Modul 7: Analisis Cluster

Tujuan Pembelajaran : Untuk mengelompokkan objek berdasarkan kesamaan sifat antar anggota kelompok yang sama dan heterogen di antara kelompok yang satu dengan lainnya

A. Latar Belakang

Analisis clustering merupakan analisis yang memisahkan data ke dalam sejumlah kelompok menurut karakterikstik masing-masing. Data yang mempunyai karakteristik yang sama akan membentuk satu cluster, sedangkan data yang memiliki karakteristik yang berbeda akan terpisah dalam cluster yang berbeda. Clustering sering disebut sebagai pembelajaran tidak terbimbing (unsupervised learning), karena dalam clustering tidak diperlukan label kelas untuk setiap data yang akan diproses. Sehingga sangat cocok untuk melakukan clustering pada data yang label kelasnya sulit didapatkan pada saat pembangkitan fitur.

Pada clustering, label kelas untuk setiap data dapat diberikan dengan mengamati hasil cluster yang terbentuk. Unsupervised seperti clustering juga sering digunakan untuk mengeksplorasi dan mengkarakteristikkan set data sebelum melakukan supervised. Hal yang harus dipahami dalam cluster yaitu kemiripan harus didefinisikan berdasarkan pada atribut objek. Difinisi kemiripan dan metode dalam data yang dikelompokkan berbeda tergantung pada algoritma clustering yang digunakan. Algoritma clustering yang berbeda juga cocok digunakan pada jenis set data yang berbeda.

B. Materi dan Prosedur

Algoritma K-Means merupakan algoritma pengelompokkan iteratif yang melakukan partisi set data ke dalam sejumlah K cluster yang sudah ditetapkan di awal. Algoritma K-Means sederhana untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, K-Means menjadi salah satu algoritma yang paling penting dalam bidang data mining (Wu dan Kumar, 2009).

Secara historis, bentuk esensial K-Means ditemukan oleh sejumlah peneliti dari lintas disiplin ilmu. Peneliti yang paling berpengaruh adalah Lloyd (1982), Forgey (1965), Friedman dan Rubin (1967), dan McQueen (1967). Algoritma K-Means berkembang hingga menjadi konteks yang lebih besar sebagai algoritma *hill-climbing*, seperti disampaikan oleh Gray dan Nuhoff (1998).

K-Means mengelompokkan set data r-dimensi, $X = \{x_i \mid i = 1, \ldots, N\}$, dimana $x_i \in \mathcal{R}^d$ yang menyatakan data ke-i sebagai “titik data”. Algoritma K-Means mengelompokkan semua titik data dalam X sehingga setiap titik x_i hanya jatuh dalam satu dari K partisi. Perlu diperhatikan bahwa titik berada dalam cluster yang mana, dilakukan dengan cara memberikan setiap titik sebuah ID cluster. Titik dengan ID cluster yang sama akan berada dalam satu cluster yang sama, sedangkan titik dengan ID cluster yang berbeda akan berada dalam cluster yang berbeda. Untuk menyatakan hal ini, biasanya dilakukan vektor keanggotaan cluster m dengan panjang N, dimana m_i bernilai ID cluster titik x_i.

Parameter yang harus dimasukkan ketika menggunakan algoritma K-Means adalah nilai K. Nilai K yang digunakan didasarkan pada informasi yang diketahui sebelumnya tentang berapa banyak cluster data yang muncul dalam X, berapa banyak cluster yang dibutuhkan untuk penerapannya, atau jenis cluster dicari dengan melakukan percobaan dengan beberapa

nilai K. Dalam K-Means, setiap cluster dari K cluster diwakili oleh titik tunggal dalam $\mathcal{R}^d$. Set representative cluster dinyatakan C = {c_j | j = 1, …, K}. Sejumlah K representative cluster tersebut disebut juga sebagai *cluster means* atau *cluster centroid*. Untuk set data dalam X dikelompokkan berdasarkan konsep kedekatan atau kemiripan. Meskipun konsep yang dimaksud untuk data-data yang berkumpul dalam satu cluster adalah data-data yang mirip, tetapi kuantitas yang digunakan untuk mengukurnya adalah ketidakmiripan. Artinya, data-data dengan ketidakmiripan (jarak) yang kecil/dekat maka lebih besar kemungkinannya untuk bergabung dalam satu cluster yang sama. Metrik yang umum digunakan untuk ketidakmiripannya adalah Euclidean.

