

Data Preprocessing

By. Gunawansyah

Data

- ✓ Data yang ada pada umumnya:
 - ❑ Banyak noise
 - ❑ Ukuran yang besar
 - ❑ Dapat merupakan campuran dari berbagai sumber
- ✓ Memahami data sangat penting untuk preprocessing

Mengapa Perlu Data Preprocessing?

a. Data dalam dunia nyata “dirty”

- ✓ Tidak lengkap: berisi data yang hilang/kosong, kekurangan atribut yang sesuai, hanya berisi data aggregate
 - ✓ e.g., occupation=“ ”

b. Banyak “noise”: berisi data yang outlier atau error


- ✓ e.g., Salary=“-10”

c. Tidak konsisten: berisi nilai yang berbeda dalam suatu kode atau nama

- ✓ e.g., Age=“42” Birthday=“03/07/1997”
- ✓ e.g., Was rating “1,2,3”, now rating “A, B, C”
- ✓ e.g., discrepancy between duplicate records

d. Data yang tidak berkualitas, akan menghasilkan kualitas mining yang tidak baik pula.

e. Data Preprocessing, cleanning, dan transformasi merupakan pekerjaan mayoritas dalam aplikasi data mining (90%).



- ✓ ***Garbage In Garbage Out:*** Tanpa tersedianya data yang berkualitas, jangan mengharapkan hasil mining yang bermutu.

- ✓ Pengambilan keputusan yang bermutu harus berbasis data yang bermutu pula.

- ✓ ***Contoh:***

Data yang ganda atau tidak lengkap mungkin akan menyebabkan kesalahan pada pola-pola sebagai bakal pengetahuan yang dihasilkan, atau bahkan menyesatkan.

Why Preprocess the Data?

Measures for **data quality**: A multidimensional view

Accuracy: correct or wrong, accurate or not

Completeness: not recorded, unavailable, ...

Consistency: some modified but some not, ...

Timeliness: timely update?

Believability: how trustable the data are correct?

Interpretability: how easily the data can be understood?

Ukuran Kualitas :

- ✓ Accuracy
- ✓ Completeness
- ✓ Consistency
- ✓ Timeliness
- ✓ Believability
- ✓ Value added
- ✓ Interpretability
- ✓ Accessibility

Major Tasks in Data Preprocessing

1. Data **cleaning**
 - ▶ Fill in **missing** values
 - ▶ Smooth **noisy** data
 - ▶ Identify or **remove outliers**
 - ▶ Resolve **inconsistencies**
2. Data **reduction**
 - ▶ **Dimensionality** reduction
 - ▶ **Numerosity** reduction
 - ▶ Data **compression**
3. Data **transformation** and data **discretization**
 - ▶ **Normalization**
 - ▶ Concept hierarchy generation
4. Data **integration**
 - ▶ Integration of **multiple databases** or files

Tugas-tugas Utama Data Preprocessing – Jia Wei Han (#1)

✓ ***Data Cleaning***

Mengisi missing values, menghaluskan (smoothing) noisy data, membereskan inkonsistensi.

✓ ***Data Integration***

Menggabung data dari berbagai sumber database yang berbeda ke dalam sebuah penyimpanan seperti data cube atau warehouse.

✓ ***Data Transformation***

Mengubah / mentransformasikan data ke dalam bentuk yang paling tepat / cocok untuk proses data mining.

✓ ***Data Reduction***

Mengurangi volume data, tetapi mempertahankan hasil analisis yang sama atau serupa.

✓ Data Cleaning

- ▶ Data pada dunia nyata cenderung incomplete (tidak lengkap), noisy (terganggu, memuat penyimpangan), dan inconsistent (tidak konsisten).
- ▶ Rutin-rutin data cleaning mencoba untuk:
 - ✓ menangani missing values (nilai-nilai yang hilang)
 - ✓ smoothing (menghaluskan) data yang ber-noise, jika outliers (data-data di luar umumnya)
 - ✓ teridentifikasi memperbaiki inkonsistensi data.

Penanganan Missing Values (#1)

1. Abaikan recordnya

- ✓ Kurang efektif saat terdapat beberapa atribut yang memiliki missing values.
- ✓ Juga saat persentase missing values harus dipertimbangkan.

2. Masukkan nilai yang hilang secara manual

- ✓ Secara umum pendekatan ini sangat menyita waktu dan tidak dapat dikerjakan dengan mudah pada large data sets dengan missing values dalam jumlah besar.

3. Gunakan konstanta umum untuk mengganti nilai yang hilang

- ✓ Ganti semua nilai-nilai atribut yang hilang dengan konstanta yang sama, seperti label "unknown" or nilai -tak terhitung.
- ✓ Meskipun metode ini mudah, tidak dianjurkan untuk digunakan, "unknown" dapat dianggap suatu attribute value.

