Data Mining

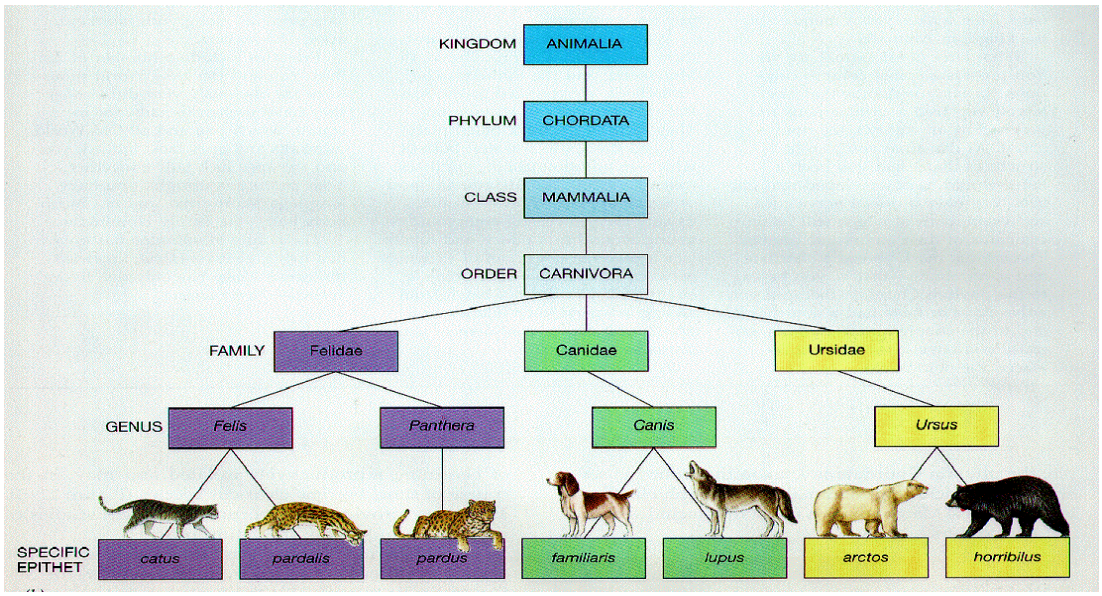
Clustering

Oleh : Suprayogi

Pendahuluan

Saat ini terjadi fenomena yaitu berupa data yang melimpah, setiap hari banyak orang yang berurusan dengan data yang bersumber dari berbagai jenis observasi dan pengukuran. Misalnya data yang menjelaskan karakteristik spesies makhluk hidup, data yang menggambarkan ciri-ciri fenomena alam, data yang berasal dari ringkasan hasil eksperimen ilmu pengetahuan, dan data yang mencatat perfoma suatu mesin. 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 criteria tertentu.

Klasifikasi berperan penting dalam sejarah panjang “human development” . Untuk mempelajari obyek baru atau memahami suatu fenomena baru seseorang selalu mencoba untuk mendeskripsikan fitur-fitur dan lebih jauh lagi membandingkan fitur tersebut dengan obyek-obyek/fenomena yang sudah dikenalnya, berdasarkan pada kesamaan / ketidaksamaan, menyimpulkan/generalisasi, berdasarkan suatu aturan-aturan tertentu. Sebagai contoh semua benda-benda alam pada dasarnya diklasifikasikan ke dalam grup:binatang,tumbuh-tumbuhan,mineral. Menurut taksonomi biologi semua binatang dikelompokkan kedalam kategori kingdom,phylum,class,order,family,genus,species dari yang umum ke spesifik.Dengan demikain terdapat binatang yang bernama tigers,lions, wolves, dogs, horses, sheeps, cats, mice dsb. Sebetulnya penamaan dan klasifikasi memiliki arti yang sama menurut Everitt et al.(2001). Dengan mengetahui informasi tentang klasifikasi tersebut, seseorang dapat menyimpukan sifat-sifat suatu obyek tertentu berdasarkan kategori dimana obyek berasal. Sebagai contoh ketika kita melihat anjing laut didarat kita dapat langsung menyimpulkan bahwa anjing laut tersebut pandai berenang, tanpa kita melihat langsung dia berenang.



taxonomi bidang biologi , sumber http://ykonline.yksd.com

Cluster

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

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.

