

## Cluster Analysis

Hery Tri Sutanto

Jurusan Matematika MIPA UNESA

### Abstrak

Dalam analisis cluster mempelajari hubungan interdependensi antara seluruh set variabel perlu diteliti. Tujuan utama analisis cluster adalah mengelompokkan obyek (elemen) seperti orang, produk (barang), toko, perusahaan ke dalam kelompok-kelompok yang relatif homogen berdasarkan pada suatu set variabel yang dipertimbangkan untuk diteliti. Obyek di dalam setiap kelompok harus relatif mirip/sama. Variabel-variabel pada cluster ini harus jauh berbeda dengan obyek dari cluster lain. Jika digunakan cara seperti ini maka analisis cluster merupakan bagian depan dari analisis faktor, dimana mereduksi (memperkecil) banyaknya obyek (responden) bukan banyaknya variabel atau atribut responden, yaitu mengelompokkan obyek-obyek tersebut kedalam cluster yang banyaknya lebih sedikit dari banyaknya obyek asli yang diteliti, misalnya dari 50 orang responden, dikelompokkan dengan 5 cluster dengan setiap cluster terdiri dari 10 orang.

Kata kunci :

Cluster, mereduksi, interdependensi dan relatif sama

#### 1. Pendahuluan

Seperti pada analisis faktor, analisis cluster juga meneliti semua hubungan interpendensi, tidak ada variabel bebas dan tak bebas dalam analisis cluster. Perbedaan variabel bebas dan tidak bebas terjadi dalam analisis regresi, analisis varians, analisis diskriminan dimana kita ingin mempelajari pengaruh setiap variabel bebas, baik sendirian maupun secara bersama-sama terhadap variabel tak bebas.

Di dalam analisis cluster hubungan interdependensi antara seluruh set variabel perlu diteliti. Tujuan utama analisis cluster adalah mengelompokkan obyek (elemen) seperti orang, produk (barang), toko, perusahaan ke dalam kelompok-kelompok yang relatif homogen berdasarkan pada suatu set variabel yang dipertimbangkan untuk diteliti. Obyek di dalam setiap kelompok harus relatif mirip/sama. Variabel-variabel pada cluster ini harus jauh berbeda dengan obyek dari

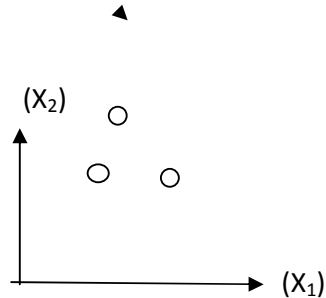
cluster lain. Jika digunakan cara seperti ini maka analisis cluster merupakan bagian depan dari analisis faktor, dimana mereduksi (memperkecil) banyaknya obyek (responden) bukan banyaknya variabel atau atribut responden, yaitu mengelompokkan obyek-obyek tersebut kedalam cluster yang banyaknya lebih sedikit dari banyaknya obyek asli yang diteliti, misalnya dari 50 orang responden, dikelompokkan dengan 5 cluster dengan setiap cluster terdiri dari 10 orang.

Pada makalah ini menguraikan konsep dasar analisis cluster. Tahap-tahap untuk menjalankan yang akan dikaji dan diilustrasikan didalam pengclusteran hirarki dengan program paket statistik yang sudah dikenal dan pengclusteran tidak hirarki yang diikuti dengan pengclusteran variabel. Pada umumnya suatu obyek dikelompokkan ke dalam suatu cluster sedemikian rupa lebih berhubungan (berkorelasi) dengan obyek lainnya di dalam clusternya dari pada dengan obyek dari cluster lain. Pembentukan cluster berdasarkan kuat tidaknya hubungan antar-obyek. Metode ini dinamakan metode hirarki.

### 1.2 Konsep dasar

Analisis cluster merupakan suatu kelas tehnik yang digunakan untuk mengklasifikasikan obyek atau kasus (responden) ke dalam kelompok yang relatif homogen yang dinamakan cluster. Obyek dalam setiap kelompok cenderung mirip satu sama lain dan berbeda jauh dengan obyek dari cluster lainnya. Di dalam pengclusteran setiap obyek hanya boleh masuk ke dalam satu cluster saja sehingga tidak terjadi tumpang tindih (overlapping atau interaction).