Pada saat data sudah dihitung ketidakmiripan terhadap setiap centroid, maka selanjutnya dipilih ketidakmiripan yang paling kecil sebagai cluster yang akan diikuti sebagai relokasi data pada cluster di sebuah iterasi. Relokasi sebuah data dalam cluster yang diikuti dapat dinyatakan dengan nilai keanggotaan a yang bernilai 0 atau 1. Nilai 0 diberikan jika data tidak menjadi anggota sebuah cluster dan 1 jika data tersebut menjadi anggota sebuah cluster. Karena K-Means mengelompokkan secara tegas data hanya pada satu cluster, maka dari nilai a sebuah data pada semua cluster, hanya satu yang bernilai 1, sedangkan lainnya 0 seperti yang dinyatakan oleh persamaan berikut:

$$a_{ij} = \begin{cases} 1, & arg\ \min\{d(x_i, c_j)\} \\ 0, & lainnya \end{cases}$$

$d(x_i, c_j)$ menyatakan ketidakmiripan (jarak) dari data ke-I ke cluster c_j.

Relokasi centroid untuk mendapatkan titik centroid C didapatkan dengan menghitung rata-rata setiap fitur dari semua data yang tergabung dalam setiap cluster yang dinyatakan oleh persamaan berikut:

$$c_j = \frac{1}{Nk} \sum_{i=1}^{Nk} x_{jl}$$

Nk adalah jumlah data yang tergabung dalam sebuah cluster.

Jika diperhatikan dari langkahnya yang selalu memilih cluster terdekat, maka dapat diketahui bahwa K-Means berusaha untuk meminimalkan fungsi objektif seperti yang dinyatakan oleh persamaan berikut:

$$J = \sum_{i=1}^{N} \sum_{l=1}^{K} a_{ic}\, d(x_i, c_l)^2$$

Dengan kata lain, K-Means berusaha untuk meminimalkan total jarak kuadrat di antara setiap titik x_i dan representasi cluster c_j terdekat. Algoritma K-Means disajikan sebagai berikut:

1. Inisialisasi; tentukan nilai K sebagai jumlah cluster yang diinginkan dan metric ketidakmiripan (jarak) yang diinginkan. Jika perlu, tetapkan ambang batas perubahan fungsi objektif dan ambang batas perubahan posisi centroid.
2. Pilih K data dari set data X sebagai centroid.
3. Alokasikan semua data ke centroid terdekat dengan metric jarak yang sudah ditetapkan (memperbarui cluster ID setiap data).
4. Hitung kembali centroid C berdasarkan data yang mengikuti cluster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu (a) perubahan fungsi objektif sudah di bawah ambang batas yang diinginkan; atau (b) tidak ada data yang

berpindah cluster; atau (c) perubahan posisi centroid sudaj di bawah ambang batas yang ditetapkan.

Algoritma K-Means mengelompokkan set data X dalam langkah iteratif. Terdapat dua langkah utama dalam algoritam K-Means,yaitu (1) penentuan kembali ID cluster dari semua titik data dalam X, dan (2) memperbarui representasi cluster (centroid) berdasarkan titik data dalam setiap cluster. Algoritma bekerja sebagai berikut. Pertama, representasi cluster diinisialisasi dengan memilih K data dalam $\mathcal{R}^{\mathrm{d}}$ secara acak. Selanjutnya, secara iteratif melakukan dua langkah berikut sampai tercapai kondisi konvergen.

Langkah 1: (*Data assignment*) Setiap data ditetapkan ke centroid terdekat dengan pemecahan hubungan apa adanya. Hasilnya berupa data yang terpartisi.

Langkah 2: (*Relocation of means*) Setiap representasi cluster direlokasikan ke pusat dengan rata-rata aritmetika dari semua data yang ditetapkan masuk ke dalamnya. Langkah ini didasarkan pada observasi bahwa data memberikan set titik. Representasi tunggal yang terbaik untuk set tersebut adalah rata-rata dari titik data.