Aplikasi Teknik Clustering

Clustering telah diterapkan diberbagai bidang seperti di jelaskan sebagai berikut:

1. Teknik

   Digunakan dalam bidang *biometric recognition & speech recognition*, analisa sinyal radar, *Information Compression*,dan *noise removal.*

2. Ilmu Komputer

   Web mining,analisa database spatial,*information retrieval,*textual document collection,dan image segmentation.

3. Medis

   Digunakan dalam mendefinisikan taxonomi dalam bidang biologi, identifikasi fungsi protein dan gen, diagnosa penyakit dan penanganannya.

4. Astronomy

   Digunakan untuk mengelompokkan bintang dan planet, menginvestigasi formasi tanah, mengelompokkan wilayah /kota, digunakan dalam studi tentang sistem pada sungai dan gunung.

5. Sosial

   Digunakan pada analisa pola perilaku,identifikasi hubungan diantara budaya yang berbeda, pembentukan sejarah evolusi bahasa, dan studi psikologi criminal.

6. Ekonomi

   Penerapan pada pengenalan pola pembelian& karakteristik konsumen, pengelompokan perusahaan, analisa trend stok.

Perbedaan dengan klasifikasi.

Pada dasarnya sistem klasifikasi berupa supervised atau unsupervised. Tergantung pada obyek-obyek data baru apakah ditempatkan pada kelas diskrit supervised atau kategori unsupervised. Pada klasifikasi supervised label pada kelas dari setiap data mengikuti fitur/variable penyerta kelas sehingga jika ada data baru yang belum diketahui kelasnya, dengan model yang sudah dibangun kita dapat memprediksi kelas dari data baru tersebut. Dalam klasifikasi un*supervised*(*clustering/segmentation/partitioning*) data yang digunakan tidak memilki label kelas seperti pada klasifikasi supervised, tetapi kemudian dikelompokkan menurut karakteristiknya.

Tujuan pengelompokan

Tujuan *clustering* (pengelompokan) data dapat dibedakan menjadi dua, yaitu pengelompokan untuk **pemahaman** dan clustering untuk **penggunaan** (Prasetyo,2012). Biasanya proses pengelompokan untuk tujuan **pemahaman** hanya sebagai proses awal untuk kemudian dilanjutkan dengan pekerjaan seperti *summarization(*rata-rata,standar deviasi*),* pelabelan kelas untuk setiap kelompok sehingga dapat digunakan sebagai data training dalam klasifikasi supervised. Sementara jika untuk **penggunaan,** tujuan utama *clustering* biasanya adalah mencari prototipe kelompok yang paling representatif terhadap data, memberikan abstraksi dari setiap obyek data dalam kelompok dimana sebuah data terletak didalamnya.

Contoh tujuan clustering untuk **pemahaman** diantaranya: dibidang Biologi (pengelompokan berdasarkan karakter tertentu secara hirarkis) , pengelompokan gen yang memiliki fungsi sama. Dibidang information retrieval (web search),bidang klimatologi (pengelompokam pola tekanan udara yang berpengaruh pada cuaca),bidang bisnis (pengelompokan konsumen yang berpotensi untuk analisa dan strategi pemasaran).

