seri tulisan data mining

.: Analisis Gerombol - Bagian 2 :.

Penggerombolan Tak Berhirarki - Algoritma k-means

Bagus Sartono

bagusco@gmail.com

June 4, 2016

Abstract

Pada seri tulisan ini akan dipaparkan beberapa hal dasar dan umum mengenai analisis gerombol (*cluster analysis*) dengan menitikberatkan pada pemahaman konsep dasar dan penjelasan proses algoritmanya. Beberapa aplikasi akan diberikan dengan uraian yang lebih terbatas. Kritik, saran dan pertanyaan terhadap materi tulisan ini dapat disampaikan melalui email pada alamat **bagusco@gmail.com**. Penulis akan sangat mengapresiasi berbagai masukan tersebut. Akhirnya, selamat membaca.

1 Pengantar

Seperti yang telah diungkapkan pada Bagian 1 dari seri tulisan mengenai Analisis Gerombol ini bahwa terdapat teknik penggerombolan yang bersifat tak berhirarki. Lebih khusus, tulisan pada bagian kedua ini akan membahas tentang algoritma yang dikenal sebagai k-means. Simbol k pada nama k-means menunjukkan banyaknya gerombol yang terbentuk.

Secara verbal sederhana algoritma ini bekerja sebagai berikut. Andaikan kita memiliki banyak objek yang masing-masing diamati pada beberapa variabel. Jika ada sebanyak p buah variabel maka kita bisa memandang masing-masing objek terletak pada ruang berdimensi p. Selanjutnya andaikan dari banyak objek tersebut kita ingin memperoleh k buah gerombol.

Yang dilakukan oleh algoritma ini pertama kali adalah menentukan secara acak koordinat titik tengah dari k buah gerombol. Titik tengah dari gerombol ini selanjutnya disebut sebagai *centroid*. Dengan kata lain pada bagian awal ditentukan koordinat dari k buah centroid yang masing-masing berupa titik dalam ruang berdimensi p.

Dari setiap objek yang ada, kita dapat menghitung jaraknya terhadap masing-masing dari k buah centroid. Melalui nilai jarak tersebut, dengan mudah kita dapat menentukan ke centroid

mana objek tersebut jaraknya paling dekat. Setiap objek selanjutnya dimasukkan ke dalam satu buah cluster yang centroidnya jaraknya paling dekat ke objek tersebut. Dengan cara ini maka setiap objek sekarang sudah memiliki 'status' masuk ke gerombol mana.

Algoritma tidak berhenti sampai disitu. Setelah jelas keanggotaan masing-masing objek, proses berikutnya adalah menghitung koordinat centroid baru dari setiap gerombol. Koordinat centroid baru dari suatu gerombol tersebut diperoleh dengan cara merata-ratakan koordinat (data) dari semua objek yang tergabung ke dalam gerombol itu. Jika pada masing-masing gerombol sudah dilakukan, maka akan diperoleh k buah koordinat centroid yang baru.

Menggunakan koordinat yang baru, selanjutnya dilakukan lagi identifikasi keanggotaan setiap objek dengan menentukan jarak dari objek ke centroid dan melihat mana yang paling dekat. Proses kemudian dilanjutkan dengan menghitung kembali koordinat centroid yang baru. Demikian seterusnya dilanjutkan dengan iterasi berikutnya dan berikutnya, dan kemudian berhenti pada saat koordinat centroid tidak mengalami perubahan (atau perubahannya sudah dianggap dapat diabaikan).

2 Ilustrasi Sederhana: data berdimensi satu

Untuk menjelaskan bagaimana algoritma yang telah dipaparkan di atas lebih kongkrit, berikut disajikan ilustrasi yang sangat sederhana yang melibatkan 9 objek dengan pengamatan dilakukan hanya pada 1 (satu) variabel. Dengan demikian data hanya memiliki dimensi satu.

Andaikan saja sembilan amatan tersebut masing-masing memiliki nilai:

$$1,\ 2,\ 2,\ 3,\ 4,\ 4,\ 8,\ 8,\ 10.$$

Sembilan objek tersebut selanjutnya ingin dikelompokkan menjadi dua grup.

Awalnya, metode k-means akan bekerja dengan menentukan centroid awal. Beberapa software melakukan inisialisasi koordinat centroid ini dengan memilih secara acak k buah amatan dari data yang ada. Ada juga yang memilih centroid awal dengan mengambi k buah amatan paling atas atau paling bawah. Misalnya saja pada ilustrasi ini dipilih centroid pertama yaitu $C_1 = 1$ dan centroid gerombol kedua $C_2 = 2$.