Algoritma K-Means mencapai kondisi konvergen ketika pengalokasian kembali titik data tidak lagi berubah. Proses dari iterasi ke iterasi hingga dicapai kondisi konvergen juga dapat diamati dari nilai fungsi objektif yang didapatkan. Pada kondisi yang semakin konvergen dapat diamati bahwa nilai fungsi objektif akan semakin menurun.

Pemilihan K titik data sebagai centroid awal juga mempengaruhi hasil clustering. Sifat ini menjadi karakteristik alami K-Means yang mengakibatkan hasil cluster yang didapat pada percobaan yang berbeda mendapatkan hasil yang berbeda pula. Kondisi seperti ini dikenal dengan solusi yang *local optima*, yang artinya algoritma K-Means sangat sensitif terhadap lokasi awal centroid. Dengan kata lain, inisialisasi set representatif cluster C yang berbeda dapat mengakibatkan hasil cluster yang berbeda, bahkan pada set data X yang sama. Penyelesaian masalah local optima dapat diselesaikan dengan menjalankan algoritma beberapa kali dengan inisial centroid yang berbeda kemudian memilih hasil yang terbaik.

Jika terdapat informasi mengenai set data, seperti jumlah partisi yang secara alami menggambarkan set data, maka informasi tersebut dapat digunakan untuk memilih nilai K yang optimal. Jika tidak ada informasi seperti itu, maka harus menggunakan beberapa kriteria lain untuk memilih K, misalnya dengan mencoba beberapa nilai K berbeda dan memilih clustering yang nilai fungsi objektifnya minimal. Alternatif yang lain adalah secara progresif meningkatkan jumlah cluster dengan gabungan kriteria pemberhentian yang cocok. Bisecting K-Means melakukan hal tersebut dengan meletakkan semua data dalam cluster tunggal dan kemudian secara rekursif memecah cluster paling pada menjadi dua cluster menggunakan 2-means.

C. Studi Kasus

Diketahui terdapat nilai indeks kedalaman kemiskinan dan indeks keparahan kemiskinan. Akan dilakukan clustering pada 2 set data tersebut. Nilai masing-masing data dapat dilihat pada tabel berikut.

Data ke-i	Indeks Kedalaman Kemiskinan (x)	Indeks Keparahan Kemiskinan (y)
1	1	1
2	4	1
3	6	1

4	1	2
5	2	3
6	5	3
7	2	5
8	3	5
9	2	6
10	3	8

Pengukuran jarak yang digunakan adalah jarak Euclidean. Jumlah cluster (K) adalah 3. Ambang batas (T) yang ditetapkan untuk perubahan fungsi objektif adalah 0,1. Langkah-langkah yang dilakukan sebagai berikut.

1. Insisialisasi

Dilakukan pemilihan K data sebagai centroid awal, misalnya dipilih data ke-2, 4, dan 6.

Centroid	x	y
1	4	1
2	1	2
3	5	3

Nilai fungsi objektif yang diberi adalah 1000 karena data belum masuk dalam cluster.

2. Iterasi 1

Menghitung jarak setiap data ke centroid terdekat. Centroid terdekat akan menjadi cluster yang diikuti oleh data tersebut. Berikut beberapa contoh perhitungan jarak ke setiap centroid pada data ke-1.

$$d(x_1, c_1) = \sqrt{\sum_{i=1}^{r}(x_{1i} - c_{1i})^2} = \sqrt{(1-4)^2 + (1-1)^2} = 3$$

$$d(x_1, c_2) = \sqrt{\sum_{i=1}^{r}(x_{1i} - c_{2i})^2} = \sqrt{(1-1)^2 + (1-2)^2} = 1$$

$$d(x_1, c_3) = \sqrt{\sum_{i=1}^{r}(x_{1i} - c_{3i})^2} = \sqrt{(1-5)^2 + (1-3)^2} = 4{,}4721$$

Data ke-i	Jarak ke centroid			Terdekat	Cluster yang diikuti
	1	2	3		
1	3,0000	1,0000	4,4721	1,0000	2
2	0,0000	3,1623	2,2361	0,0000	1
3	2,0000	5,0990	2,2361	2,0000	1
4	3,1623	0,0000	4,1231	0,0000	2
5	2,8284	1,4142	3,0000	1,4142	2
6	2,2361	4,1231	0,0000	0,0000	3
7	4,4721	3,1623	3,6056	3,1623	2
8	4,1231	3,6056	2,8284	2,8284	3
9	5,3852	4,1231	4,2426	4,1231	2
10	7,0711	6,3246	5,3852	5,3852	3

Selanjutnya dihitung centroid yang baru untuk setiap cluster berdasarkan data yang bergabung pada setiap clusternya. Untuk cluster 1, terdapat 2 data yang tergabung ke dalamnya.

