis data dapat dibedakan menjadi tiga kelompok, yaitu data record, data berbasis grafik (*graph data*), dan data terurut (*ordered data*). Pada perkuliahan ini kita hanya akan menggunakan data record untuk proses data mining.

Kebanyakan metode data mining mengasumsikan bahwa set data yang diproses adalah kumpulan baris data (*record*/*entries*/*objects*), dimana setiap barisnya terdiri atas sejumlah fitur (atribut) yang tetap. Dalam set data berbentuk data record, tidak ada hubungan antara baris data dengan baris data yang satu dengan baris data yang lainnya dan juga dengan set data yang lain. Setiap baris data berdiri sendiri sebagai sebuah data individu.

Dalam sistem basis data, umumnya ada sejumlah tabel yang saling berhubungan menggunakan suatu kunci (kunci utama, kunci tamu). Akan tetapi dalam set data record, diasumsikan bahwa hanya ada satu tabel yang berisi sejumlah baris data. Oleh karena itu, biasanya set data yang diolah dalam data mining adalah keluaran dari sistem data warehouse yang menggunakan *query* untuk melakukan pengambilan data dari sejumlah tabel dalam sistem basis data. Ada beberapa contoh set data yang masuk dalam jenis data record, diantaranya data matrik (*matrix data*), data keranjang belanja (*market basket data*), dan data dokumen.

1. Data Matrik

Jika set data berisi kumpulan data yang mempunyai sejumlah atribut (fitur) numerik yang sama, set data tersebut dapat dipandang sebagai vektor (data) dalam wilayah multidimensi, dimana masing-masing dimensi menyatakan satu atribut yang berbeda. Sekumpulan data matrik dapat diinterpretasikan sebagai mantrik M x N, dimana M adalah jumlah baris (satu baris menyatakan satu record/objek) dan N adalah jumlah kolom (dimana satu kolom menyatakan satu atribut/fitur). Contoh data matrik dapat dilihat pada tabel 2 dibawah ini.

Tabel 2. Contoh Data Matrik

Tinggi	Berat	Sepatu	Celana
168	60	38	30
175	85	42	35
170	77	39	39

2. Data Keranjang Belanja/Transaksi

Data keranjang belanja (data transaksi) adalah set data yang setiap recordnya berisi sejumlah item dan jumlah item untuk sebuah transaksi bisa berbeda dengan transaksi yang lain. Contohnya bisa dilihat pada kasus keranjang belanja di pasar atau supermarket, dimana setiap pembeli melakukan pembelian barang yang jumlah dan jenisnya bisa berbeda dengan pembeli yang lain. Contoh data transaksi dapat dilihat pada tabel 3 dibawah ini.

Tabel 3. Contoh Data Transaksi

TID	Item
1	Susu, Bedak, Sabun
2	Susu, Mentega
3	Bedak, Gula, Sabun, Roti

3. Data Dokumen

Setiap dokumen merupakan satu vektor 'term'. Tiap term merupakan satu komponen (atribut) dari vektor tersebut. Nilai dari setiap komponen menyatakan berapa kali kemunculan term tersebut dalam suatu dokumen. Contoh dari data dokumen dapat dilihat pada tabel 4 dibawah ini.

Tabel 4. Contoh Data Dokumen

	Sistem	Informasi	Data	Proses
Makalah 1	18	19	15	9
Makalah 2	6	6	4	8
Makalah 3	14	0	2	12

V. Kualitas Data

Permasalahan kualitas data ditinjau dari aspek pengukuran data dan pengumpulan data.

1. Kesalahan Pengukuran (*Measurement Error*)

Kesalahan ini mengacu pada permasalahan hasil dari proses pengukuran. Problem yang umum terjadi adalah nilai yang dicatat berbeda dari nilai sebenarnya untuk beberapa tingkat. Pada atribut kontinu, beda numerik dari hasil pengukuran dengan nilai sebenarnya disebut dengan *error*. Yang termasuk dalam jenis kesalahan pengukuran adalah:

a. *Noise*

Merupakan komponen random dari suatu error pengukuran. Noise berkaitan dengan modifikasi dari nilai asli. Contoh: distorsi atau penyimpangan dari suara orang saat berbicara di telepon yang jaringannya buruk.

b. Bias

Suatu variasi pengukuran dari kuantitas yang sedang diukur dengan pengurangan antara mean dan nilai kuantitas yang diketahui.

c. *Precision*

Kedekatan dari pengukuran berulang (dari kuantitas yang sama) satu dengan yang lainnya. Diukur dengan standar deviasi.

d. *Accuracy*

Kedekatan pengukuran terhadap nilai sebenarnya dari kuantitas yang sedang diukur.