Kemudian pada iterasi pertama akan ditentukan amatan mana saja yang lebih dekat dengan centroid gerombol pertama dan amatan mana saja yang lebih dekat ke centroid gerombol kedua. Karena datanya hanya berdimensi satu, maka kedekatan dapat dengan mudah menggunakan selisih dua nilai saja. Misalnya, amatan dengan nilai 1 cenderung lebih dekat dengan $C_1 = 1$ daripada ke $C_2 = 2$ karena selisihnya 0 dibandingkan selisihnya 1 untuk ke gerombol 2. Sebaliknya amatan dengan nilai 8 cenderung lebih dengan dengan $C_2 = 2$ yang selisihnya 6

dibandingkan dengan gerombol dengan centroid $C_1 = 1$ yang selisihnya 7. dengan demikian dua gerombol yang terbentuk adalah

$$\text{Gerombol } 1 = 1, \text{ Gerombol } 2 = 2, 2, 3, 4, 4, 8, 8, 10$$

sehingga diperoleh dua centroid yang baru yaitu $C_1 = 1$ dan $C_2 = 5.125$ yang masing-masing diperoleh dari rata-rata nilai anggota-anggotanya.

Selanjutnya pada iterasi berikutnya dilakukan penggerombolan ulang setiap amatan dengan menggunakan centroid C_1 dan C_2 yang baru. Hasilnya adalah

$$\text{Gerombol } 1 = 1, 2, 2, 3, \text{ Gerombol } 2 = 4, 4, 8, 8, 10.$$

Amatan dengan nilai 3 tadinya ada di gerombol 2 namun sekarang masuk ke gerombol pertama karena jaraknya dengan C_1 adalah 2.00 sedangkan jaran ke C_2 lebih jauh yaitu 2.125. Sementara amatan dengan nilai 8 tetap di gerombol 2 karena jarak ke C_2 adalah 2.875 dan ini lebih dekat dibandingkan dengan jarak ke C_1. Dengan keanggotaan yang baru tersebut diperoleh koordinat centroid baru yaitu $C_1 = 2$ dan $C_2 = 6.8$.

Iterasi yang ketiga selanjutnya memperoleh keanggotaan gerombol yang baru menjadi

$$\text{Gerombol } 1 = 1, 2, 2, 3, 4, 4, \text{ Gerombol } 2 = 8, 8, 10$$

yang kemudian membentuk koordinat centroid baru yaitu $C_1 = 2.67$ dan $C_2 = 8.67$.

Hasil di atas kemudian berlanjut ke iterasi keempat dengan keanggotaan

$$\text{Gerombol } 1 = 1, 2, 2, 3, 4, 4, \text{ Gerombol } 2 = 8, 8, 10$$

dan centroid $C_1 = 2.67$ dan $C_2 = 8.67$. Perhatikan bahwa hasil dari iterasi keempat ini memberikan koordinat centroid yang sama dengan iterasi sebelumnya. Dengan demikian proses iterasi k-means berhenti dan menghasilkan dua gerombol seperti yang diperoleh pada iterasi terakhir ini.

3 Algoritma k-means

Jika pada bagian sebelumnya telah disebutkan secara verbal algoritma dari teknik k-means. Kita coba tuliskan dalam notasi matematisnya.

Andaikan terdapat n buah amatan yang masing-masing diamati pada p buah peubah. Setiap amatan dapat dituliskan dalam bentuk vektor baris yang berisi p nilai yaitu $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$ yang merupakan sebuah titik di ruang berdimensi p, dengan $i = 1, 2, \ldots, n$. Apabila ingin dibuat k buah gerombol dengan centroid $\mathbf{C}_1, \ldots \mathbf{C}_k$, maka tahapan-tahapan yang dilakukan adalah:

1. Inisialisasi koordinat centroid

   Berikan sembarang nilai (koordinat) terhadap $\mathbf{C}_1, \ldots \mathbf{C}_k$. Kita dapat saja mengambil k buah amatan dari n amatan yang ada baik secara acak maupun subjektif. Kita dapat pula mensuplai dengan k buah titik berdasarkan informasi-informasi yang diperoleh dari proses terdahulu. Pada kasus n yang cukup besar, beberapa peneliti menyarankan mengambil sampel acak dari n amatan tersebut dan melakukan *hierarchical clustering* untuk selanjutnya menggunakan centroid hasil penggerombolan tersebut sebagai titik awal centroid.

2. Perbaikan nilai Centroid

   (a) Hitung jarak setiap amatan ke masing-masing centroid

   Proses ini akan menghasilkan matriks $n \times p$ dimana baris ke-i dan kolom ke-j-nya merupakan jarak dari amatan $\mathbf{x}_i$ ke centroid $\mathbf{C}_j$, untuk setiap pasangan $i = 1, \ldots, n$ dan $j = 1, \ldots, k$.

