**ANALISIS CLUSTER**

###### D E F I N I S I

Analisis kelompok (*cluster analysis*) merupakan salah satu teknik statistik multivariat untuk mengidentifikasi sekelompok obyek yang memiliki kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok obyek lainnya. Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variasi data obyek.

**T U J U A N**

Tujuan analisis kelompok adalah untuk mengelompokkan sekumpulan data obyek ke dalam beberapa kelompok yang memiliki karakteristik tertentu dan dapat dibedakan satu sama lain atau untuk mendapatkan *cluster* yang mempunyai obyek dengan memperlihatkan *small within–cluster variation relative to the between-cluster variation*. Hal ini dilakukan untuk analisis dan interpretasi lebih lanjut sesuai dengan tujuan penelitian yang dilakukan.

Terkait dengan tujuannya, maka secara logika *cluster* yang baik adalah *cluster* yang mempunyai :

* Homogenitas (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within cluster*)
* Heterogenitas (perbedaan) yang tinggi antar *cluster* yang satu dengan *cluster* lainnya (*between cluster*).

**T A H A P A N A N A L I S I S**

Tahapan analisis kelompok dapat dilihat pada Gambar berikut ini.



Gambar Tahapan Analisis Kelompok (*Cluster Analysis*)

Proses pengelompokan mulai dengan mengambil p ukuran dari setiap n obyek. nxp matrik data mentah ditransformasikan ke nxn matrik kesamaan atau ukuran jarak dimana kesamaan atau jarak dihitung antara pasangan obyek dari variabel p. Selanjutnya algoritma pengelompokan dipilih yang mendefinisikan aturan mengenai bagaimana mengelompokkan obyek ke dalam *sub group* berdasarkan kesamaan antar obyek.

Adapun langkah-langkah analisis kelompok adalah sebagai berikut:

Langkah 1: Pengambilan p pengukuran variabel pada n obyek pengamatan

Langkah 2 : n obyek diukur pada variabel p.

|  |  |
| --- | --- |
|  | Variables |
|  |  | *X1* | *X2* | *X3* | *.* | *.* | *.* | Xp |
| **Objects** | O1 |  |  |  |  |  |  |  |
| O2 |  |  |  |  |  |  |  |
| O3 |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| On |  |  |  |  |  |  |  |

# Langkah 3: Transformasikan ke nxn kesamaan (jarak) matriks

|  |  |
| --- | --- |
|  | Objects |
|  |  | O1 | O2 | O3 | . | . | . | On |
| **Objects** | O1 |  |  |  |  |  |  |  |
| O2 |  |  |  |  |  |  |  |
| O3 |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| On |  |  |  |  |  |  |  |

Langkah 4: Formasi *cluster*

Langkah 5: Profil *cluster*

|  |  |  |
| --- | --- | --- |
|  |  | Variables |
|  |  | *X1* | *X2* | *X3* | *.* | *.* | *.* | Xp |
| **Clusters** | C1 |  |  |  |  |  |  |  |
| C2 |  |  |  |  |  |  |  |
| C3 |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| . |  |  |  |  |  |  |  |
| Ck |  |  |  |  |  |  |  |

Langkah 6&7: melakukan interpretasi terhadap profil *cluster* dan validasi model.

**A S U M S I M O D E L**

1. Data yang digunakan dapat berupa data interval, frekuensi, dan biner. Set data objek harus memiliki variabel dengan tipe yang sejenis, tidak campur antar tipe yang satu dengan lainnya.
2. Semua keputusan baik proses pembentukan cluster hingga interpretasi dan validasinya bergantung pada tujuan penelitian

**K O N S E P D A S A R**

* Konsep dasar penggunaan analisis kelompok adalah konsep pengukuran jarak (*distance*) dan kesamaan (*similarity*).
* *Distance* adalah ukuran tentang jarak pisah antar obyek, sedangkan *similarity* adalah ukuran kedekatan.
* Konsep ini penting karena pengelompokan pada analisis kelompok didasarkan pada kedekatan. Pengukuran jarak tersebut dapat dipisahkan di dua kelas berdasarkan kualitas data yang tersedia.
* Apabila datanya adalah berupa data metrik maka *distance-type measure* dapat digunakan. Sedangkan apabila berupa data yang berkualitatif lebih cocok menggunakan *matching-type measure*.

**DISTANCE-TYPE MEASURES**

Dalam *distance-type measures* kita mengasumsikan bahwa data dikumpulkan dari n obyek atau individu. Masing-masing obyek diwakili dengan vektor observasi X’ = (X1, X2, ..., Xp) pada p variabel. Notasi X’ = (X1, X2, ..., Xp) digunakan untuk menunjukkan ukuran yang dikumpulkan pada obyek ke-i. Beberapa ukuran jarak adalah :

* Kasus khusus dari **Minkowski** metrik didefinisikan dengan :



dimana  adalah jarak antara dua obyek i dan j dan k variabel

* ***Euclidian*** jarak antara obyek i dan j diperoleh kalau r=2



* ***Absolute or city-block*** metrik diperoleh kalau r=1



* **Mahalanobis** jarak D2, dihitung dengan:



Dimana  adalah *pooled within-group covariance* matriks, dan dan vektor masing-masing ukuran pada obyek i dan j.

