Terapan Data Mining
Klasterisasi Metode K-Means

Nurhayati, Ph.D.
Aryajaya Alamsyah, S.Kom

Konsep Dasar Klasterisasi.

- Klasterisasi adalah salah satu teknik data mining untuk mengelompokan data tanpa kelas atau label yang ditentukan.
- Klasterisasi bertujuan untuk mengelompokan data-data dengan karakteristik yang mirip ke dalam satu klaster dan data dengan karakteristik yang berbeda ke klaster yang lain.
- Klasterisasi yang baik adalah memaksimalkan kesamaan antar data dalam satu klaster dan meminimalkan kesamaan antar klaster.
- Klasterisasi terbagi menjadi 3 pendekatan yaitu partisi, hirarki dan kepadatan.

Tabel 1. Metode Klasterisasi Berdasarkan Pendekatan

Pendekatan	Metode Klasterisasi			
Partisi	K-Means	K-Medoids	Fuzzy C-Means (FCM)	Dll.
Hirarkis	Birch	Chameleon	Hirarkis Claster	Dll.
Kepadatan	DBSCAN	Optics	Denclue	Dll.

*Metode K-Means yang akan dibahas lebih detail.

- Secara umum langkah-langkah dalam analisa klaster.
  1. **Seleksi Fitur**. Jika suatu dataset memiliki banyak sekali variabel, maka langkah awal adalah pemilihan variabel yang layak digunakan untuk analisis lebih lanjut. Penjelasan lebih detail terdapat di pertemuan 4 (pra-proses data mining).
  2. **Normalisasi Data**. Jika pengukuran setiap variabel berbeda satuan, maka dilakukan normalisasi data sehingga setiap variabel dapat dibandingkan. Penjelasan lebih detail terdapat di pertemuan 4 (pra-proses data mining).
  3. **Jarak antar objek data**. Dasar analisa klaster adalah mencari kemiripan. Menentukan tingkat kemiripan digunakan suatu ukuran kemiripan yang berdasarkan jarak antar data berdasarkan variabel yang ada. Terdapat beberapa rumus untuk mengukur jarak antar data seperti Euclidean Distance, Manhattan Distance, Minkowski Distance, dll.
  4. **Proses Klaster**. Pada contoh ini akan menggunakan **K-Means**.
  5. **Evaluasi Klaster**. Evaluasi klaster meliputi tiga parameter yaitu penilaian tendensi klaster, penentuan jumlah klaster, pengukuran kualitas klaster

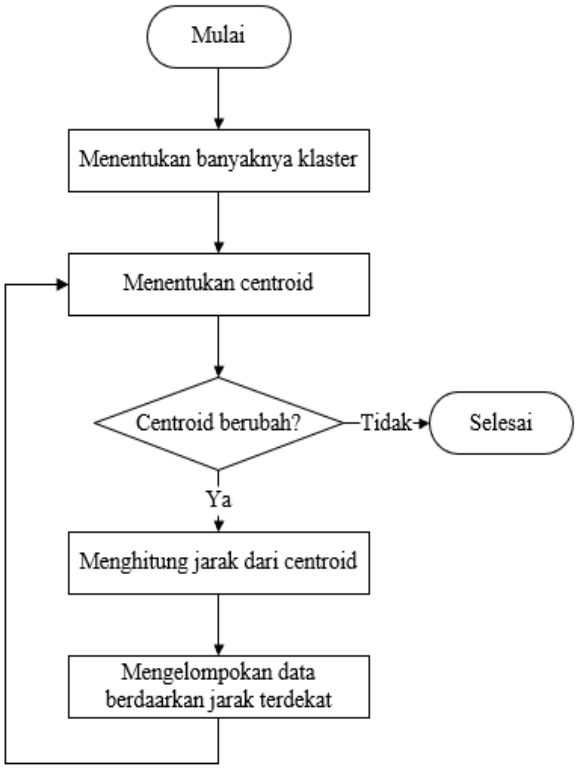
Perbedaan Klasifikasi dan Klasterisasi

Tabel 2. Perbedaan klasifisikasi dengan klasterisasi

Klasifikasi	Klasterisasi
Data berlabel	Data tidak memiliki label
Jumlah grup (class) sudah pasti (diketahui sejak awal)	Jumah grup (class) belum pasti (tidak diketahui sejak awal)
Tujuan: membentuk class berdasarkan feature/variabel yang disediakan	Tujuan: membentuk class berdasarkan pola kemiripan (pettern similaritas) antar data
Algoritma: Decision Tree, Naïve Bayes, SVM, K-NN, dll.	Algoritma: K-Means, K-Medoid, FCM, Chameleon, DBSCAN, dll

Algoritma K-Means

Algoritma K-means merupakan model partisi. Model partisi adalah model yang menggunakan *centroid* untuk membuat *cluster*. *Centroid* adalah "titik tengah" suatu *cluster*. *Centroid* berupa nilai yang bisa ditetukan secara random atau menggunakan rumus tersendiri seperti elbow. Suatu objek data termasuk dalam suatu *cluster* jika memiliki jarak terpendek terhadap *centroid cluster* tersebut.



Secara umum algoritma K-means adalah :

1. Menentukan banyaknya cluster (k).
2. Menentukan centroid.
3. Apakah centroid-nya berubah?
   a. Jika ya, hitung jarak data dari centroid.
   b. Jikat tidak, selesai.
4. Mengelompokkan data berdasarkan jarak terdekat.

Contoh perhitungan algoritma K-Means.

Tabel 3. Contoh dataset

N	a	b
1	1	1
2	2	1
3	4	3
4	5	4

Langkah 1 Tentukan nilai k

Tentukan banyaknya cluster adalah dua (k = 2) yang akan dibuat. Banyaknya cluster harus lebih kecil dari pada banyaknya data (k < n).