   (b) Tentukan keanggotaan gerombol dari setiap amatan

   Keanggotaan gerombol dari setiap amatan ditentukan berdasarkan centroid mana yang terdekat jaraknya. atau

$$M_i = \arg_{j=1,\ldots,k} d(\mathbf{x}_i, \mathbf{C}_j) \tag{1}$$

   (c) Lakukan updating terhadap koordinat centroid

   Centroid baru dari suatu gerombol diperoleh dengan merata-ratakan semua anggota gerombol.

3. Penghentian iterasi

   Tahapan nomor (2) diulangi beberapa kali sampai centroid yang baru tidak mengalami perubahan dibandingkan centroid dari iterasi sebelumnya.

4 Ilustrasi Sederhana: data berdimensi dua

Mari kita bahas satu buah ilustrasi sederhana yang lain menggunakan data berdimensi dua, atau data dengan dua buah variabel. Ilustrasi ini sengaja dibuat dengan dua variabel agar dapat dengan mudah divisualisasikan. Data yang digunakan terdiri atas 22 amatan dengan peubah tinggi badan dan berat badan, seperti yang ditampilkan pada Tabel 1. Tampilan visual dari dari posisi titik-titik amatan disajikan pada plot tebaran (*scatter plot*) pada Gambar 1. Menggunakan data ini akan diterapkan algoritma k-means untuk membentuk 3 (tiga) buah gerombol.

Tahap pertama dari k-means adalah inisialisasi centroid. Misalnya saja tiga buah amatan dipilih secara acak dan masing-masing dari ketiganya digunakan sebagai centroid awal. Tabel

Table 1: Data untuk ilustrasi algoritma k-means dengan dua peubah

ID	tinggi badan suami	tinggi badan istri
1	175	175
2	178	178
3	175	166
4	180	179
5	185	181
6	178	163
7	175	158
8	181	154
9	169	155
10	171	154
11	177	156
12	158	155
13	158	152
14	175	149
15	172	163
16	158	151
17	153	155
18	150	150
19	157	152
20	154	150
21	157	145
22	160	154

2 menampilkan tiga buah centroid pada tahap inisialisasi ini. Sementara Gambar 2 menyajikan tampilan posisi centroid tersebut. Masing-masing amatan selanjutnya dimasukkan ke dalam gerombol yang berpadanan dengan centroid terdekat. Perbedaan warna pada Gambar 2 merupakan perbedaan gerombol yang dihasilkan.

Table 2: Centroid pada tahap inisialisasi

ID	tinggi badan suami	tinggi badan istri
centroid 1	158	151
centroid 2	157	152
centroid 3	160	154

Lihat bahwa posisi ketiga centroid saling berdekatan. Warna merah adalah amatan dengan

