

Clustering

Suprayogi

Cluster

Suatu cluster merupakan sekelompok entitas yang memiliki kesamaan dan memiliki perbedaan dengan entitas dari kelompok lain (Everitt, 1980)

Aplikasi Teknik Clustering

- Teknik
Digunakan dalam bidang *biometric recognition & speech recognition*, analisa sinyal radar, *Information Compression*, dan *noise removal*
- Ilmu Komputer
Web mining, analisa database spatial, *information retrieval*, textual document collection, dan image segmentation
- Medis
Digunakan dalam mendefinisikan taxonomi dalam bidang biologi, identifikasi fungsi protein dan gen, diagnosa penyakit dan penanganannya
- Sosial
Digunakan pada analisa pola perilaku, identifikasi hubungan diantara budaya yang berbeda, pembentukan sejarah evolusi bahasa, dan studi psikologi kriminal.
- Ekonomi
Penerapan pada pengenalan pola pembelian & karakteristik konsumen, pengelompokan perusahaan, analisa trend stok

Pendahuluan

- Salah satu aktifitas analisis data adalah klasifikasi atau pengelompokan data ke dalam beberapa kategori/*cluster*. Obyek-obyek/data yang dikelompokkan ke dalam suatu group memiliki ciri-ciri yang sama berdasarkan kriteria tertentu

Algoritma Clustering

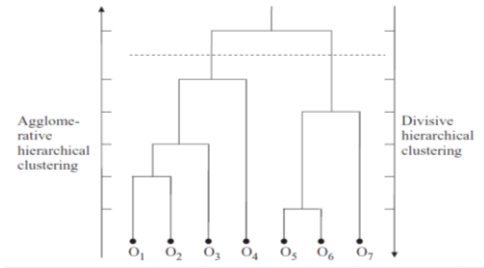
- Algoritma Clustering bekerja dengan mengelompokkan obyek-obyek data (pola, entitas, kejadian, unit, hasil observasi) ke dalam sejumlah *cluster* tertentu (Xu and Wunsch, 2009).
- Dengan kata lain algoritma Clustering melakukan pemisahan/ pemecahan/ segmentasi data ke dalam sejumlah kelompok (*cluster*) menurut karakteristik tertentu.

Jenis-jenis Clustering

Menurut:

- **Struktur kelompok** (*hierarchical* dan *partitioning*)
- **Keanggotaan data dalam kelompok** (eksklusif dan tumpang tindih)
- **Kekompakan data dalam kelompok** (komplet dan parsial)

Gambar Hierarchical clustering sumber (Xu & Wunsch, 2009)



Algoritma K-Means

- Dalam machine-learning dan statistik K-Means merupakan metode analisis kelompok yang mengarah pada pembagian N obyek pengamatan ke dalam K kelompok (cluster).
- Setiap obyek dimiliki oleh sebuah kelompok dan metode ini mencoba untuk menemukan pusat dari kelompok (centroid) dalam data sebanyak iterasi perbaikan yang dilakukan.

Algoritma K-Means

1. Tentukan jumlah cluster
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat cluster (centroid) menggunakan mean utk masing-masing kelompok
4. Alokasikan masing-masing data ke centroid terdekat
5. Kembali ke langkah 3 jika masih ada data yang berpindah cluster, **atau** jika nilai centroid diatas nilai ambang, **atau** jika nilai pada fungsi obyektif yang digunakan masih diatas ambang

Jarak antar data dengan centroid

- Euclidean

$$\sqrt{\sum_{j=1}^p |X_{2j} - X_{1j}|^2}$$

- Manhattan

$$\sum_{j=1}^p |X_{2j} - X_{1j}|$$

- Minkowsky

$$\lambda \sqrt[\lambda]{\sum_{j=1}^p |X_{2j} - X_{1j}|^\lambda}$$

Pengalokasian data ke dalam cluster

$$a_{ij} = \begin{cases} 1 & d = \min\{D(X_i, C_k)\} \\ 0 & \text{lainnya} \end{cases}$$

a_{ij} adalah nilai keanggotaan titik X_i ke centroid C_k , d adalah jarak terpendek dari data X_i ke k kelompok setelah dibandingkan, dan C_1 adalah centroid ke-1.

Studi Kasus Clustering dengan algoritma K-Means

- BPR ABC memiliki data nasabah yang pernah memperoleh kredit, data berupa jumlah rumah dan mobil yang dimiliki pelanggan

Nasabah	Jumlah Rumah	Jumlah Mobil
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

Hasil yang diharapkan

Mengelompokkan nasabah yang memenuhi sifat berikut:

- Nasabah yang jumlah rumah dan mobilnya hampir sama akan berada pada kelompok nasabah yang sama.
- Nasabah yang jumlah rumah dan mobilnya cukup berbeda akan berada pada kelompok nasabah yang berbeda.

Algoritma K-Means

- Langkah 1:** Tentukan jumlah cluster yang diinginkan (misal:k=3)
- Langkah 2:** Pilih centroid awal secara acak : Pada langkah ini secara acak akan dipilih 3 buah data sebagai centroid, misalnya: data {B,E,F}

$$M1=(3,3) ,M2=(1,2),M3=(4,2)$$

3. Langkah 3: Hitung jarak dengan centroid

Data nasabah A : (1,3)
centroid M1: (3,3), centroid M1: (1,2), centroid M3: (4,2)

Nasabah	Jarak ke centroid cluster1	Jarak ke centroid cluster2	Jarak ke centroid cluster3	Jarak terdekat
A	2	1	3.162	C2
B	0	2.236	1.414	C1
C	1	3.162	1	C3
D	2	4.123	1.414	C3
E	2.236	0	3	C2
F	1.414	3	0	C3
G	2.828	1	3.162	C2
H	2.236	1.414	2.236	C2