Contoh kasus:

Terdapat berat standar laboratorium suatu benda adalah 1 gram dan kita akan menghitung *precision* dan bias dari skala benda dari hasil pengukuran yang baru. Kita melakukan pengukuran sebanyak lima kali dan memperoleh {1.015, 0.990, 1.013, 1.001, 0.986}.

Jawab:

Mean = (1.015 + 0.990 + 1.013 + 1.001 + 0.986) / 5 = 1.001

Bias = 1.001 – 1 = 0.001

Precision=

$$\sqrt{\frac{(1.015-1.001)^2+(0.990-1.001)^2+(1.013-1.001)^2+(1.001-1.001)^2+(0.986-1.001)^2}{4}}$$

2. Kesalahan Pengumpulan (Collection Error)

Mengacu pada kesalahan-kesalahan (*error*) seperti hilangnya objek data atau nilai atribut, atau lingkup objek data yang tidak tepat. Yang temasuk dalam kesalahan pengumpulan diantaranya:

a. Outliers

Merupakan objek data dengan sifat yang berbeda sekali dari kebanyakan objek data dalam data-set. Misalkan, terdapat data penelitian tentang tinggi anak siswa SMA yakni 160cm sampai 180cm. Tetapi dalam data tersebut terdapat anak yang mempunyai

tinggi 140cm. Data anak dengan tinggi 140cm tersebut yang disebut data outlier, karena berbeda sangat jelas.

Terdapat beberapa hal yang mempengaruhi munculnya data outlier antara lain:

1) Kesalahan dalam pemasukan data
2) Kesalahan dalam pengambilan *sample*
3) Memang ada data-data ekstrim yang tidak dapat dihindarkan keberadaannya.

b. Missing value

Merupakan nilai dari suatu atribut yang tidak ditemukan. Asalannya terjadinya *missing value* adalah:

1) Informasi tidak diperoleh (misal, orang-orang menolak untuk memberikan data umur dan berat bedan)
2) Atribut yang mungkin tidak bisa diterapkan ke semua kasus (misal, pendapatan tahunan tidak bisa diterapkan pada seseorang yang pengangguran)

Penanganan *missing values* adalah dengan:

1) Mengurangi objek data
2) Memperkirakan *missing values*
3) Mengabaikan *missing values* pada saat analisis
4) Mengganti dengan semua nilai yang mungkin (tergantung probabilitasnya)

c. *Duplicate data*

Set data mungkin terdiri dari objek data yang ganda (duplikat), atau hampir selalu terjadi duplikasi antara satu dengan yang lainnya. Persoalan utama ketika menggabungkan data dari sumber-sumber yang bervariasi (heterogen). Contoh: orang yang sama dengan alamat email yang lebih dari satu. Pembersihan data (*data cleaning*) merupakan proses yang berkaitan dengan permasalahan data yang duplikat.

VI. Pemrosesan Awal Data

Set data yang akan diproses dengan metode-motode data mining sering kali harus melalui pemrosesan awal. Langkah ini masuk ke dalam tahapan KDD sebelum proses data mining. Beberapa permasalahan seperti jumlah populasi data yang besar, banyaknya data yang menyimpang (anomali data), dimensi yang terlalu tinggi, banyaknya fitur yang tidak berkontribusi besar, dan lain-lain merupakan pemicu munculnya pemrosesan awal data (*pre-processing*) yang harus diterapkan pada set data sebelum akhirnya digunakan dalam proses data mining. Beberapa pekerjaan yang umum dilakukan sebagai pemrosesan awal pada set data akan dibahas pada bab-bab di bawah ini.

6.1. Agregasi

Agregasi (*aggregation*) adalah proses mengkombinasikan dua atau lebih objek ke dalam sebuah objek tunggal. Aggregasi data sangat berguna ketika pada set data ada sejumlah nilai dalam satu fitur yang sebenarnya satu kelompok, yang tidak akan menyimpang dari deskripsi fitur tersebut jika nilainya digabungkan. Agregasi yang dapat dilakukan adalah *sum* (jumlah), *average* (rata-rata), *min* (terkecil), *max* (terbesar).

Sebagai contoh adalah data transaksi pembelian di beberapa cabang distributor. Setiah hari masing-masing cabang melakukan banyak sekali transaksi. Semua transaksi tersebut akan menghasilkan data yang besar dan komplek. Oleh sebab itu data tersebut akan lebih sederhana tetapi tetap tidak menghilangkan deskripsinya apabila disajikan dalam bentuk gabungan setiap harinya di masing-masing cabang. Dengan begitu, pemrosesan data dalam data mining akan relatif lebih sederhana dan komputasinya menjadi lebih cepat. Selain itu dampaknya adalah penggunaan perangkat penyimpanan menjadi lebih sedikit atau kecil. Lebih jelasnya dapat dilihat pada tabel-tabel di bawah ini.