Gambar 1



Dilihat apa yang diclusterkan, maka analisa cluster dibagi menjadi dua, yaitu:

- a. Pengclusteran observasi
- b. Pengclusteran variabel

Secara umum ada dua metode di dalam analisa cluster yaitu

- a. Metode hirarki, yaitu hasil pengclusterannya disajikan secara berjenjang dari  $n$ ,  $(n-1)$  sampai 1 cluster yang termasuk dalam metode

hirarki single linkage, complete linkage, average linkage, median linkage, dan centroid linkage

- b. Metode non hirarki adalah banyaknya cluster sudah diketahui dan biasanya metode ini dipakai dalam mengcluster data yang berukuran besar yaitu metode K “means”.

Untuk menyatakan suatu observasi atau variabel mempunyai sifat yang lebih dekat dengan observasi tertentu dari pada observasi yang lain digunakan fungsi jarak. Suatu fungsi dinamakan jarak jika mempunyai sifat

- a.  $d_{jj} \geq 0$  dan  $d_{ii} = 0$  jika  $i = j$
- b.  $d_{ij} = d_{ji}$  (simetris)
- c.  $d_{ik} \leq d_{ij} + d_{jk}$  panjang salah satu suatu segitiga lebih kecil atau sama dengan jumlah dua sisi yang lain

Beberapa macam jarak yang dipakai dalam analisa cluster

1 Jarak Euclidean dengan rumus  $d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2}$

2 Jarak Manhattan dengan rumus  $d_{ij} = \sum_{k=1}^p |X_{ik} - X_{jk}|$

3 Jarak Pearson dengan rumus  $d_{ij} = \sqrt{\sum_{k=1}^p \frac{(X_{ik} - X_{jk})^2}{\text{Var}(X_{ik})}}$

4 Jarak Korelasi dengan rumus  $d_{ij} = 1 - r_{ij}$

5. Jarak Mutlak Korelasi dengan rumus  $d_{ij} = 1 - |r_{ij}|$

Metode-metode pengclusteran hirarki dibedakan dengan dasar konsep jarak antar cluster. Metode-metode yang digunakan untuk menentukan jarak antara cluster

1. Single linkage mempunyai jarak antar cluster (i,j) dengan k  $d_{(i,j)k} = \min(d_{ik}, d_{jk})$
2. Complete linkage mempunyai jarak antara cluster (i,j) dengan k  $d_{(i,j)k} = \max(d_{ik}, d_{jk})$

3. Average linkage mempunyai jarak antara cluster (i,j) dengan k  
 $d_{(i,j)k} = \text{average}(d_{ik}, d_{jk})$

4. Median linkage mempunyai jarak antara cluster (i,j) dengan k  
 $d_{(i,j)k} = \text{median}(d_{ik}, d_{jk})$

Langkah-langkah dalam membuat cluster data dengan menggunakan metode hirarki

- Tentukan matriks jarak antar data yang diclusterkan
- Tentukan dua data yang mempunyai jarak terkecil kemudian gabungkan dua data ini ke dalam satu cluster
- Modifikasi matriks jarak sesuai aturan jarak antar cluster yang sesuai dengan metode yang dipakai
- Lakukan langkah 2 dan 3 sampai matriks jarak berorder 1X1

Langkah- langkah dalam membuat cluster data dengan metode non hirarki K “Means” adalah

- Mulai
- Tentukan k buah pusat awal
- Tentukan jarak setiap data ke titik pusat
- Lakukan pengclusteran setiap data ke pusat terdekat
- Tentukan nilai pusat baru sebagai rata-rata data dalam cluster
- Lakukan langkah 3-5 sampai nilai pusat cluster tidak berubah lagi
- Selesai.

#### Contoh

Seorang ahli pertanian ingin mengelompokkan (cluster) 6 jenis jagung berdasarkan sifat produksinya (hasil produksi). Keenam jenis jagung itu adalah jenis 1,2,3,4,5 dan 6 dibudidayakan di lima tempat atau lokasi ,yaitu  $L_1, L_2, L_3, L_4, L_5$

Tabel 1

Rata-rata Hasil Produksi (kwintal per ha) dari 6 jenis jagung di 5 Lokasi yang Berbeda

Lokasi Percobaan			Jenis jagung			
	1	2	3	4	5	6
$L_1$	51,7	37,1	25,5	47,4	39,5	36,1
$L_2$	30,2	29,5	32,9	29,5	47,6	40,3
$L_3$	22,6	17,7	35,6	32,8	25,8	28,8
$L_4$	23,7	23,9	32,2	34,7	30,6	30,6
$L_5$	36,5	30,1	35,1	28,2	34,5	34,5

Kemudian kita menentukan matriks jarak berdasarkan rumus

$$d_{ij} = 1 - r_{ij}^2$$

Dimana

$$r_{ij} = \frac{n \sum Y_i Y_j - (\sum Y_i) (\sum Y_j)}{\sqrt{\{n \sum Y_i^2 - (\sum Y_i)^2\} \{n \sum Y_j^2 - (\sum Y_j)^2\}}}$$