Penanganan Missing Values (#2)

4. Gunakan rata-rata nilai atribut untuk mengganti nilai yang hilang

- ✓ Misalkan pendapatan rata-rata dari customer AllElectronics adalah \$28,000. Gunakan nilai ini untuk mengganti nilai pendapatan yang kosong.

5. Gunakan rata-rata nilai atribut dari semua sample yang berada pada kelas yang sama

- ✓ Sebagai contoh, jika customer dikelompokkan berdasar tingkat resiko kreditnya, ganti nilai yang hilang dengan rata-rata nilai pendapatan dari semua customer pada tingkat resiko kredit yang sama.

6. Gunakan nilai yang "*paling mungkin*" untuk digantikan pada nilai yang hilang

- ✓ Dapat diperoleh melalui regresi, atau beberapa metode inferensi seperti formula Bayes atau induksi decision tree.

Penanganan Noisy Data (#1)

“Apa yang dimaksud noise ?”

Noise adalah kesalahan yang terjadi secara random atau karena variasi yang terjadi dalam pengukuran variabel.

► ***Solusi:***

Dengan *smoothing* (penghalusan data).

► ***Beberapa pendekatan Smoothing:***

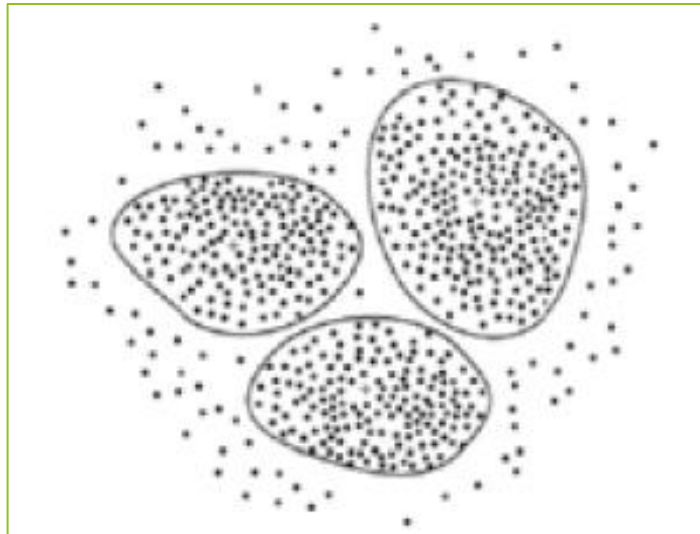
- ✓ Binning
- ✓ Clustering
- ✓ Regression

#Binning

- ✓ Metode-metode binning menghaluskan nilai pada data yang terurut dengan "berkonsultasi" dengan data "tetangganya", yaitu nilai-nilai di sekitarnya.
- ✓ Nilai-nilai yang terurut didistribusikan ke dalam sejumlah "buckets" atau bins.
- ✓ Penghalusan data secara lokal.
- ✓ Pada contoh ini, data pertama kali diurutkan, dan kemudian dipartisi ke dalam bins dengan kedalaman yang sama, misal 3 (setiap bin berisi tiga nilai).
- ✓ Binning juga digunakan sebagai suatu teknik diskretisasi.

#Clustering

- ✓ Data Outliers (di luar nilai yang wajar) dapat dideteksi dengan clustering yang mengelompokkan nilai-nilai yang sama dalam sebuah group (cluster)
- ✓ Secara intuitif, nilai yang berada di luar semua cluster yang terbentuk, dapat dipertimbangkan sebagai outliers.



#Regression

Smoothing dapat dilakukan dengan *fitting (mengepaskan)* data pada sebuah function yang diperoleh dengan perhitungan regresi.

- ✓ **Regresi Linier** melibatkan penemuan garis “terbaik” untuk mencocokkan dua variabel, sehingga satu variabel dapat digunakan untuk meramalkan yang lain.
- ✓ **Multiple Regresi Linier** adalah perluasan dari regresi linear, dimana lebih dari dua variabel dilibatkan dan data disesuaikan pada permukaan multi dimensi.

Penanganan Inkonsistensi Data

- ✓ Beberapa data yang tidak konsisten dapat diperbaiki **secara manual** dengan menggunakan referensi-referensi eksternal.
Contoh: Kesalahan-kesalahan yang dibuat pada data entry dapat diperbaiki dengan melakukan pelacakan pada kertas.
- ✓ Mungkin juga terdapat inkonsistensi **yang disebabkan oleh integrasi data**, dimana atribut-atribut yang diberikan ternyata memiliki nama-nama yang berbeda karena berasal dari database-database yang berbeda.
- ✓ **Redundancy** (perulangan) atribut juga mungkin terjadi

✓ Data Integration (Integrasi Data)

Sering terjadi tugas analisis melibatkan integrasi data, yaitu penggabungan data dari beberapa data stores ke dalam sebuah tempat penyimpanan data seperti data cube atau warehousing.

