Analisis Kelompok
(Cluster Analysis)

Sundari Mega Purnamasari (18209007)
Program Studi Sistem dan eknologi Informasi
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
sundari.mega@students.itb.ac.id

***Abstract*— *Cluster analysis* atau pengelompokan adalah teori mengenai serangkaian pengamatan pada himpunan bagian. *Clustering* merupakan teknik umum untuk analisis data statistik yang digunakan dalam berbagai bidang, termasuk *machine learning, data mining* (penggalian data) , pengenalan pola, analisis citra, dan bioinformatika. Metode ini juga tidak hanya mengelompokkan objek tetapi juga fitur dari objek tersebut. Pada makalah ini akan dipaparkan mengenai dasar teori dari metode ini dan bagaimana algoritma yang tepat dalam melakukan pengelompokkan ini serta seperti apa struktur bentukan yang terjadi dari hasil analisis data. Selain itu, pada makalah ini juga terdapat ilustrasi atau contoh permasalahan yang akan diselesaikan menggunakan metode *cluster analysis* ini. Permasalahan dari ilustrasi tersebut akan diselesaikan menggunakan formula-formula yang dipaparkan pada teori dasar dan dilakukan dengan pendekatan melalui metode hierarki dengan cara penggabungan dan pemecahan. Dengan begitu kita dapat mengetahui bagaimana implementasi dari formula dan algoritma pada metode ini sehingga didapat pemahaman dasar yang cukup untuk mengeksplorasi *cluster analysis* lebih jauh lagi.**

Kata kunci*— *cluster analysis, algoritma, hierarki, statistic

.

I. PENDAHULUAN

Cluster analysis adalah analisis statistika yang bertujuan untuk mengelompokkan data sedemikian sehingga data yang berada dalam kelompok yang sama mempunyai sifat yang relatif homogen daripada data yang berada dalam kelompok yang berbeda. Ditinjau dari hal-hal yang dikelompokkan, *cluster analysis* dibagi menjadi dua macam, yaitu :

1. Pengelompokkan observasi
2. Pengelompokkan variable

Secara umum, *cluster analysis* memiliki dua metode, yaitu :

1. Metode hierarki.

Metode ini digunakan untuk mencari struktur pengelompokkan dari objek-objek. Jadi, hasil pengelompokkannya disajikan secara hierarki atau berjenjang. Metode hierarki ini terdiri dari dua cara,yaitu :

a) *Agglomerative* (penggabungan).

Cara ini digunakan jika masing-masing objek dianggap satu kelompok kemudian antar kelompok yang jaraknya berdekatan bergabung menjadi satu kelompok.

b) *Divise* (pemecahan).

Cara ini dgunakan jika pada awalnya semua objek berada dalam satu gerombol. Setelah itu, sifat paling beda dipisahkan dan membentuk satu gerombol yang lain. Proses tersebut berlanjut sampai semua objek tersebut masing-masing membentuk satu gerombol.

2. Metode tak hierarki.

Metode ini digunakan apabila jumlah kelompok yang diinginkan diketahui dan biasanya dipakai untuk mengelompokkan data yang ukurannya besar.

II. TEORI DASAR

Dalam proses penggabungan kelompok dengan metode hierarki selalu diikuti dengan perbaikan matriks jarak. Suatu fungsi disebut jarak jika mempunyai sifat tak negative ($d_{ij} \geq 0$) dan ($d_{ij} = 0$) jika i = j, simetri ($d_{ij} = d_{ji}$), panjang salah satu sisi segitiga selalu lebih kecil atau sama dengan jumlah dua sisi yang lain ($d_{ij} \leq d_{ik} + d_{jk}$).

Beberapa macam jarak yang biasa dipakai di dalam analisis kelompok :

1. Jarak Euclidean
   Rumusnya:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \{x_{ik} - x_{jk}\}^2}$$

Sebuah tinjauan *cluster analysis* dalam penelitian kesehatan psikologi menemukan bahwa pengukuran jarak yang paling umum dalam penelitian adalah jarak Euclidian atau kuadrat jarak Euclidian.