**TEKNIK/ALGORITMA ANALISIS KELOMPOK**

Menurut Dillon, dua tipe teknik pengelompokan yang sering digunakan adalah:

1. *hierarchical techniques*, yang mengelompokan kelompoknya sendiri dengan membentuk konstruksi hierarki atau berdasarkan tingkatan tertentu

Proses pengelompokan dilakukan secara bertingkat dan bertahap. Pada teknik ini terdapat metoda *agglomerative* dan *divisive*:

* 1. Untuk metoda *agglomerative* dimulai dengan kenyataan bahwa setiap obyek membentuk *cluster*nya masing-masing. Kemudian, dua obyek dengan jarak terdekat bergabung. Selanjutnya obyek ketiga akan bergabung dengan *cluster* yang ada atau bersama obyek yang lain, membentuk *cluster* baru. Hal ini tetap memperhitungkan jarak kedekatan antar obyek. Proses akan terus berlanjut hingga akhirnya terbentuk satu *cluster* yang terdiri dari keseluruhan obyek.

Terdapat beberapa metoda yang termasuk ke dalam metoda *agglomerative*, yaitu sebagai berikut:

* *Single Linkage* atau *Nearest-Neighbor*

Metode ini menggunakan prinsip jarak minimum, yang diawali dengan mencari dua obyek yang memiliki jarak terdekat. Keduanya membentuk *cluster* yang pertama. Pada langkah selanjutnya terdapat 2 kemungkinan :

* + Obyek ketiga akan tergabung dengan *cluster* yang telah terbentuk
	+ Dua obyek lain akan membentuk *cluster* baru.

Proses ini akan terus berlanjut sampai akhirnya terbentuk *cluster* tunggal. Pada metoda ini, jarak antar *cluster* didefinisikan sebagai jarak terdekat antar anggotannya.

* *Complete Linkage* atau *Furthest Neighbor*

Metode ini merupakan kebalikan dari pendekatan yang digunakan pada single linkage. Dalam hal ini prinsip jarak yang digunakan adalah jarak terjauh antar obyek.

* *Average Linkage*

Metode ini mengikuti prosedur yang sama dengan kedua metoda sebelumnya. Prinsip ukuran jarak yang digunakan adalah jarak rata-rata antar tiap pasangan obyek yang mungkin.

* *Metode Ward’s Error Sum of Squares*

Ward mengajukan suatu metoda pembentukan *cluster* yang didasari oleh hilangnya informasi akibat penggabungan obyek menjadi *cluster*. Hal ini diukur dengan jumlah total dari deviasi kuadrat pada mean *cluster* untuk tiap observasi. *Error Sum of Squares* (ESS) digunakan sebagai fungsi obyektif. Dua obyek akan digabungkan apabila memiliki fungsi obyektif terkecil diantara kemungkinan yang ada.

ESS = 

 Dimana :  = nilai untuk obyek ke-i pada *cluster* ke-j

 k = jumlah total *cluster* pada tiap langkah

  = jumlah obyek pada *cluster* ke-j

1. *partitioning techniques*, yang membentuk kelompok dengan mengoptimasi beberapa kriteria pengelompokan. Metode dalam *partitioning techniques* adalah *K-means clustering* dan metode berdasarkan *trace*.

**I N T E R P R E T A S I**

* Pada tahap ini hasil pengelompokan berupa cluster-cluster diinterpretasikan sesuai nilai karakteristik yang terkandung dalam obyek-obyeknya. Biasanya interpretasi berdasarkan nilai rata-rata dari karakteristik obyek dalam cluster.
* Pada tahap ini yang perlu diperhatikan ialah karakteristik apa yang membedakan masing-masing cluster lalu sesuai dengan tujuan penulisan label apa yang dapat diberikan kepada masing-masing cluster tersebut.
* Interpretasi dapat dilakukan dari grafik dendogram maupun analisis koefisien aglomerasi
* Dendogram merupakan suatu diagram pohon yang menggambarkan penggabungan atau pemisahan obyek menjadi cluster dalam tiap tahap pemrosesan.

**V A L I D A S I**

Pada tahap ini dilakukan pengujian terhadap cluster yang telah dibentuk. Uji yang dapat dilakukan seperti membandingkan hasil yang diperoleh dengan algoritma yang berbeda. Misalnya bila pertama kita menggunakan algoritma hierarkikal maka dicoba menggunakan algoritma non hierarkikal apakah hasilnya mirip atau tidak. Bila hasilnya berbeda secara signifikan, maka cluster tersebut masih belum valid karena tidak dapat diterapkan secara umum. Pada prosedur pengelompokkan dengan melihat tingkat perubahan koefisien agglomerasi pada algoritma hierarkikal sebenarnya juga merupakan langkah uji terhadap cluster yang dibentuk.