Tabel 5. 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

Misalnya kita menggunakan agregasi sum pada kolom total, dikelompokan berdasarkan kolom tanggal dan kolom IDT dapat dihilangkan sehingga hasilnya tampak seperti pada tabel 6 di bawah ini.

Tabel 6. 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

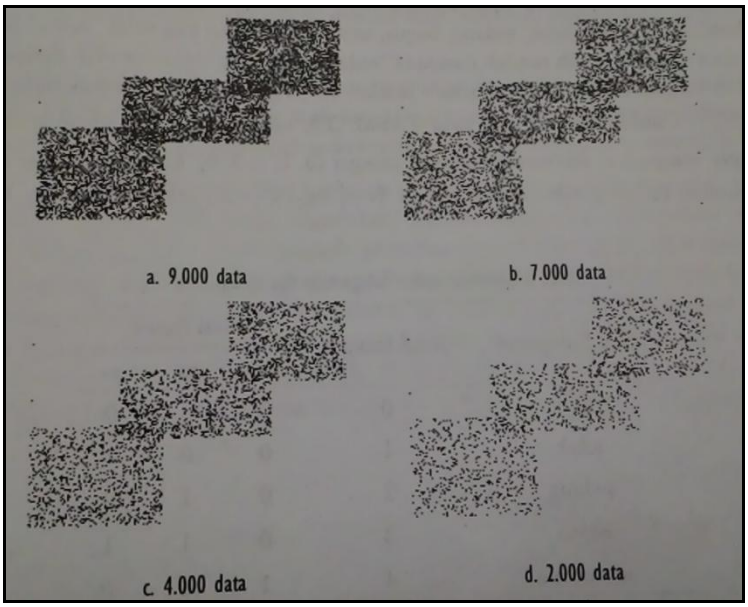
Ada beberapa alasan mengapa sebaiknya melakukan agregasi, diantaranya adalah:

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

6.2. Penarikan Contoh

Kunci utama dalam penarikan contoh (*sampling*) adalah bahwa sampel data akan bekerja hampir sama dengan seluruh data jika sample tersebut mampu mewakili (representatif) seluruh data. Sample disebut representatif jika diperkirakan mempunyai sifat yang sama dengan seluruh data, biasanya diukur dengan rata-rata (*mean*) pada sample dan data asli. Jika sama atau sangat mendekati, sample tersebut bisa dikatakan bagus. Tetapi, penggunaan sample yang baik juga tidak menjamin bahwa hasil pemrosesan data mining pada sample juga sama bagusnya dengan pemrosesan pada seluruh data asli.

Ada dua tipe penarikan contoh yang sering digunakan yaitu penarikan contoh tanpa pengembalian dan penarikan contoh dengan pengembalian. Pada teknik yang pertama, setiap data yang sudah terambil untuk digunakan sebagi sample tidak dikembalikan lagi ke data aslinya, sedangkan pada teknik kedua setiap data yang sudah terambil untuk digunakan sebagai sampel dikembalkan ke data asli. Akibatnya, sebuah data memiliki kemungkinan untuk muncul lebih dari satu kali dalam sampel. Sebagai contoh dapat dilihat pada gambar 2 di bawah ini.



Gambar 2. Struktur Data Yang Hilang Karena Penarikan Contoh

Pada gambar 2 di atas dapat dilihat contoh proses *sampling* secara acak pada set data dua dimensi yang berisi 9.000 data. Secara kasat mata penarikan contoh 7.000 data masih memberikan bentuk data yang menyerupai bentuk distribusi data yang asli. Ketika menggunakan 4.000 data, bentuk aslinya masih terlihat. Akan tetapi, ketika menggunakan 2.000 data, hasil penyampelan mulai terdistorsi dan bentuk asli data mulai tidak tampak.

6.3. Binerisasi dan Diskretisasi

Beberapa algoritma data mining, khususnya algoritma klasifikasi membutuhkan data dalam bentuk atribut kategorikal. Sedangkan algoritma asosiasi memerlukan data dalam bentuk atribut biner. Transformasi data dari tipe kontinu dan diskret ke atribut biner disebut binerisasi (*binarization*) sedangkan transformasi data dari atribut kontinu ke atribut kategoris disebut diskretisasi (*discretization*).

Cara pertama untuk melakukan binerisasi adalah dari M macam nilai kategoris, masing-masing diberikan nilai yang unik dengan nilai integer dalam jangkauan [0,M-1]. Jika atribut ordinal, urutan nilai kategorisnya harus diperhatikan. Misalnya untuk nilai kategoris kualitas = {rusak, jelek, sedang, bagus, sempurna}, nilai-nilai tersebut memiliki urutan nilai dari rendah ke tinggi (kalau dari contoh nilai kualitas tersebut dari kiri ke kanan). Jumlah bit yang dibutuhkan untuk binerisasi adalah N=$\lceil log_2(M) \rceil$.