Kita hitung dulu koefisien korelasi antara jenis i dan j dimana  
i=1,2,3,4,5,6

j=1,2,3,4,5,6

Hasil perhitungan koefisien korelasi  $r_{ij}$  untuk i,j=1,2,3,4,5,6 sebagai berikut

$$\begin{array}{lll} r_{11} = 1,000 & r_{22} = 1,000 & r_{33} = 1,000 \\ r_{11} = 1,000 & r_{22} = 1,000 & r_{33} = 1,000 \\ r_{12} = 0,992 & r_{23} = -0,754 & r_{34} = -0,922 \\ r_{13} = -0,794 & r_{24} = 0,509 & r_{35} = -0,400 \\ r_{14} = 0,685 & r_{25} = 0,707 & r_{36} = -0,352 \\ r_{15} = 0,461 & r_{26} = 0,734 & r_{44} = 1,000 \\ r_{16} = 0,504 & r_{35} = 1,000 & r_{66} = 1,000 \\ r_{45} = 0,050 & r_{56} = 0,990 & r_{66} = 1,000 \\ r_{46} = 0,004 \end{array}$$

Dengan menggunakan nilai koefisien korelasi sederhana antara i dan j, maka nilai jarak antara jenis jagung i dan j berikut ini:

$$d_{ij} = 1 - r_{ij}^2 \quad i,j=1,2,3,4,5,6$$

$$d_{11} = 1 - r_{11}^2 = 0$$

$$d_{12} = 1 - r_{12}^2 = 0,152$$

$$d_{13} = 1 - r_{13}^2 = 0,370$$

$$d_{14} = 1 - r_{14}^2 = 0,531$$

$$d_{15} = 1 - r_{15}^2 = 0,787$$

$$d_{16} = 1 - r_{16}^2 = 0,746$$

$$d_{22} = 1 - r_{22}^2 = 0$$

$$d_{23} = 1 - r_{23}^2 = 0,431$$

$$d_{24} = 1 - r_{24}^2 = 0,741$$

$$d_{25} = 1 - r_{25}^2 = 0,500$$

$$d_{26} = 1 - r_{26}^2 = 0,461$$

$$d_{33} = 1 - r_{33}^2 = 0$$

$$d_{34} = 1 - r_{34}^2 = 0,150$$

$$d_{35} = 1 - r_{35}^2 = 0,840$$

$$d_{36} = 1 - r_{36}^2 = 0,876$$

$$d_{44} = 1 - r_{44}^2 = 0$$

$$d_{45} = 1 - r_{45}^2 = 0,998$$

$$d_{46} = 1 - r_{46}^2 = 0,999$$

$$d_{55} = 1 - r_{55}^2 = 0$$

$$d_{56} = 1 - r_{56}^2 = 0,020$$

$$d_{66} = 1 - r_{66}^2 = 0$$

Dengan menggunakan nilai jarak diatas diperoleh matriks jarak  $D_i$

Dibawah ini

$$D_1 = \begin{bmatrix} 0 & 0,152 & 0,370 & 0,531 & 0,787 & 0,746 \\ 0,152 & 0 & 0,431 & 0,741 & 0,500 & 0,461 \\ 0,370 & 0,431 & 0 & 0,150 & 0,840 & 0,876 \\ 0,531 & 0,741 & 0,150 & 0 & 0,998 & 0,999 \\ 0,787 & 0,500 & 0,840 & 0,998 & 0 & 0,020 \\ 0,746 & 0,461 & 0,876 & 0,999 & 0,020 & 0 \end{bmatrix}$$

Minimum jarak 0,020

Dengan menggunakan setiap jenis jagung sebagai suatu kelompok, maka suatu kriteria pengelompokkan adalah menggabungkan dua jenis {5,6} yang memiliki nilai jarak minimum yaitu nilai  $d_{56} = 0,020$  maka kedua jenis jagung itu mempunyai kedekatan sifat hasil produksi sehingga jenis jagung 5 dan 6 dikelompokkan ke dalam satu kelompok yang diberi nama kelompok {5,6}.