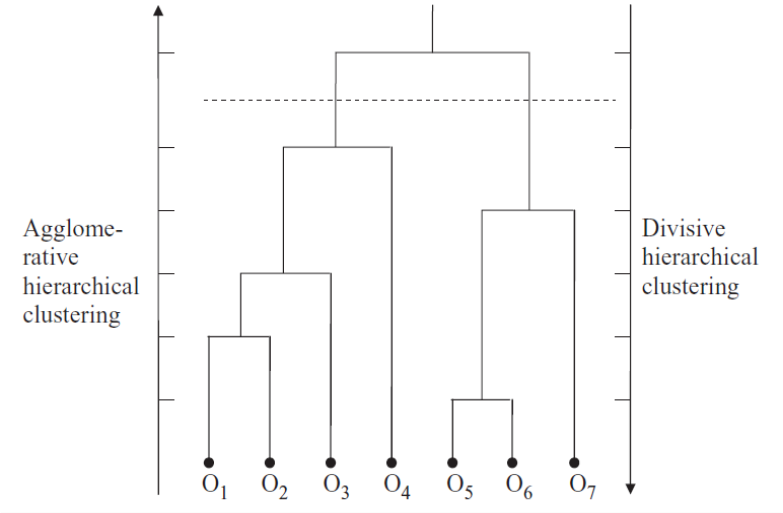
Contoh tujuan clustering untuk **penggunaan** dibidang *summarization,* dengan semakin besarnya jumlah data maka ongkos melakukan peringkasan semakin mahal (berat&kompleks), maka perlu diterapkan pengelompokan data untuk membuat prototipe yang dapat mewakili keseluruhan data yang akan digunakan. Kompresi , data yang terletak dalam satu cluster dapat dikompresi dengan diwakili oleh indeks prototipe yang dikaitkan dengan kelompok ,teknik kompresi ini dikenal sebagai *quantization vector.*

Jenis-jenis pengelompokan

Clustering dapat dibedakan menurut **struktur kelompok ,keanggotaan data dalam kelompok,** dan **kekompakan data dalam kelompok**. Menurut **struktur kelompok** clustering dibagi menjadi dua yaitu *hierarchical* dan *partitioning.*

Hierarchical clustering adalah metode *clustering* yang mengelompokkan data dengan urutan partisi berkalang, metode ini dikelompokkan menjadi dua metode yaitu *agglomerative* dan *divisive*, metode *agglomerative* berawal dari obyek-obyek individual dimana pada awalnya banyaknya cluster sama dengan banyaknya obyek. Pertama-tama obyek-obyek yang paling mirip dikelompokkan, dan kelompok-kelompok awal ini digabungkan sesuai dengan kemiripannya. Akhirnya sewaktu kemiripan berkurang, semua subkelompok digabungkan menjadi satu cluster tunggal. Sementara Metode *Hierarchical divisive*

merupakan proses kebalikan dari agglomerative , keduanya mengorganisasi data ke dalam struktur hirarki berbasis matrix *proximity*, hasil dari dari *Hierarcichal Clustering* digambarkan dalam bentuk *binary tree* ataupun *dendogram*, root merupakan keseluruhan dataset dan tiap cabang merupakan data point, *clustering* akhir dapat diperoleh dari pemotongan dendogram pada level-level yang sesuai.



Gambar Hierarchical clustering sumber (Xu & Wunsch, 2009)

Berbeda dengan klastering hirarki yang menghasilkan suatu tingkatan berurutan klaster dengan cara penggabungan secara iterative atau pemisahan, *partitional clustering* mengelompokkan datapoint kedalam k klaster tanpa struktur hirarki(Xu & Wunsch, 2009), metode ini membagi set data ke dalam sejumlah kelompok yang tidak saling overlap antara satu kelompok dengan kelompok lainnya, artinya setiap data hanya menjadi satu kelompok, termasuk dalam metode ini adalah K-Means dan DBSCAN.