Sebagai contoh dapat dilihat pada tabel 7 di bawah ini, dimana nilai kategoris kualitas = {rusak, jelek, sedang, bagus, sempurna} dikonversi menjadi nilai integer {0, 1,

2, 3, 4}. Karena ada lima macam nilai kategoris, jumlah bit yang dibutuhkan adalah N=$\lceil log_2(5) \rceil = 3$, yaitu menjadi tiga atribut biner x1, x2, x3.

Tabel 7. 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

Sedangkan untuk melakukan diskretisasi terdiri atas dua langkah. Pertama, memutuskan berapa jumlah kategori yang harus digunakan. Langkah kedua, menentukan bagaimana memetakan nilai-nilai dari atribut kontinyu ke nilai kategori. Sebagai contoh nilai yang ada pada tabel 8 diubah menjadi atribut katarogikal dengan nilai: rendah, sedang, tinggi.

Tabel 8. 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]

Rendah: range [114-167]

Rendah: range [168-220]

6.4. Pemilihan Fitur

Salah satu cara untuk mengurangi dimensi set data adalah dengan memilih fitur yang tepat atau hanya menggunaka atribut-atribut yang diperlukan. Secara konseptual, pemilihan sub set fitur merupakan suatu proses pencarian terhadap semua kemungkinan sub set fitur.

Dalam memilih fitur perlu diperhatikan fitur-fitur yang memiliki duplikasi informasi yang tergantung dalam satu dan lebih atribut lain. Selain itu fitur-fitur yang tidak relevan yaitu fitur-fitur yang tidak mengandung informasi berguna untuk tugas data mining secara langsung. Sebagai contoh NIM setiap mahasiswa sering tidak relevan untuk memprediksi IPK mahasiswa.

6.5. Transfromasi Atribut

Transformasi atribut berfungsi untuk memetakan keseluruhan himpunan nilai dari atribut yang diberikan ke suatu himpunan nilai-nilai pengganti yang baru sedemikian hingga nilai yang lama dapat dikenali dengan satu dari nilai-nilai baru tersebut.

Sebagian fungsi dari transformasi atribut adalah standarisasi dan normalisasi. Tujuan dari standarisasi dan normalisasi adalah untuk membuat keseluruhan nilai mempunyai suatu sifat khusus. 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:

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

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

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. Maka dihasilkan hasil standarisasi pada tabel 9 di bawah ini.

Tabel 9. 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
		$\sigma_A = 6.1$	

Sedangkan untuk 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*. Teknik linear yang lain adalah dengan menskalakan jangkauan setiap fitur dalam jangkauan [0,1] (persamaan 4) atau [-1,1] (persamaan 5).

$$x_{ik} = \frac{x_{ik} - \min(x_k)}{\max(x_k) - \min(x_k)} \qquad \text{(persamaan 4)}$$

$$x_{ik} = \frac{2x_{ik} - (\max(x_k) + \min(x_k))}{\max(x_k) - \min(x_k)} \qquad \text{(persamaan 5)}$$

Sebagai contoh ada data $X = \{x_1, x_2, x_3, x_4, x_5\}^T$, dimana untuk $x_1 = \{0, 2, 1\}$, $x_2 = \{1, 7, 1\}$, $x_3 = \{2, 6, 3\}$, $x_4 = \{5, 1, 4\}$, $x_5 = \{3, 3, 4\}$. Jika diperhatikan, jangkauan nilai untuk fitur pertama adalah [0,5], fitur kedua [1,7], fitur ketiga [1,4]. Masing-masing fitur memiliki jangkauan yang tidak sama. Data tersebut disajikan pada tabel 10 di bawah ini.

Tabel 10. 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

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

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

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

$$\sigma_1 = 1.9235$$

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

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

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

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

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

Setelah dihitung pada 3 fitur didapatkan hasil normalisasi sebagai berikut:

Tabel 11. 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

Jika dinormalisasi menggunakan pendekatan linear dengan jangkau [0,1] didapatkan:

Tabel 12. 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

Jika dinormalisasi menggunakan pendekatan linear dengan jangkau [-1,1] didapatkan:

Tabel 13. 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

VII. Daftar Pustaka

[1] Astuti, F.A. 2013. Data Mining. Yogyakarta: Andi.

[2] Kusrini & Taufiz, E.L. 2009. Algoritma Data Mining. Yogyakarta: Andi.

[3] Prasetyo, E. 2012. Data Mining: Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: Andi.

[4] Prasetyo, E. 2014. Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB. Yogyakarta: Andi.

VIII. Materi Berikutnya

Pokok Bahasan	Similaritas dan Dissimilaritas
Sub Pokok Bahasan	1. Ketidakmiripan dan kemiripan data satu atribut 2. Ketidakmiripan data multiatribut 3. Kemiripan data multiatribut