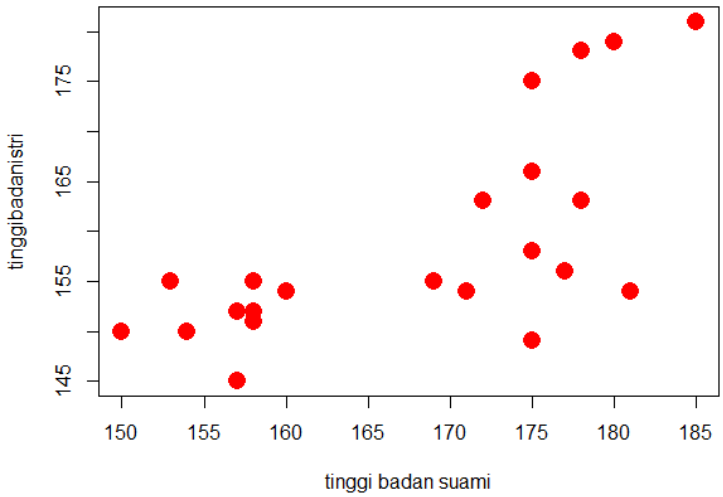

Figure 1: Data ilustrasi

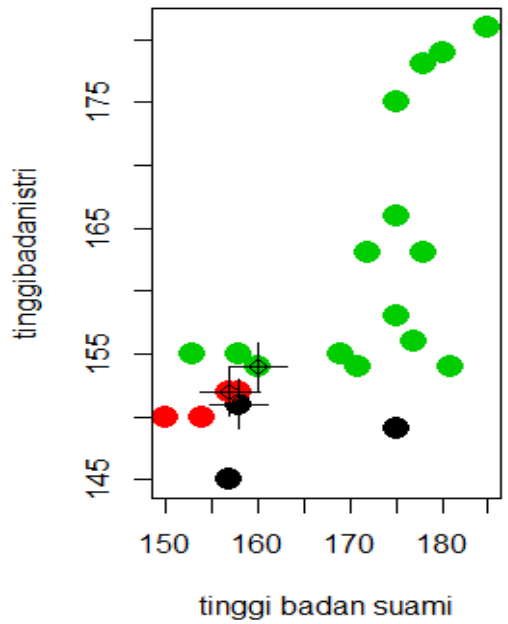

Figure 2: Iterasi 0

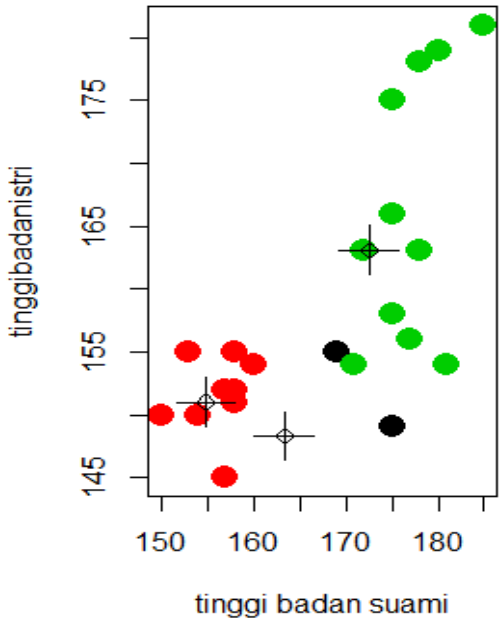

Figure 3: Iterasi 1

lebih dekat ke centroid paling kiri, sedangkan warna hitam adalah amatan yang paling dekat dengan centroid yang posisinya paling bawah. Amatan-amatan dengan warna hijau merupakan amatan yang paling dekat dengan centroid yang berada di kanan atas. Perhatikan bahwa centroid yang paling kanan (centroid gerombol hijau) selanjutnya akan berpindah lebih ke kanan atas lagi karena posisi centroid yang baru akan merupakan rata-rata dari setiap amatan. Jika amatan-amatan yang berwarna hijau di rata-ratakan tentu saja akan cenderung memiliki rata-rata di sekitar koordinat 175an untuk sumbu horizontal dan 163an untuk sumbu vertikalnya. Perubahan yang sama juga akan dialami oleh dua centroid yang lain. Hasil koordinat centroid yang baru disajikan pada Tabel 3 yang kemudian posisinya digambarkan pada Gambar 3.

Table 3: Centroid pada iterasi 1

ID	tinggi badan suami	tinggi badan istri
centorid 1	163.3	148.3
centroid 2	154.8	151.0
centroid 3	172.5	163.1

Selanjutnya cerita yang sama akan terjadi kembali bahwa setiap amatan kemudian diukur jaraknya dengan ketiga centroid yang baru dan kemudian ditentukan membershipnya berdasarkan pada centroid mana dia paling dekat. Perhatikan bahwa beberapa amatan yang dikiri dengan

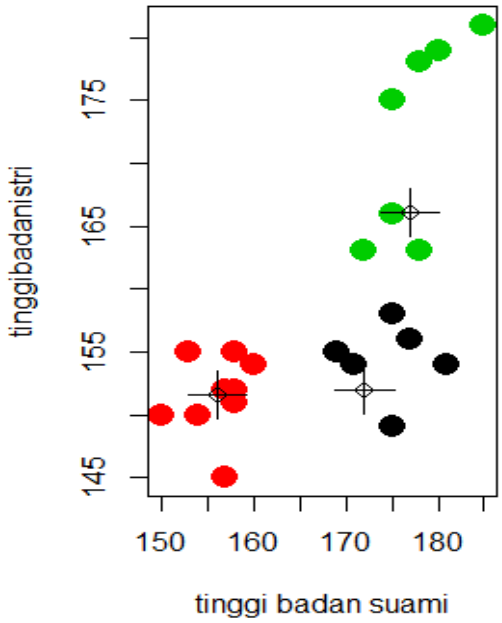

Figure 4: Iterasi 2

warna merah lebih banyak dibanding dengan sebelumnya. Sementara yang berwarna hijau berkurang karena beberapa amatan yang berada di bawah akan menyeberang ke gerombol lain mengingat centroid hijau yang baru tadi bergeser ke atas.

Dengan adanya perubahan membership ini tentu saja akan kembali terjadi pergeseran koordinat centroid. Perhatikan gerombol amatan yang berwarna hitam. Ada dua amatan yang masuk ke gerolmbol tersebut. Centroid baru bagi gerombol warna hitam ini akan bergeser berada tepat di antara kedua amatan itu, yang berarti centroidnya akan bergerak ke kanan dan ke atas menuju titik kira-kira 170 pada sumbu horizontal dan 152 pada sumbu vertikal. Centroid dari gerombol hijau juga akan bergeser sedikit naik. Hasil koordinat yang baru disajikan pada Tabel 4 dan visualisasi posisinya ada pada Gambar 4.