Untuk melaksanakan pengelompokkan berikutnya, maka kita dapat menentukan jarak antara kelompok {5,6} dan kelompok jenis jagung yang tersisa 1,2,3,4 . Jarak minimum (neighbourhood) dan kelompok berikutnya dengan menggunakan rumus

$$d_{(x,y)z} = d\{(x,y),z\} = \min\{d(x,z), d(y,z)\}$$

Sehingga diperoleh

$$d_{(5,6)1} = d\{d(5,1), d(6,1)\} = \min\{0,787, 0,746\} = 0,746$$

$$d_{(5,6)2} = d\{d(5,2), d(6,2)\} = \min\{0,500, 0,461\} = 0,461$$

$$d_{(5,6)3} = d\{d(5,3), d(6,3)\} = \min\{0,840, 0,876\} = 0,840$$

$$d_{(5,6)4} = d\{d(5,4), d(6,4)\} = \min\{0,998, 0,999\} = 0,998$$

Dengan membuang baris dan kolom dari matriks  $D_1$  yang berhubungan dengan varitas 5 dan 6, kemudian melalui penambahan baris dan kolom untuk kelompok {5,6}, maka akan memperoleh matriks jarak yang baru  $D_2$  berikut ini

$$D_2 = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & (5,6) \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ (5,6) \end{matrix} & \begin{bmatrix} 0 & 0,152 & 0,370 & 0,531 & 0,746 \\ & 0 & 0,431 & 0,741 & 0,461 \\ & & 0 & 0,150 & 0,840 \\ & & & 0 & 0,998 \\ & & & & 0 \end{bmatrix} \end{matrix}$$

Dari matriks  $D_2$  diperoleh nilai jarak minimum  $d(3,4)=0,150$  yang berarti ada kedekatan sifat produksi jenis 3 dan 4. Sehingga perlu menggabungkan jenis 3 dan 4 pada jarak  $d(3,4)=0,150$  ke dalam satu kelompok baru yang diberi nama kelompok  $\{3,4\}$ .

Kita ulangi proses pengelompokkan berikutnya dengan cara menentukan jarak antara kelompok  $\{3,4\}$ ,  $\{5,6\}$ , 1,2 dengan menggunakan rumus

$$d_{(x,y)z} = d\{(x,y),z\} = \min\{d(x,z), d(y,z)\}$$

Sehingga diperoleh:

$$d_{(3,4)1} = d\{d(3,1), d(4,1)\} = \min\{0,370, 0,531\} = 0,370$$

$$d_{(3,4)2} = d\{d(3,2), d(4,2)\} = \min\{0,431, 0,741\} = 0,431$$

$$d_{(3,4)(5,6)} = d\{d(3, (5,6)), d(4, (5,6))\} = \min\{0,840, 0,998\} = 0,840$$

Kemudian diadakan penghapusan baris dan kolom dari matriks  $D_2$  yang berkaitan dengan jenis 3 dan 4, setelah itu menambahkan baris dan kolom untuk kolom(3,4) maka akan diperoleh matriks jarak baru  $D_3$  dibawah ini:

$$D_3 = \begin{matrix} & \begin{matrix} 1 & 2 & (3,4) & (5,6) \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ (3,4) \\ (5,6) \end{matrix} & \begin{bmatrix} 0 & 0,152 & 0,370 & 0,746 \\ & 0 & 0,431 & 0,461 \\ & & 0 & 0,840 \\ & & & 0 \end{bmatrix} \end{matrix}$$

Dari matriks  $D_3$  terdapat nilai jarak minimum  $d(1,2)=0,152$  kedalam satu kelompok yang diberi nama kelompok  $\{1,2\}$ .

Pengelompokkan diulangi lagi untuk mengelompokkan semua jenis jagung ke dalam satu kelompok secara bersama yang menandakan algoritma telah berakhir

Kemudian dikelompokkan dengan menghitung nilai jarak antara kelompok  $\{1,2\}$  yang baru terbentuk dan kelompok  $\{3,4\}$  dan  $\{5,6\}$ . Perhitungan jarak menggunakan rumus dibawah ini :

$$d_{(x,y)z} = d\{(x,y),z\} = \min\{d(x,z), d(y,z)\}$$

Sehingga diperoleh

$$d_{(1,2)(3,4)} = \min\{d[1, (3,4)], d[2, (3,4)]\} = \min\{0,370, 0,431\} = 0,370$$

$$d_{(1,2)(5,6)} = \min\{d[1, (5,6)], d[2, (5,6)]\} = \min\{0,746, 0,461\} = 0,461$$