Data anggota	Fitur x	Fitur y
2	4	1
3	6	1
Nk	Jumlah x	Jumlah y
2	10	2
Rata-rata	5,0000	1,0000

Untuk cluster 2, terdapat 5 data yang tergabung ke dalamnya.

Data anggota	Fitur x	Fitur y
1	1	1
4	1	2
5	2	3
7	2	5
9	2	6
Nk	Jumlah x	Jumlah y
5	8	17
Rata-rata	1,6000	3,4000

Untuk cluster 3, terdapat 3 data yang tergabung ke dalamnya

Data anggota	Fitur x	Fitur y
6	5	3
8	3	5
10	3	8
Nk	Jumlah x	Jumlah y
3	11	16
Rata-rata	3,6667	5,3333

Rata-rata dari 3 cluster tersebut ditunjukkan dalam tabel berikut.

Centroid	x	Y
1	5,0000	1,0000
2	1,6000	3,4000
3	3,6667	5,3333

Nilai fungsi objektif yang didapatkan dari Euclidean kuadrat antara setiap data dengan centroid dari cluster yang diikuti.

Data ke-i	Cluster 1	Cluster 2	Cluster 3
1	0	6,1200	0
2	1,0000	0	0
3	1,0000	0	0
4	0	2,3200	0
5	0	0,3200	0
6	0	0	7,2222
7	0	2,7200	0
8	0	0	0,5556
9	0	6,9200	0
10	0	0	7,5556

Didapatkan nilai fungsi objektif J = 35,7333

Perubahan fungsi objektif didapat = 1000 – 35,7333 = 964,2667

Perubahan fungsi objektif masih di atas ambang batas yang ditetapkan, sehingga proses dilanjutkan ke iterasi berikutnya.

3. Iterasi 2

Menghitung jarak setiap data ke centroid terdekat. Centroid terdekat akan menjadi cluster yang diikuti oleh data tersebut.

Data ke-i	Jarak ke centroid			Terdekat	Cluster yang diikuti
	1	2	3		
1	4,0000	2,4739	2,4739	2,4739	2

2	1,0000	3,3941	1,0000	1,0000	1
3	1,0000	5,0120	1,0000	1,0000	1
4	4,1231	1,5232	1,5232	1,5232	2
5	3,6056	0,5657	0,5657	0,5657	2
6	2,0000	3,4234	2,0000	2,0000	1
7	5,0000	1,6492	1,6492	1,6492	2
8	4,4721	2,1260	0,7454	0,7454	3
9	5,8310	2,6306	1,7951	1,7951	3
10	7,2801	4,8083	2,7487	2,7487	3

Selanjutnya dihitung centroid yang baru untuk setiap cluster berdasarkan data yang bergabung pada setiap clusternya. Untuk cluster 1, terdapat 3 data yang tergabung ke dalamnya.

Data anggota	**Fitur x**	**Fitur y**
2	4	1
3	6	1
6	5	3
Nk	Jumlah x	Jumlah y
3	15	5
Rata-rata	5,0000	1,6667

Untuk cluster 2, terdapat 4 data yang tergabung ke dalamnya.

Data anggota	**Fitur x**	**Fitur y**
1	1	1
4	1	2
5	2	3
7	2	5
Nk	Jumlah x	Jumlah y
4	6	11
Rata-rata	1,5000	2,7500

Untuk cluster 3, terdapat 3 data yang tergabung ke dalamnya

Data anggota	**Fitur x**	**Fitur y**
8	3	5
9	2	6
10	3	8
Nk	Jumlah x	Jumlah y
3	8	19
Rata-rata	2,6667	6,3333

Rata-rata dari 3 cluster tersebut ditunjukkan dalam tabel berikut.

Centroid	**x**	**Y**
1	5,0000	1,6667
2	1,5000	2,7500
3	2,6667	6,3333

Nilai fungsi objektif yang didapatkan dari Euclidean kuadrat antara setiap data dengan centroid dari cluster yang diikuti.