Table 4: Centroid pada iterasi 2

ID	tinggi badan suami	tinggi badan istri
centroid 1	172.0	152.0
centroid 2	156.1	151.6
centroid 3	177.0	166.1

Jika kita perhatikan Gambar 4 maka terlihat bahwa centroid dari gerombol hijau tidak persis di tengah dan cenderung terlalu ke bawah. Maka kita akan mengharapkan selanjutnya akan

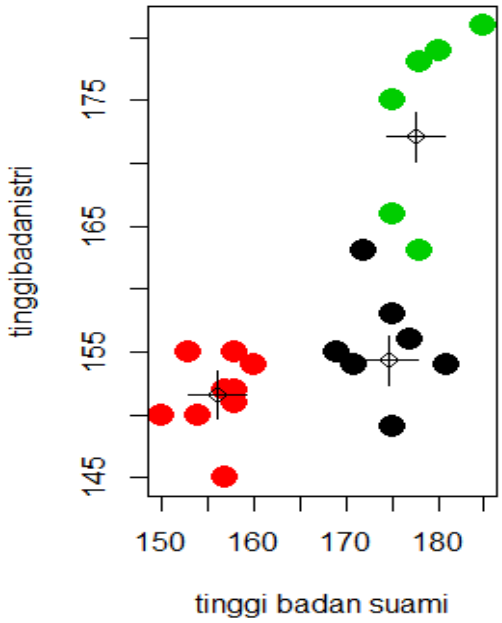

Figure 5: Iterasi 3

terjadi koreksi terhadap posisi centroid itu dengan posisi baru semesetinya akan bergeser ke atas agar lebih berimbang. Demikian juga dengan centroid dari gerombol yang berwarna hitam yang cenderung terlalu ke kiri.

Pada iterasi ketiga, koordinat centroid mengalami perubahan dan koordinat centroid yang baru selanjutnya ditampilkan pada Tabel 5. Tampilan visual dari posisi centroid yang baru diberikan pada Gambar **??**. Terlihat bahwa perubahan posisi centroid bagi gerombol hijau dan hitam terjadi seperti yang kita bayangkan sebelumnya, sementara centroid yang merah tidak berubah.

Table 5: Centroid pada iterasi 3

ID	tinggi badan suami	tinggi badan istri
centroid 1	174.7	154.3
centroid 2	156.1	151.6
centroid 3	177.6	172.1

Berdasarkan hasil perubahan centroid pada iterasi ketiga, terjadi pula pergeseran keanggotaan cluster. Kali iini tidak banyak perubahannya dimana satu amatan dari cluster hijau berpindah ke cluster hitam. Pergerseran itu selanjutnya akan menyebabkan perubahan centroid hitam dan hijau (centroid cluster 1 dan centroid cluster 3), seperti yang ditunjukkan pada Tabel

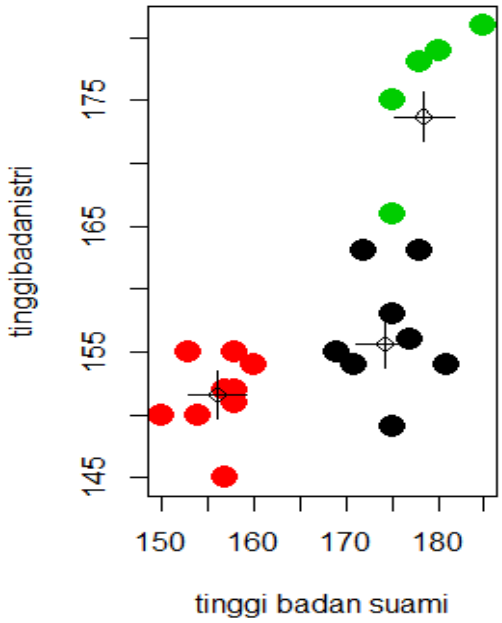

Figure 6: Iterasi 4

6 dan visualisasinya pada Gambar 6.

Table 6: Centroid pada iterasi 4

ID	tinggi badan suami	tinggi badan istri
centroid 1	174.3	155.6
centroid 2	156.1	151.6
centroid 3	178.5	173.7

Iterasi keempat menghasilkan sedikit perubahan saja karena hanya ada satu amatan yang bergeser dari hijau ke hitam, sedangkan selebihnya tidak mengalami perubahan. Dengan hasil tersebut, diperoleh bahwa sedikit pergeserah centroid seperti yang disajikan pada Tabel 7 dan visualisasinya disajikan pada Gambar 7.

