MODUL MATA KULIAH

DATA MINING

(MIK 620 SESI 10)

DISUSUN OLEH

NOVIANDI, M.Kom
NIDN. 0318018202

PROGRAM STUDI MANAJEMEN INFORMASI KESEHATAN

FAKULTAS ILMU-ILMU KESEHATAN

UNIVERSITAS ESA UNGGUL

2018

BAB III
TEKNIK-TEKNIK PRAPROSES DATA

Capaian pembelajaran

1. Mahasiswa mampu melakukan salah satu teknik dalam praproses data yaitu pemberihan data dan integrasi data.

Pra-proses Data

Pra-proses dilakukan karena dimungkinkan *data set* yang tidak lengkap, mengandung *noise* atau *outlier,* data tidak konsisten, atau ada data yang berulang. Tujuan penting dari pra-proses data adalah untuk meningkatkan kualitas data, sehingga proses data mining juga menghasilkan pengetahuan baru yang lebih baik. Tugas utama dalam pra-proses data adalah pembersihan data, integrasi data, transformasi data, reduksi data dan diskretisasi data.

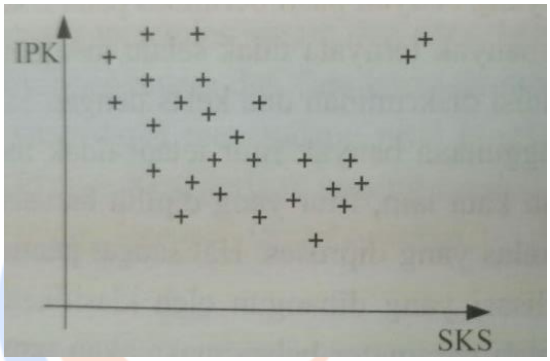
Outlier disebut juga *noise* didefinisikan sebagai titik yang terletak sangat jauh dari rata-rata variabel random pada umumnya yang berkorelasi dengan titik tersebut. Jumlah *outlier* biasanya sedikit, dan *outlier* biasanya dibuang dari data yang diproses. Pendeteksian *outlier* dapat dilakukan dengan menggunakan metode-metode seperti:

a. Pendekatan statistic
b. K-Nearest Neighbor
c. Pemeriksaan kerapatan
d. DBSCAN
e. *Outlier Removal Clustering*
f. Dan lain-lain

Outlier tidak selalu merupakan data dengan perilaku menyimpang yang akhirnya harus dibuang. Adakalanya *outlier* adalah data yang memang akan dicari karena keistimewaan perilakunya.

Contoh:

Kasus data akademik mahasiswa tingkat akhir (SKS banyak) dengan IPK tinggi.



Gambar 2.1 Pola data dengan keberadaan *outlier*

Normalisasi Data

Normalisasi dalam kegiatan data mining merupakan proses penskalaan nilai atribut dari data sehingga bisa jatuh pada range tersebut. Ada beberapa metode yang digunakan untuk proses normalisasi, yaitu:

1. Min-Max
2. Z-Score
3. Decimal Scaling
4. Sigmoidal
5. Dll

Min-Max

Metode min-max merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli.

Rumus:

Newdata = (data-min) * (newmax-newmin) / (max-min) + newmin

Newdata = Data hasil normalisasi
Min = Nilai minimum dari data per kolom
Max = Nilai maximum dari data per kolom
Newmin = adalah batas minimum yang kita berikan
Newmax = adalah batas maximum yang kita berikan

Z-Score

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

Rumus

newdata = (data-mean) / std

newdata = Data hasil normalisasi

Mean = Nilai rata-rata dari data per kolom

std = Nilai dari standard deviasi

Decimal Scaling

Metode Decimal Scaling merupakan metode normalisasi dengan menggerakkan nilai desimal dari data ke arah yang diinginkan.

Rumus

newdata = data / 10^i

newdata = Data hasil normalisasi

i = Adalah nilai scaling yang kita inginkan

Sigmoidal

Sigmoidal merupakan metode normalization melakukan normalisasi data secara nonlinier ke dalam range -1 - 1 dengan menggunakan fungsi sigmoid.