2. Jarak Manhattan
   Rumusnya :

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

3. Jarak Pearson
   Rumusnya :

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{\mathrm{var}(x_k)}}$$

4. Jarak Korelasi
   Rumusnya :

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

5. Jarak Mutlak Korelasi
   Rumusnya :

$$d_{ij} = 1 - |r_{ij}|$$

Metode-metode pengelompokkan hierarki dibedakan berdasarkan konsep jarak antar kelompok, penentuan jarak antar kelompok untuk metode-metode tersebut adalah :

1. Metode *single linkage*

Metode ini menegelompokkan dua objek yang mempunyai jarak terdekat terlebih dahulu.
Jarak antar kelompok (i,j) dengan k adalah :

$$d_{(i,j)k} = \min(d_{ik}, d_{jk})$$

2. Metode *complete linkage*

Metode ini akan mengelompokkan dua objek yang mempunyai jarak terjauh terlebih dahulu.
Jarak antar kelompok (i,j) dengan k adalah :

$$d_{(i,j)k} = \max(d_{ik}, d_{jk})$$

3. Metode *average linkage*

Metode ini akan mengelompokkan objek berdasarkan jarak rata-rata yang didapat dengan melakukan rata-rata semua jarak objek terlebih dahulu.
Jarak antar kelompok (i,j) dengan k adalah :

$$d_{(i,j)k} = \text{average}(d_{ik}, d_{jk})$$

4. Metode *median linkage*

Pada metode ini, jarak antara dua *cluster* adalah jarak dia antara *centroid cluster* tersebut. *Centroid* adalah rata-rata jarak yang ada pada sebuah cluster yang didapat dengan melakukan rata-rata pada semua anggota suatu *cluster* tertentu. Dengan metode ini, setiap terjadi cluster baru, akan terjadi perhitungan ulang *centroid* hingga terbentuk *cluster* tetap.
Jarak antar kelompok (i,j) dengan k adalah :

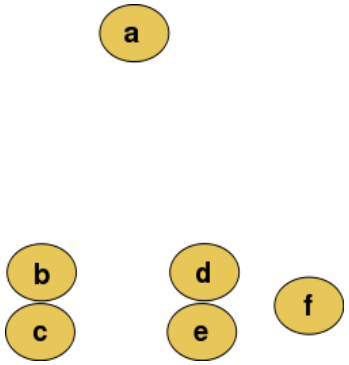
$$d_{(i,j)k} = \text{median}(d_{ik}, d_{jk})$$

Hasil dari analisis akan disajikan dalam bentuk struktur pohon yang disebut *dendogram*. Pemotongan *dendogram* dapat dilakukan pada selisih jarak penggabungan yang terbesar. Akar pohon terdiri dari *cluster* tunggal yang berisi semua pengamatan, dan daun sesuai dengan pengamatan individu.

Algoritma untuk mengelompokkan hierarki pada umumnya menggunakan cara *agglomerative*, yaitu dimulai dari daun dan secara berurutan menggabungkan *cluster* bersama, atau pemecahan yang dimulai dari akar dan dibagi secara rekursif.

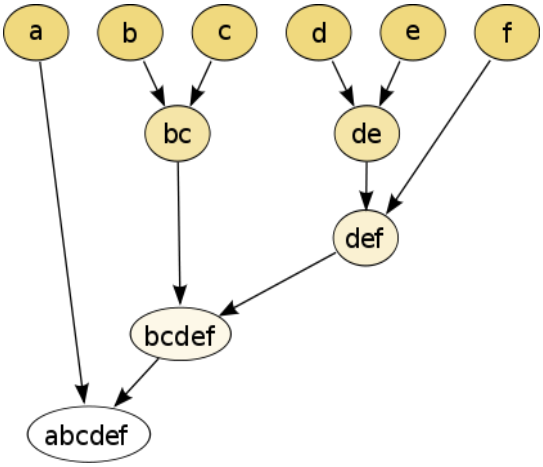
Pemotongan pada ketinggian tertentu akan memberikan *clustering* pada presisi yang dipilih. Sebagai contoh, pemotongan setelah baris kedua akan menghasilkan *cluster* {a}{bc}{de}{f}. Pemotongan setelah baris ketiga akan menghasilkan *cluster* {a}{bc}{def}, yang merupakan *clustering* kasar dengan sejumlah *cluster* yang lebih besar.