Data ke-i	**Cluster 1**	**Cluster 2**	**Cluster 3**
1	0	3,3125	0
2	1,4444	0	0

3	1,4444	0	0
4	0	0,8125	0
5	0	0,3125	0
6	1,7778	0	0
7	0	5,3125	0
8	0	0	1,8889
9	0	0	0,5556
10	0	0	2,8889

Didapatkan nilai fungsi objektif J = 19,7500

Perubahan fungsi objektif didapat = 35,7333 – 19,7500 = 15,9833

Perubahan fungsi objektif masih di atas ambang batas yang ditetapkan, Sehingga proses dilanjutkan ke iterasi berikutnya.

4. Iterasi 3

Menghitung jarak setiap data ke centroid terdekat. Centroid terdekat akan menjadi cluster yang diikuti oleh data tersebut.

Data ke-i	Jarak ke centroid			Terdekat	Cluster yang diikuti
	1	2	3		
1	4,0552	1,8200	5,5877	1,8200	2
2	1,2019	3,0516	5,4975	1,2019	1
3	1,2019	4,8283	6,2893	1,2019	1
4	4,0139	0,9014	4,6428	0,9014	2
5	3,2830	0,5590	3,3993	0,5590	2
6	1,3333	3,5089	4,0689	1,3333	1
7	4,4845	2,3049	1,4907	1,4907	3
8	3,8873	2,7042	1,37444	1,3744	3
9	5,2705	3,2882	0,7454	0,7454	3
10	6,6416	5,4601	1,6997	1,6997	3

Selanjutnya dihitung centroid yang baru untuk setiap cluster berdasarkan data yang bergabung pada setiap clusternya. Untuk cluster 1, terdapat 3 data yang tergabung ke dalamnya.

Data anggota	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
Nk	Jumlah x	Jumlah y
3	15	5
Rata-rata	5,0000	1,6667

Untuk cluster 2, terdapat 3 data yang tergabung ke dalamnya.

Data anggota	Fitur x	Fitur y
1	1	1
4	1	2
5	2	3
Nk	Jumlah x	Jumlah y
3	4	6

Rata-rata	1,3333	2,0000

Untuk cluster 3, terdapat 4 data yang tergabung ke dalamnya

Data anggota	**Fitur x**	**Fitur y**
7	2	5
8	3	5
9	2	6
10	3	8
Nk	Jumlah x	Jumlah y
4	10	24
Rata-rata	2,5000	6,0000

Rata-rata dari 3 cluster tersebut ditunjukkan dalam tabel berikut.

Centroid	**x**	**y**
1	5,0000	1,6667
2	1,3333	2,0000
3	2,5000	6,0000

Nilai fungsi objektif yang didapatkan dari Euclidean kuadrat antara setiap data dengan centroid dari cluster yang diikuti.

Data ke-i	**Cluster 1**	**Cluster 2**	**Cluster 3**
1	0	1,1111	0
2	1,4444	0	0
3	1,4444	0	0
4	0	0,1111	0
5	0	1,4444	0
6	1,7778	0	0
7	0	0	1,2500
8	0	0	1,2500
9	0	0	0,2500
10	0	0	4,2500

Didapatkan nilai fungsi objektif J = 14,3333

Perubahan fungsi objektif didapat = 19,7500 – 14,3333 = 5,4167

Perubahan fungsi objektif masih di atas ambang batas yang ditetapkan, sehingga proses dilanjutkan ke iterasi berikutnya.

5. Iterasi 4

Menghitung jarak setiap data ke centroid terdekat. Centroid terdekat akan menjadi cluster yang diikuti oleh data tersebut.

Data ke-i	**Jarak ke centroid**			**Terdekat**	**Cluster yang diikuti**
	1	**2**	**3**		
1	4,0552	1,0541	5,2202	1,0541	2
2	1,2019	2,8480	5,2202	1,2019	1
3	1,2019	4,7726	6,1033	1,2019	1
4	4,0139	0,3333	4,2720	0,3333	2
5	3,2830	1,2019	3,0414	1,2019	2

Data ke-i	Jarak ke centroid			Terdekat	Cluster yang diikuti
	1	2	3		
6	1,3333	3,8006	3,9051	1,3333	1
7	4,4845	3,0732	1,1180	1,1180	3
8	3,8873	3,4319	1,1180	1,1180	3
9	5,2705	4,0552	0,5000	0,5000	3
10	6,6416	6,2272	2,0616	2,0616	3

Selanjutnya dihitung centroid yang baru untuk setiap cluster berdasarkan data yang bergabung pada setiap clusternya. Untuk cluster 1, terdapat 3 data yang tergabung ke dalamnya.