Keanggotaan nasabah:
Cluster 1 = {B},cluster 2={A,E,G,H},cluster 3= {C,D,F}

Rasio

Between Cluster Variation/Within Cluster Variation

$$\begin{aligned} \text{centroid } M1 &= (3,3) ,M2=(1,2),M3=(4,2) \\ d(m1,m2) &= \sqrt{(3-1)^2 + (3-2)^2} = 2.236 \\ d(m1,m3) &= \sqrt{(3-4)^2 + (3-2)^2} = 1.414 \\ d(m2,m3) &= \sqrt{(1-4)^2 + (2-2)^2} = 3 \end{aligned}$$

nasabah	Jarak ke centroid terdekat
A	1
B	0
C	1
D	1.414
E	0
F	0
G	1
H	1.414

$$BCV=d(m1,m2)+d(m1,m3)+d(m2,m3) = 2.236+1.414+3 = 6,650$$

$$WCV=1^2+0^2+1^2+1.414^2+0^2+0^2+1^2+1.414^2=7$$

$$\text{Rasio} = BCV/WCV = 6.650 / 7 = 0.950$$

lanjutkan ke langkah berikutnya

4. Langkah 4: Pembaruan centroid

Cluster 1		
Nasabah	Jml Rumah	Jml Mobil
B	3	3
Mean	3	3
Cluster 2		
Nasabah	Jml Rumah	Jml Mobil
A	1	3
E	1	2
G	1	1
H	2	1
Mean	1.25	1.75
Cluster 3		
Nasabah	Jml Rumah	Jml Mobil
C	4	3
D	5	3
F	4	2
Mean	4.33	2.67

$$m1=(3,3),m2=(1.25,1.75),m3=(4.33,2.67)$$

5. Langkah 3: Kembali kelangkah 3 – iterasi 2

- Cluster 1 = {B},cluster 2={A,E,G,H},cluster 3= {C,D,F}

Nasabah	Jarak ke centroid cluster1	Jarak ke centroid cluster2	Jarak ke centroid cluster3	Jarak terdekat
A	2	1.275	3.380	C2
B	0	1.768	1.374	C1
C	1	3.021	0.471	C3
D	2	3.953	0.745	C3
E	2.236	0.354	3.399	C2
F	1.414	2.813	0.745	C3
G	2.828	0.791	3.727	C2
H	2.236	1.061	2.867	C2

$$BCV=d(m1,m2)+d(m1,m3)+d(m2,m3) = 6,741$$

$$WCV=1.275^2+0^2+0.471^2+0.745^2+0.354^2+0.745^2+0.791^2+1.061^2=4.833$$

$$\text{Rasio} = BCV/WCV = 6.741 / 4.833 = 1.394$$

Krn 1.394>0.950 maka lanjutkan

6. Langkah ke 4 – iterasi 3 (pembaruan centroid)

$m_1=(3,3), m_2=(1.25,1.75), m_3=(4.33,2.67)$

Cluster 1		
Nasabah	Jml Rumah	Jml Mobil
B	3	3
Mean	3	3

Cluster 2		
Nasabah	Jml Rumah	Jml Mobil
A	1	3
E	1	2
G	1	1
H	2	1
Mean	1.25	1.75

Cluster 3		
Nasabah	Jml Rumah	Jml Mobil
C	4	3
D	5	3
F	4	2
Mean	4.33	2.67

7. Langkah ketiga iterasi-3

Cluster 1 = {B}, cluster 2 = {A, E, G, H}, cluster 3 = {C, D, F}

$BCV = d(m_1, m_2) + d(m_1, m_3) + d(m_2, m_3) = 6,741$

$WCV = 1.275^2 + 0^2 + 0.471^2 + 0.745^2 + 0.354^2 + 0.745^2 + 0.791^2 + 1.061^2 = 4.833$

Sehingga Besar Rasio = $BCV/WCV = 6.741 / 4.833 = 1.394$

Nasabah	Jarak ke centroid cluster1	Jarak ke centroid cluster2	Jarak ke centroid cluster3	Jarak terdekat
A	2	1.275	3.350	C2
B	0	1.768	1.374	C1
C	1	3.021	0.471	C3
D	2	3.953	0.745	C3
E	2.236	0.354	3.399	C2
F	1.414	2.813	0.745	C3
G	2.828	0.791	3.727	C2
H	2.236	1.061	2.867	C2

Krn 1.394 <= 1.394 pd iterasi sbmnya maka selesai

Hasil Akhir

- cluster 1 = {B}
- cluster 2 = {A, E, G, H}
- cluster 3 = {C, D, F}

Algoritma Clustering Lainnya

- Algoritma K-Means merupakan bagian dari algoritma partitioning clustering, algoritma partitional clustering yang lain diantaranya: Mixture-Based Density, Graph Theory-Based Clustering, Fuzzy Clustering.
- Sementara Metode Clustering yang lain selain partitional diantaranya: Hierarchical Clustering, Neural Network-Based Clustering, Kernel-based Clustering, dan Sequential Data Clustering (Xu and Wunsch, 2009)

BUNGA IRIS

ID	a1	a2	a3	a4
id_1	5.1	3.5	1.4	0.2
id_2	4.9	3.0	1.4	0.2
id_3	4.7	3.2	1.3	0.2
id_4	4.6	3.1	1.5	0.2
id_5	5.0	3.6	1.4	0.2
id_6	5.4	3.9	1.7	0.4
id_7	4.6	3.4	1.4	0.3
id_8	5.0	3.4	1.5	0.2
id_9	4.4	2.9	1.4	0.2