Rumus

newdata = (1-e^(-x))/(1+ e^(-x))

x = (data-mean)/std

e = nilai eksponensial (2,718281828)

Metode ini sangat berguna pada saat data-data yang ada melibatkan data-data *outlier*. Data *outlier* data yang keluar jauh dari jangkauan data lainnya.

Seperti bahasa pemrograman pada umumnya, R memiliki nlai-nilai khusus yang merepresentasikan pengecualian-pengecualian untuk tipe data normal lainnya. Nilai tersebut yaitu:

- NA, *not available.*

  NA biasanya digunakan untuk menggantikan nilai *missing.* Pada R, operasi dasar yang ada dapat memproses dataset yang berisikan nilai NA. Perintah dibawah ini menjelaskan cara mengembalikan nilai NA sebagai hasil dari suatu operasi walaupun *input* dari *argument* tersebut tidak terdapat NA.

```
> NA + 1
[1] NA
> sum(c(NA, 1, 2))
[1] NA
> median(c(NA,1,2,3), na.rm = TRUE)
[1] 2
> length(c(NA, 2, 3, 4))
[1] 4
> 3==NA
[1] NA
> NA==NA
[1] NA
> TRUE | NA
[1] TRUE
```

- NULL

  Berarti nilai yang kosong dan memiliki panjang 0.

```
> length(c(1,2,NULL,4))
[1] 3
> sum(c(1,2,NULL, 4))
[1] 7
> x <- NULL
> c(x,2)
[1] 2
```

1. Eksplorasi data

   Dalam R terdapat beberapa cara untuk melakukan eksplorasi data, yaitu dengan mengetahui **tipe data dari setiap atribut** dan mengetahui **persebaran data setiap atribut.**

```
> data<-airquality
> str(data)
'data.frame':  153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6
...
```

```
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...

> summary(data)
     Ozone           Solar.R           Wind             Temp
Month
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :5
6.00   Min.   :5.000
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:7
2.00   1st Qu.:6.000
 Median : 31.50   Median :205.0   Median : 9.700   Median :7
9.00   Median :7.000
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :7
7.88   Mean   :6.993
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:8
5.00   3rd Qu.:8.000
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :9
7.00   Max.   :9.000
 NA's   :37       NA's   :7
      Day
 Min.   : 1.0
 1st Qu.: 8.0
 Median :16.0
 Mean   :15.8
 3rd Qu.:23.0
 Max.   :31.0

> View(data)
```

Untuk mengetahui jumlah data yang missing, dapat dilakukan dengan cara berikut:

a. Install paket **mice**

b. Gunakan fungsi md.pattern(dataset)

```
> library(mice)
> data <- airquality
> md.pattern(airquality)
    Wind Temp Month Day Solar.R Ozone
111    1    1     1   1       1     1  0
 35    1    1     1   1       1     0  1
  5    1    1     1   1       0     1  1
  2    1    1     1   1       0     0  2
       0    0     0   0       7    37 44
```

Dari hasil diatas dapat dilihat distribusi dari nilai *missing* disetiap atribut.

2. Pembersihan data

Proses mendeteksi dan mengoreksi (menghapus) *record* yang tidak akurat dari *set record*, tabel atau database yang tidak komplit, *incorrect*, *inaccurate* kemudian menggantikan, memodifikasi atau menghapus data tersebut.

```
> data <- airquality
> data$Solar.R[is.na(data$Solar.R)] <- mean(data$Solar.R, na
.rm = TRUE)
> md.pattern(data)
    Solar.R Wind Temp Month Day Ozone
116       1    1    1     1   1     1  0
 37       1    1    1     1   1     0  1
          0    0    0     0   0    37 37
```

Untuk atribut dengan tipe data kategorikal dapat menggunakan fungsi modus, yaitu:

names(sort(-table(x))[1]