Data anggota	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
Nk	Jumlah x	Jumlah y
3	15	5
Rata-rata	5,0000	1,6667

Untuk cluster 2, terdapat 3 data yang tergabung ke dalamnya.

Data anggota	Fitur x	Fitur y
1	1	1
4	1	2
5	2	3
Nk	Jumlah x	Jumlah y
3	4	6
Rata-rata	1,3333	2,0000

Untuk cluster 3, terdapat 4 data yang tergabung ke dalamnya

Data anggota	Fitur x	Fitur y
7	2	5
8	3	5
9	2	6
10	3	8
Nk	Jumlah x	Jumlah y
4	10	24
Rata-rata	2,5000	6,0000

Rata-rata dari 3 cluster tersebut ditunjukkan dalam tabel berikut.

Centroid	x	y
1	5,0000	1,6667
2	1,3333	2,0000
3	2,5000	6,0000

Nilai fungsi objektif yang didapatkan dari Euclidean kuadrat antara setiap data dengan centroid dari cluster yang diikuti.

Data ke-i	Cluster 1	Cluster 2	Cluster 3
1	0	1,1111	0
2	1,4444	0	0
3	1,4444	0	0
4	0	0,1111	0
5	0	1,4444	0
6	1,7778	0	0
7	0	0	1,2500
8	0	0	1,2500
9	0	0	0,2500
10	0	0	4,2500

Didapatkan nilai fungsi objektif J = 14,3333
Perubahan fungsi objektif didapat = 14,3333 – 14,3333 = 0,0000
Perubahan fungsi objektif sudah di bawah ambang batas yang ditetapkan, berarti kondisi cluster sudah mencapai konvergen dan proses iterasi pun berhenti.

Diagram proses clustering dari tahap inisialisasi sampai pada tahap iterasi keempat ditunjukkan pada Gambar berikut.

D. Latihan Mandiri

Lakukan analisis cluster untuk mengelompokkan data-data yang tersedia di syam-ok dan berikan interpretasinya. Variabel yang digunakan adalah nama kabupaten/kota, indeks kedalaman kemiskinan, dan indeks keparahan kemiskinan.

E. Rangkuman

Analisis clustering merupakan analisis yang memisahkan data ke dalam sejumlah kelompok menurut karakterikstik masing-masing. Data yang mempunyai karakteristik yang sama akan membentuk satu cluster, sedangkan data yang memiliki karakteristik yang berbeda akan terpisah dalam cluster yang berbeda. Algoritma K-Means mengelompokkan semua titik data dalam X sehingga setiap titik x_i hanya jatuh dalam satu dari K partisi. Parameter yang harus dimasukkan ketika menggunakan algoritma K-Means adalah nilai K. Nilai K yang digunakan didasarkan pada informasi yang diketahui sebelumnya tentang berapa banyak cluster data yang muncul dalam X, berapa banyak cluster yang dibutuhkan untuk penerapannya, atau jenis cluster dicari dengan melakukan percobaan dengan beberapa nilai K.

Algoritma K-Means disajikan sebagai berikut:

1. Inisialisasi; tentukan nilai K sebagai jumlah cluster yang diinginkan dan metric ketidakmiripan (jarak) yang diinginkan. Jika perlu, tetapkan ambang batas perubahan fungsi objektif dan ambang batas perubahan posisi centroid.
2. Pilih K data dari set data X sebagai centroid.
3. Alokasikan semua data ke centroid terdekat dengan metric jarak yang sudah ditetapkan (memperbarui cluster ID setiap data).
4. Hitung kembali centroid C berdasarkan data yang mengikuti cluster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu (a) perubahan fungsi objektif sudah di bawah ambang batas yang diinginkan; atau (b) tidak ada data yang

berpindah cluster; atau (c) perubahan posisi centroid sudaj di bawah ambang batas yang ditetapkan.