Menurut **keanggotaan data dalam kelompok,** pengelompokan dibagi menjadi dua yaitu ekslusif dan tumpang tindih. Dalam kategori ekslusif sebuah data hanya menjadi anggota satu kelompok saja dan tidak bisa menjadi anggota kelompok lainnya. Metode yang termasuk kategori ini adalah **K-Means** dan **DBSCAN**, sedangkan yang masuk kategori overlap adalah metode clustering yang membolehkan sebuah data menjadi anggota di lebih dari satu kelompok, misalnya Fuzzy C-Means dan *Hierarchical Clustering.*

Sementara menurut kategori kekompokan, clustering terbagi menjadi dua yaitu komplet dan parsial. Jika semua data bisa bergabung menjadi satu (dlm konsep *partitioning*), bisa dikatakan semua data kompak menjadi satu kelompok. Namun jika ada satu atau beberapa data yang tidak ikut bergabung dalam kelompok mayoritas, data tersebut dikatakan memiliki perilaku menyimpang atau dikenal dengan istilah outlier/noise/"uninterested background". Beberapa metode yang dapat mendeteksi outlier ini diantaranya adalah DBSCAN dan K-Means (dengan sejumlah komputasi tambahan).

K-Means

Dalam machine-learning dan statistic K-Means merupakan metode analisis kelompok yang mengarah pada pembagian N obyek pengamatan ke dalam K kelompok (cluster), dimana setiap obyek dimiliki oleh sebuah kelompok dengan *mean* (rata-rata) dan metode ini mencoba untuk menemukan pusat dari kelompok (centroid) dalam data sebanyak iterasi perbaikan yang dilakukan. Metode ini berusaha membagi data kedalam kelompok sehingga data yang berkarakteristik sama dimasukkan ke dalam satu kelompok sementara data yang berkarakteristik berbeda dimasukkan dalam kelompok yang lain. Adapun tujuan dari clustering/pengelompokan data ini adalah meminimalkan fungsi obyektif yang diset dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi didalam suatu kelompok dan memaksimalkan variasi antar kelompok. Clustering menggunakan metode K-Means secara umum dilakukan dengan algoritma sbb:

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**

Pada langkah 3 , lokasi centroid setiap kelompok diambil dari rata-rata semua nilai data pada setiap fiturnya. Jika M menyatakan jumlah data, i menyatakan fitur/variable/atribut ke-i dan p menyatakan dimensi dari data, untuk menghitung centroid fitur ke i digunakan formula:

$C_i = \frac{1}{M}\sum_{j=1}^{M} xj$ ……(1)

Jarak antara data dan centroid diukur dengan beberapa cara diantaranya :

Euclidean

$D(X_2,X_1) = ||\ X_2\text{-}X_1\ ||_2 = \sqrt{\sum_{i=1}^{p} |X_{2j} - X_{1j}|^2}$ .....(2)

D adalah jarak antara data X2 dan X1, dan |.| adalah nilai mutlak.

Pengukuran jarak pada ruang jarak Manhattan menggunakan formula:

$D(X_2,X_1) = ||\ X_2\text{-}X_1\ ||_2 = \sum_{j=1}^{p} |X_{2j} - X_{1j}|$.....(3)

Pengukuran jarak pada ruang jarak Minkowsky:

$D(X_2,X_1) = ||\ X_2\text{-}X_1\ ||\lambda = \lambda \sqrt{\sum_{i=1}^{p} \left| X_{2j} - X_{1j} \right|^{\lambda}}$ .....(4)

Namun demikian cara yang paling banyak digunakan adalah Euclidean dan manhattan. Euclidean menjadi pilihan jika ingin dicari jarak terpendek antara dua titik, sedangkan Manhattan memberikan jarak terjauh anatara dua titik.

Pada langkah 4 pengalokasian kembali data ke dalam masing-masing kelompok kedalam K-Means didasarkan pada perbandingan jarak antara data dengan centroid setiap kelompok yang ada. Pengalokasian ini dapat dirumuskan sbb:

$$a_{ij} = \begin{cases} 1 & d=\min\{D(X_i,C_1)\} \\ 0 & \text{lainnya} \end{cases} \quad \text{..... (5)}$$

a_{ij} adalah nilai keanggotaan titik X_i ke centroid C_1, d adalah jarak terpendek dari data Xi ke k kelompok setelah dibandingkan, dan C1 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.

Tabel Data nasabah

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

Clustering yang diharapkan mampu menghasilkan kelompok nasabah yang memenuhi sifat berikut:

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

   Berikut langkah-langkah clustering menggunakan algoritma K-Means.