Table 7: Centroid pada iterasi 5

ID	tinggi badan suami	tinggi badan istri
centroid 1	174.8	156.5
centroid 2	156.1	151.6
centroid 3	178.6	175.8

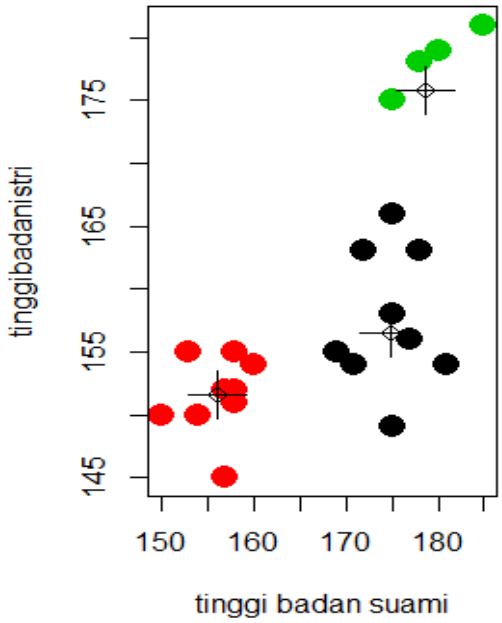

Figure 7: Iterasi 5

Centroid dari hasil pengkelasan baru yang digambarkan oleh Gambar 7 mengalami sedikit pergeseran menjadi koordinat pada Tabel 8. Ini merupakan centroid final karena pada iterasi berikutnya tidak terjadi lagi pergeseran.

Table 8: Centroid pada iterasi 6

ID	tinggi badan suami	tinggi badan istri
centroid 1	174.8	157.6
centroid 2	156.1	151.6
centroid 3	179.5	178.3

Hasil akhir dari algoritma ini menghasilkan centroid pada Tabel 8 dan keanggotaan dari masing-masing amatan disajikan pada Tabel 9, dengan Gerombol 1 beranggotakan 9 amatan, Gerombol 2 beranggotakan 9 amatan, dan Gerombol 3 beranggotakan 4 amatan.

5 Algoritma k-Means di R

Software R menyediakan fungsi dengan nama kmeans pada package stats untuk menjalankan algoritma k-means. Perintah sederhana yang dapat diberikan adalah:

$$\text{kmeans(x, centers, iter.max} = 10)$$

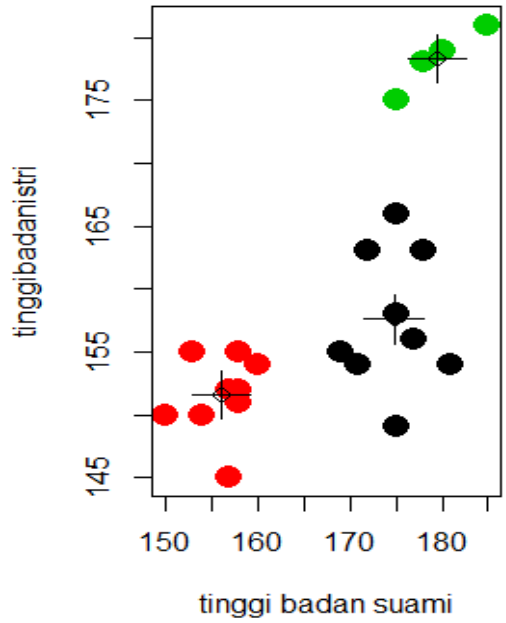

Figure 8: Iterasi 6 (terakhir)

dengan x adalah matriks data numerik atau data frame yang semua kolom-kolomnya bersifat numerik, centers menyebutkan banyaknya gerombol yang akan dibuat atau menyebutkan koordinat awal centroid, dan iter.max menyebutkan banyaknya iterasi maksimum dengan default 10 iterasi.

Sebagai contoh, andaikan data pada ilustrasi di bagian sebelumnya kita simpan pada file dengan nama ilustrasi2.csv maka pertama yang perlu kita lakukan adalah membaca data tersebut. Perintah yang bisa digunakan adalah fungsi read.csv dan selanjutnya menyebutkan nama data yang akan dibaca. Data itu selanjutnya disimpan pada suatu *data frame* dengan nama ilustrasi. Perintah head ilustrasi adalah perintah untuk menampilkan beberapa baris pertama dari data frame ilustrasi.

```
> ilustrasi <- read.csv("D:/ilustrasi2.csv", header=T)
> head(ilustrasi)
  tinggibadan tinggibadanistri
1         175              175
2         178              178
3         175              166
4         180              179
5         185              181
6         178              163
```