Dengan penghapusan baris dan kolom dari matriks  $D_3$  yang berhubungan dengan jenis 1 dan 2, kemudian setelah menambah baris dan kolom untuk kolom {1,2} akan diperoleh matriks jarak baru  $D_4$  berikut ini:

$$d_{(1,2),(3,4)} = \min\{d[1, (3,4)], d[2, (3,4)]\} = \min\{0,370, 0,431\} = 0,370$$

$$d_{(1,2),(3,4)} = \min\{d[1, (3,4)], d[2, (3,4)]\} = \min\{0,370, 0,431\} = 0,370$$

$$D_4 = \begin{matrix} & \begin{matrix} (1,2) & (3,4) & (5,6) \end{matrix} \\ \begin{matrix} (1,2) \\ (3,4) \\ (5,6) \end{matrix} & \begin{bmatrix} 0 & 0,370 & 0,461 \\ & 0 & 0,840 \\ & & 0 \end{bmatrix} \end{matrix}$$

Dari matriks  $D_4$  terdapat jarak minimum  $d(3,4)=0,370$  dengan menggabungkan jenis 1,2,3 dan 4 kedalam kelompok baru dengan nama {1,2,3,4} menggunakan rumus dibawah ini

$$d_{(x,y)z} = d\{(x,y), z\} = \min\{d(x,z), d(y,z)\}$$

Sehingga diperoleh

$$\begin{aligned} d_{(1,2,3,4),(5,6)} &= d\{(1,2,3,4), (5,6)\} = \min\{d[(1,2), (5,6)], d[(3,4), (5,6)]\} \\ &= \min\{0,461, 0,840\} = 0,461 \end{aligned}$$

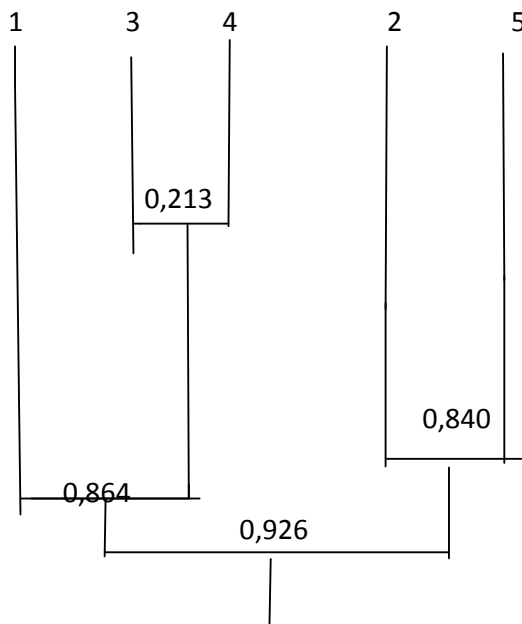
Dengan penghapusan baris dan kolom dari matriks  $D_4$  terdapat jarak minimum yang nilainya 0,370 yang berhubungan kelompok (1,2) dan (3,4) kemudian menambahkan baris dan kolom untuk kelompok (1,2,3,4) yang baru terbentuk, sehingga diperoleh matriks jarak baru:

$$D_5 = \begin{matrix} & \begin{matrix} (1,2,3,4) & (5,6) \end{matrix} \\ \begin{matrix} (1,2,3,4) \\ (5,6) \end{matrix} & \begin{bmatrix} 0 & 0,461 \\ & 0 \end{bmatrix} \end{matrix}$$

Matriks  $D_5$  merupakan matriks jarak terakhir dengan semua jenis jagung akan mengelompok menjadi satu kelompok {1,2,3,4,5,6} dengan jarak  $d=0,461$

Hasil pengelompokkan 6 jenis jagung dengan menggunakan metode Single Linkage dalam analisis cluster hirarki dapat disajikan dengan dendogram dibawah ini.





Gambar 1

Dari dendrogram diatas dengan jarak minimum  $d=0,40$  maka akan memperoleh dua cluster jenis jagung  $\{1,2,3,4\}$  dan cluster jenis jagung  $\{5,6\}$ .

Pada dasarnya batas nilai jarak  $d$  yang dipilih akan tergantung sejauh mana kekuatan yang diinginkan dengan memperhatikan semakin besar nilai jarak  $d$  akan semakin kuat pengelompokkan (cluster) itu. Jika dipilih  $d=0,50$  maka keenam jenis jagung cukup dikelompokkan satu cluster saja, tetapi jika dipilih  $d=0,30$  keenam jenis jagung dikelompokkan menjadi cluster, cluster  $\{1,2\}$ , cluster  $\{3,4\}$  dan cluster  $\{5,6\}$ .

#### Daftar Pustaka

1. Jupranto,2004. Analisis Multivariat Arti dan Interpretasi,Penerbit Rineka Cipta,Jakarta.
2. Gaspersz,Vincent,1995. Teknik Analisis Dalam Penelitian Percobaan, Penerbit Tarsito Bandung, Bandung.