Langkah 2 Tentukan centroid setiap cluster.

Untuk menentukan *centroid* awal (*initial centroid*) banyak metode yang dapat digunakan. Di sini metode yang digunakan adalah mengambil data dari data sumber, secara acak atau random (Sel yang berwarna kuning dan hijau di Tabel 3).

Tabel 4. Centroid pada iterasi ke-0

	a	b
Claster 1	1	1
Claster 2	2	1

Untuk pengulangan berikutnya (pengulangan ke-1 sampai selesai), centroid baru dihitung dengan menghitung nilai rata-rata data pada setiap cluster. Jika centroid baru berbeda dengan centroid sebelumnya, maka proses dilanjutkan ke langkah berikutnya. Namun jika centroid yang baru dihitung sama dengan centroid sebelumnya, maka proses clustering selesai.

Langkah 3. Hitung jarak data dengan centroid. Rumus-rumus untuk menghitung jarak yaitu:

1. Euclidean
2. Manhattan
3. Minkowski.

Rumus yang digunakan di sini adalah rumus *Euclidean Distance*:

$$d(x_i, c_i) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2}$$

Dimana

d = jarak

c = centroid

x = data

j = banyaknya data

Jarak data dengan *cluster* 1 adalah :

$$d(x_1, c_1) = \sqrt{(a_1 - c_{1a})^2 + (b_1 - c_{1b})^2} = \sqrt{(1-1)^2 + (1-1)^2} = 0$$

$$d(x_2, c_1) = \sqrt{(a_2 - c_{1a})^2 + (b_2 - c_{1b})^2} = \sqrt{(2-1)^2 + (1-1)^2} = 1$$

$$d(x_3, c_1) = \sqrt{(a_3 - c_{1a})^2 + (b_3 - c_{1b})^2} = \sqrt{(4-1)^2 + (3-1)^2} = 3{,}605551$$

$$d(x_4, c_1) = \sqrt{(a_4 - c_{1a})^2 + (b_4 - c_{1b})^2} = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

Jarak data dengan *cluster* 2 adalah :

$$d(x_1, c_2) = \sqrt{(a_1 - c_{2a})^2 + (b_1 - c_{2b})^2} = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

$$d(x_2, c_2) = \sqrt{(a_2 - c_{2a})^2 + (b_2 - c_{2b})^2} = \sqrt{(2-2)^2 + (1-1)^2} = 0$$

$$d(x_3, c_2) = \sqrt{(a_3 - c_{2a})^2 + (b_3 - c_{2b})^2} = \sqrt{(4-2)^2 + (3-1)^2} = 2{,}828427$$

$$d(x_4, c_2) = \sqrt{(a_4 - c_{2a})^2 + (b_4 - c_{2b})^2} = \sqrt{(5-2)^2 + (4-1)^2} = 4{,}242641$$

Tabel 5. Hasil menghitung jarak data dengan centroid

n	a	b	dc1	dc2
1	1	1	0	1
2	2	1	1	0
3	4	3	3.605.551	2.828.427
4	5	4	5	4.242.641

Langkah 4

Kelompokkan data sesuai dengan cluster-nya, yaitu data yang memiliki jarak terpendek.

1. d(x1, c1) < d(x1, c2) maka x1 masuk ke dalam *cluster* 1.
2. Pada Tabel 5, data n = 1 masuk ke dalam *cluster* 1 karena dc1 < dc2, sedangkan
3. data n = 2, 3, 4 masuk ke dalam *cluster* 2 karena dc2 < dc1.

Tabel 5. Hasil menghitung jarak data dengan centroid dan pengelompokan data

n	A	b	dc1	dc2	c1	c2
1	1	1	0	1	Ok	
2	2	1	1	0		Ok
3	4	3	3.605.551	2.828.427		Ok
4	5	4	5	4.242.641		Ok

Langkah 5 Proses kembali lagi ke langkah 2

Untuk hasil clustering yang lebih lengkap, berikut tabel-tabel hasil analisis dan perhitungan dari awal sampai selesai :

Inisialisasi awal (dataset awal)

n	a	b
1	1	1
2	2	1
3	4	3
4	5	4

Tentukan centroid

	a	b
c1	1	1
c2	2	1

Hitung jarak antar data dan kelompokan data

n	a	b	dc1	dc2	c1	c2
1	1	1	0	1	Ok	
2	2	1	1	0		Ok
3	4	3	3.605.551	2.828.427		Ok
4	5	4	5	4.242.641		Ok

Iterasi 1 (centroid berubah)

Nilai pada sel diperoleh dari menghitung rata-rata pada tabel di atasnya sesuai dengan warna sel

	a	b
c1	1	1
c2	3.666.667	2.666.667

Hitung jarak antar data dan kelompokan data

n	a	b	dc1	dc2	c1	c2
1	1	1	0	314.466	Ok	
2	2	1	1	2.357.023	Ok	
3	4	3	3.605.551	0.471405		Ok
4	5	4	5	1.885.618		Ok

Iterasi 2 (centroid berubah)

Nilai pada sel diperoleh dari menghitung rata-rata pada tabel di atasnya sesuai dengan warna sel

	a	b
c1	1.5	1
c2	4.5	3.5

Hitung jarak antar data dan kelompokan data

N	a	b	dc1	dc2	c1	c2
1	1	1	0.5	4.301.163	Ok	
2	2	1	0.5	2.357.023	Ok	
3	4	3	3.201.562	0.471405		Ok
4	5	4	4.609.772	1.885.618		Ok

Iterasi 3 (centroid tetap) sehingga proses K-Means Selesai

	a	b
c1	1.5	1
c2	4.5	3.5