Untuk lebih jelasnya, akan diperlihatkan gambar sebagai berikut.



Gambar 1.

Lalu pengelompokan *dendogram* akan menjadi seperti gambar berikut.



Gambar 2

Dalam contoh ini, kita memiliki enam elemen, yaitu {a}{b}{c}{d}{e}{f}. Langkah pertama adalah menentukan elemen untuk menggabungkan sebuah *cluster*. Biasanya, dalam penggabungan ini diambil dua elemen terdekat sesuai dengan jarak yang dipilih.

Secara bebas kita juga dapat membuat matriks jarak pada tahap ini dengan angka dalam baris *ke-j* kolom *ke-i* adalah jarak antara *j dan* elemen *i.. K*emudian, setelah *clustering* berlangsung, baris dan kolom menjadi kelompok dengan jarak yang sudah diperbarui. Ini adalah cara yang umum untuk mengimplementasikan jenis *clustering* dan berguna untuk menyembunyikan jarak antara *cluster*.

Setiap algomerasi terjadi pada jarak antar cluster yang lebih besar daripada algomerasi sebelumnya, dan

clustering berhenti jika *cluster* terlalu jauh untuk digabung atau ketika ada jumlah angka *cluster* yang cukup kecil.

Metode yang merupakan metode tak hierarki adalah metode *k-means*. algoritma *k-means* memberikan poin pada *cluster* dengan pusat yang terdekat. Pusat ini adalah rata-rata dari semua titik dalam *cluster*.

Contohnya, kumpulan data memiliki tiga dimensi dan *cluster* ini memiliki dua titik :

X = (x1, x2, x3) dan Y = (y1, y2, y3). Kemudian Z centroid menjadi Z = (z1, z2, z3), dimana

$$z1 = \frac{x1+y1}{2}, \quad z2 = \frac{x2+y2}{2}, \quad \text{dan} \quad z3 = \frac{x3+y3}{2}$$

Keuntungan utama dari algoritma ini adalah kesederhanaan dan kecepatan yang memungkinkan untuk pengoperasian di dataset yang besar.

III. DATA DAN HASIL ANALISIS

Pada sub bab ini akan diberikan contoh permasalahan dari teori metode *cluster analysis*. Contoh permasalahan yang akan diberikan adalah contoh yang sederhana mengenai pengelompokan suatu himpunan bagian menjadi suatu kesatuan. Pengelompokan akan dilakukan berdasarkan sifat-sifatnya apakah setiap elemen ekivalen dengan elemen lainnya. Jika terdapat elemen yang memiliki kesamaan terdekat, maka elemen-elemen tersebut akan digabungkan dalam suatu kelompok.

Tahap-tahap pengelempokkan data dengan menggunakan metode hierarki adalah :

1. Tentukan matriks jarak antar data yang dikelompokkan.
2. Tentukan dua data yang mempunyai jarak terkecil kemudin gabungkan dua data ini ke dalam satu kelompok.
3. Modifikasi matriks jarak sesuai aturan jarak antar kelompok yang sesuai dengan metode pengelompokan yang dipakai.
4. Lakukan langkah 2 dan 3 samapai matriks jarak berukuran 1x1

Tahap-tahap pengelompokkan data dengan menggunakan metode tak hierarki *k-means* adalah :