Sumber-sumber data dapat berupa multiple database, data cubes, atau flat files.

Beberapa Issues dalam Integrasi Data (#1)

1. Entity Identification Problem

- ✓ Bagaimana seorang data analis atau komputer dapat yakin bahwa *customer_id* dalam sebuah database dan *cust_number* dalam database lain sebenarnya menunjuk pada entity yang sama?
- ✓ Database dan data warehouse biasanya mempunyai **metadata - data tentang data**. Metadata dapat digunakan untuk membantu menghindari kesalahankesalahan dalam integrasi.

2. Redudancy (Perulangan)

- ✓ Beberapa perulangan dapat dideteksi dengan **correlation analysis (analisa korelasi)**. Contoh: Saat diberikan dua buah atribut, dapat diukur berapa kuat sebuah atribut berpengaruh pada atribut lainnya.

3. Detection and Resolution of Data Value Conflicts (Deteksi dan Penyelesaian Konflik data)

- ✓ Sebagai contoh, untuk entiti yang sama, nilai-nilai atribut dari beberapa sumber yang berbeda dapat berbeda. Hal ini dapat menyebabkan perbedaan dalam skala representasi.
- ✓ Sebuah atribut weight dapat disimpan dalam unit metric pada satu sistem dan unit British imperial pada sistem lainnya. Gram vs. Ounce.
- ✓ Harga hotel diukur melalui sistem mata uang yang berbeda, termasuk juga pelayanan yang berbeda (seperti free breakfast) dan pajak.

✓ Data Transformation

Tujuannya:

diharapkan lebih efisien dalam proses data mining dan mungkin juga agar pola yang dihasilkan lebih mudah dipahami.

Strategi:

- ▶ Smoothing
- ▶ Attribute (feature) construction
- ▶ Aggregation
- ▶ Normalization
- ▶ Discretization

Normalisasi Data

adalah proses penskalaan nilai atribut dari data sehingga bisa jatuh pada range tertentu.

Contoh:

Misalnya berkenaan dengan pencatatan tingkat kematian penduduk di Indonesia perbulannya berdasarkan jenis umur. Secara Soft Computation Research Group, EEPISITS sederhana, disana ada 3 dimensi data, yaitu bulan (1-12), umur (0-150 misalnya), dan jumlah kematian (0-jutaan).

Kalau kita bentangkan range masing-masing dimensi, maka kita akan mendapatkan ketidak-seimbangan range pada dimensi yang ketiga, yaitu jumlah kematian.

Normalisasi Data

- ✓ Unit ukuran dapat mempengaruhi analisis data.
- ✓ Unit yang lebih kecil akan menghasilkan rentang nilai yang besar
 - ✓ Atribut akan memiliki “bobot” yang lebih besar dari atribut lain sehingga data perlu dinormalisasi atau dibakukan.
- ✓ Hasil suatu normalisasi adalah $[-1, 1]$ atau $[0.0, 1.0]$
- ✓ Diperlukan dalam **klasifikasi** (termasuk neural network dan nearest network) dan **clustering**.
- ✓ Metode Normalisasi data :
 - a) Min-max
 - b) Z-score
 - c) Decimal scaling

Normalization method (Min-Max)

Min-Max merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli.

Rumus:

$$X_i = \frac{(X_i - X_{i\min})}{(X_{i\max} - X_{i\min})}$$

Keuntungan dari metode ini adalah keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses normalisasi. Tidak ada data bias yang dihasilkan oleh metode ini.

Kekurangannya adalah jika ada data baru, metode ini akan memungkinkan terjebak pada "out of bound" error.

Normalization method (Z-Score)

Z-score merupakan metode normalisasi yang berdasarkan mean (nilai rata-rata) dan standard deviation (deviasi standar) dari data.

Rumus:

$$\text{newdata} = (\text{data} - \text{mean}) / \text{std}$$

Metode ini sangat berguna jika kita tidak mengetahui nilai aktual minimum dan maksimum dari data.

Normalization method (Decimal Scaling)

Metode ini melakukan normalisasi dengan menggerakkan nilai desimal dari data ke arah yang diinginkan.

Rumus:

$$\text{newdata} = \text{data} / 10^i$$

dimana i adalah nilai integer untuk menggerakkan nilai desimal ke arah yang diinginkan.

TERIMA KASIH