1. **Langkah 1**: Tentukan jumlah cluster yang diinginkan (misl:k=3)
2. **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 …… (iterasi 1)

Pada langkah ini setiap data akan ditentukan centroid terdekatnya, dan data tersebut akan ditetapkan sebagai anggota kelompok yang terdekat dengan centroid.

Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sbb:

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

DM1 =$\sqrt{(1-3)^2+(3-3)^2}$ = 2

DM2 =$\sqrt{(1-1)^2+(3-2)^2}$ = 1

DM3 =$\sqrt{(1-4)^2+(3-2)^2}$ = 3.162

Tabel hasil perhitungan jarak antara masing-masing data dengan centroid

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

Dari tabel diatas didapatkan keanggotaan nasabah sbb:

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

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*) :

Karena 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

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

Dalam hal ini d(mi,mj) menyatakan jarak Euclidean dari m ke mj

Menghitung WCV

Yaitu dengan memilih jarak terkecil antara data dengan centroid pada masing-masing cluster:

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

$WCV=1^2+0^2+1^2+1.\ 414^2+0^2+0^2+1^2+1.414^2=7$

Sehingga Besar Rasio = BCV/WCV = 6.650 / 7 = 0.950

Karena langkah ini merupakan iterasi 1 maka lanjutkan ke langkah berikutnya

4. **Langkah 4:** Pembaruan centroid dengan menghitung rata-rata nilai pada masing-masing cluster.

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

Sehingga didapatkan centroid baru yaitu : m1=(3,3),m2=(1.25,1.75),m3=(4.33,2.67)

5. **Langkah 3:** (Iterasi-2) Kembali kelangkah 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. Selanjutnya pada langkah ini dilakukan penempatan lagi data dalam centroid terdekat sama seperti yang dilakukan dilangkah-3. Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sbb:

   Data : (1,3) , m1=(3,3),m2=(1.25,1.75),m3=(4.33,2.67)

   DM1 =$\sqrt{(1-3)^2+(3-3)^2}$ = 2

   D M2 = $\sqrt{(1-1.25)^2+(3-1.75)^2}$ = 1.275

   DM3 = $\sqrt{(1-4.33)^2+(3-2.67)^2}$ = 3.350

Nasabah	Jarak ke centroid custer1	Jarak ke centroid custer2	Jarak ke centroid custer3	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

Dari tabel diatas didapatkan keanggotaan nasabah sbb:

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

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*) :

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

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

Bila dibandingkan maka rasio sekarang (1.394) lebih besar dari rasio sebelumnya (0.950) oleh karena itu algoritma dilanjutkan kelangkah berikutnya

6. Langkah ke 4 – iterasi 3

Pada langkah ini dilakukan pembaruan centroid lagi:

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

Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sbb:

Data nasabah A : (1,3) , m1=(3,3),m2=(1.25,1.75),m3=(4.33,2.67)

DM1 =$\sqrt{(1-3)^2+(3-3)^2}$ = 2

D M2 = $\sqrt{(1-1.25)^2+(3-1.75)^2}$ = 1.275

DM3 = $\sqrt{(1-4.33)^2+(3-2.67)^2}$ = 3.350

Nasabah	Jarak ke centroid custer1	Jarak ke centroid custer2	Jarak ke centroid custer3	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

Dari tabel diatas didapatkan keanggotaan nasabah sbb:

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

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*) :

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

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

Bila dibandingkan maka rasio sekarang (1.394) sudah tidak lagi lebih besar dari rasio sebelumnya (1.394) oleh karena itu algoritma akan dihentikan.

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).

Daftar Pustaka

Rui Xu,Donald and C. Wunsch, Clustering ,John Wiley & Sons, INC,2009

Eko Prasetyo, Data Mining konsep dan aplikasi menggunakan Matlab,Andi Offset,2012

Sani Susanto and Dedy suryadi, Pengantar Data Mining, Andi Offset,2010