Selanjutnya perintah untuk menjalankan algoritma k-means adalah

Table 9: Hasil akhir penggerombolan untuk ilustrasi algoritma k-means dengan dua peubah

ID	tinggi badan suami	tinggi badan istri	kode gerombol
1	175	175	3
2	178	178	3
3	175	166	1
4	180	179	3
5	185	181	3
6	178	163	1
7	175	158	1
8	181	154	1
9	169	155	1
10	171	154	1
11	177	156	1
12	158	155	2
13	158	152	2
14	175	149	1
15	172	163	1
16	158	151	2
17	153	155	2
18	150	150	2
19	157	152	2
20	154	150	2
21	157	145	2
22	160	154	2

```
> hasilgerombol <- kmeans(ilustrasi, centers=3, iter.max =10)
```

yang menjalankan algoritma k-means untuk menghasilkan 3 (tiga) buah gerombol. Selanjutnya hasil dari algoritma itu disimpan pada objek dengan nama *hasilgerombol*.

Untuk melihat keanggotaan dari setiap gerombol gunakan perintah

```
> hasilgerombol$cluster
 [1] 1 1 2 1 1 2 2 2 2 2 2 3 3 2 2 3 3 3 3 3 3 3
```

yang memberikan kode gerombol untuk setiap amatan. Karena ada 22 amatan maka ada 22 angka 1 sampai 3. Cara membacanya adlah bahwa amatan no 1, 2, 3, 5 dimasukkan pada Gerombol 1. Sedangkan amatan nomor 3 dimasukkan pada Gerombol 2 bersama dengan amatan nomor 6, 7 dan sebagainya. Banyaknya anggota setiap gerombol dapat ditampilkan menggunakan perintah

```
> hasilgerombol$size
[1] 4 9 9
```

Sedangkan centroid dari masing-masing cluster diperoleh dengan menggunakan perintah

```
> hasilgerombol$centers
  tinggibadan tinggibadanistri
1    179.5000        178.2500
2    174.7778        157.5556
3    156.1111        151.5556
```

6 Mengevaluasi Kebaikan Penggerombolan

Terdapat beberapa cara yang dapat dilakukan untuk mengevaluasi kebaikan penggerombolan. Beberapa orang menggunakan istilah *clustering validity* terhadap proses ini. Banyak usulan nilai statistik atau indeks yang digunakan, tetapi kesemuanya menggunakan prinsip bahwa antar objek dalam satu cluster bersifat dekat/mirip, dan antar cluster bersifat jauh/tak-mirip. Dari sekian banyak nilai/kriteria evaluasi, yang akan dibahas di bagian ini adalah kriteria yang menggunakan konsep jumlah kuadrat (sum of squares), sedangkan berbagai jenis kriteria evaluasi yang lain akan dibahas pada tulisan di bagian lain berikutnya (mudah-mudahan akan segera terselesaikan).

Tiga nilai jumlah kuadrat yang bisa dihitung dari suatu hasil penggerombolah adalah

1. Jumlah Kuadrat Dalam Gerombol (Within-Cluster Sum of Squares) yang dinotasikan SS_w. Ini merupakan ukuran kedekatan antar objek dalam gerombol. Nilai SS_w yang kecil mengindikasikan kesamaan/kemiripan antar objek dalam gerombol. Dari **setiap** gerombol bisa dihitung nilai JKDG ini menggunakan formula:

$$SS_w = \sum_i \sum_j (x_{ij} - \bar{x}_j)^2 \tag{2}$$

   dengan i adalah indeks amatan dan j adalah indeks variabel, dan $\bar{x}_j$ adalah rata-rata nilai variabel ke-j pada gerombol tertentu.

2. Jumlah Kuadrat Antar Gerombol (Between Cluster SUm of Squares) yang dinotasikan SS_b. Ini merupakan ukuran keterpisahan antar gerombol. Nilai SS_b yang besar mengindikasikan hasil penggerombolan yang semakin baik. Dari satu hasil penggerombolan dapat dihitung nilai SS_b dengan formula

$$SS_b = \sum_k \sum_j \bar{x}_{jk} - \bar{\bar{x}}_j)^2 \tag{3}$$

dengan k adalah indeks gerombol, dan j adalah indeks variabel. Nilai $\bar{x}_{jk}$ merupakan nilai rata-rata variabel ke-j pada gerombol ke-k dan $\bar{\bar{x}}_j$ adalah nilai rata-rata variabel ke-j dari semua amatan.

3. Jumlah Kuadrat Total (Total Sum of Squares) yang dinotasikan SS_t diperoleh dari $SS_t = \sum_k SS_w + SS_b$

Pada R, hasil dari algoritma k-means, kita bisa menampilkan SS_w dari masing-masing gerombol dan SS_b menggunakan perintah

```
> hasilgerombol$withinss
[1]  71.7500 347.7778 157.1111
> hasilgerombol$betweenss
[1] 4213.952
```

Sedangkan total dari ketiga SS_w diperoleh dengan perintah

```
> hasilgerombol$tot.withinss
[1] 576.6389
```

7 Menentukan Banyaknya Gerombol

Pada saat kita menggunakan algoritma k-means, kita diharapkan menentukan di awal berapa banyaknya gerombol yang akan dibuat. Dalam beberapa kasus, analis tidak sulit menentukan ini karena banyaknya gerombol dapat disesuaikan dengan kebutuhan yang ada. Namun dalam banyak kasus yang lain, analis kesulitan menentukan nilai banyaknya gerombol ini.

Indikator jumlah kuadrat dalam gerombol dapat dijadikan petunjuk untuk menentukan banyaknya gerombol. Nilai total dari jumlah kuadrat dalam gerombol ini akan mengecil dengan semakin banyaknya gerombol yang dibuat. Jika proses penggerombolan dengan algoritma k-means diulang-ulang untuk berbagai k (banyaknya gerombol) maka kita akan memperoleh gambaran berapa banyak gerombol yang memadai. Caranya adalah menentukan pada kondisi berapa gerombol yang kalau ditambah lagi akan diikuti dengan penurunan nilai jumlah kuadrat dalam gerombol yang tidak signifikan.

Berikut ini adalah perintah yang dapat digunakan untuk menghasilkan plot nilai total jumlah kuadrat dalam gerombol untuk berbagai kondisi banyaknya gerombol mulai dari 1 hingga 10 gerombol. Program R tersebut berupa fungsi yang diberi nama wss dengan input data yang digunakan dan banyaknya gerombol maksimum (dengan nilai default 15 gerombol). Selanjutnya fungsi akan bekerja dengan menjalankan algoritma kmeans untuk 1, 2, 3, dst gerombol dan menyimpan nilai total jumlah kuadrat. Nilai itulah yang kemudian digambarkan.

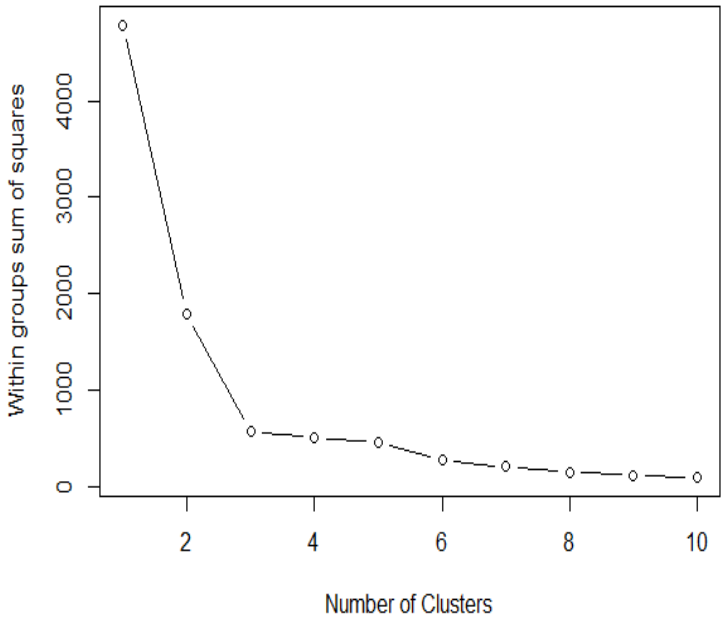

Figure 9: Jumlah kuadrat dalam gerombol untuk berbagai banyaknya gerombol

```
wssplot <- function(data, nc=15, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")}

wssplot(ilustrasi, nc=10)
```

Gambar 9 merupakan grafik nilai jumlah kuadrat untuk penggerombolan menjadi 1 hingga 10 gerombol dari data ilustrasi. Terlihat bahwa ketika banyaknya gerrombol dari 1 menjadi 2 dan dari 2 menjadi 3 gerombol terjadi penurunan jumlah kuadrat yang drastis. Namun penurunan itu kemudian tidak lagi signifikan ketika banyaknya gerombol berubah dari 3 gerombol menjadi 4 gerombol. Berdasarkan hasil tersebut kita dapat menyimpulkan bahwa 3 (tiga) gerombol sudah memadai.

8 Penutup

Telah dipaparkan penjelasan mengenai algoritma k-means untuk analisis gerombol yang disertai dengan ilustrasi sederhana. Mudah-mudahan tulisan ini bermanfaat.