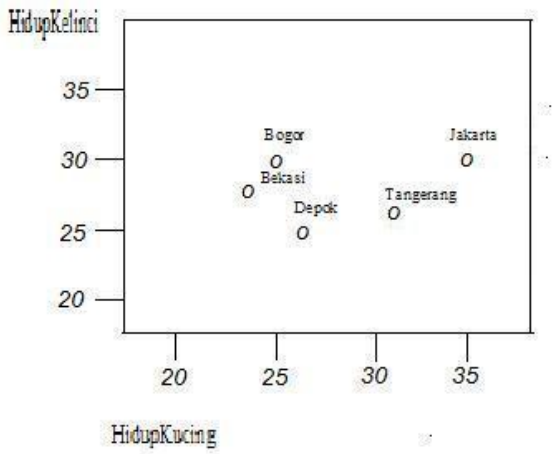
1. Mulai
2. Tentukan k buah pusat awal.
3. Tentukan jarak setiap data ke tiap pusat.
4. Lakukan pengelompokkan setiap data ke pusat terdekat.
5. Tentukan nilai pusat baru sebagai rata-rata data dalam kelompok.
6. Lakukan langkah 3-5 sampai nilai pusat kelompok tak berubah lagi.
7. Selesai

Dibawah ini akan diberikan tabel yang merupakan contoh ilustrasi dari *cluster analysis* menggunakan data harapan hidup kucing dan harapan hidup kelinci di kota jabodetabek.

	HidupKucing	HidupKelinci
Jakarta	30	35
Bogor	25	30
Depok	26	25
Tangerang	31	26
Bekasi	24	28

Tabel 1

Dari table diatas dapat dibuat diagram persebaran harapan hidup kucing dan harapan hidup kelinci di setiap kota seperti pada gambar berikut.



Gambar 3

Langkah awal analisis kelompok metode hierarki adalah membentuk matriks jarak antar observasi :
Menghitung matriks jarak berdasarkan kuadrat jarak Euclidian, misalkan d(Jakarta, Bogor) = $(30-25)^2 + (35-30)^2 = 50$

	Bogor	Depok	Tangerang	Bekasi
Jakarta	50	104	82	85
Bogor		26	52	**5**
Depok			26	13
Tangerang				53

Tabel 2

Setelah matriks jarak sudah terbentuk, maka langkah berikutnya adalah menentukan dua observasi yang mempunyai jarak terdekat dan kemudian digabung dalam satu kelompok. Dari tabel diatas dapat dilihat bahwa kota yang jaraknya paling dekat adalah Kota Bogor dan Bekasi. Oleh karena itu, pada tabel baru yang akan dibuat Kota Bogor dan Bekasi digabung dalam satu sel (satu kelompok).

	(Bgr, Bks)	Depok	Tangerang
Jakarta	…	104	82
(Bgr,Bks)		…	…
Depok			26

Tabel 3

Selanjutnya, dibuat tabel yang mengisi Jarak antara Jakarta dengan (Bogor, Bekasi) untuk berbagai metode.

Nomor	Metode	Jarak antara kelompok (Bgr,Bks) dengan Indonesia
1	*Single linkage*	min($d_{jak\ bgr}$, $d_{jak\ bks}$) = min (50, 85) = 50
2	*Complete linkage*	max($d_{jak\ bgr}$, $d_{jak\ bks}$) = max (50, 85) = 85
3	*Average linkage*	Average($d_{jak\ bgr}$, $d_{jak\ bks}$) = average(50, 85) = 67,5
4	*median linkage*	Median($d_{jak\ bgr}$, $d_{jak\ bks}$) = median(50, 85) = 67,5

Tabel 4

Kemudian pilih salah satu metode saja, misalkan kita menggunakan metode *single linkage* untuk semua *cluster*, maka akan diperoleh matriks jaraknya sebagai berikut.

	(Bgr, Bks)	Depok	Tangerang
Jakarta	50	104	82
(Bgr, Bks)		**13**	52
Depok			26

Tabel 5

Setelah tahap ini, dapat dilihat dari tabel diatas bahwa observasi yang mempunyai jarak paling dekat adalah Kota Depok dengan Bogor dan Bekasi, sehingga ketiga kota ini digabung seperti pada tabel berikut.

	(Bgr, Bks, Dpk)	Tangerang
Jakarta	50	82
(Bgr, Bks, Dpk)		**26**

Tabel 6

Pada tahap ini, jarak paling dekat adalah 26 sehingga Tangerang bergabung dengan kelompok (Bogor, Bekasi, Depok) sehingga matriks jarak berubah menjadi seperti pada tabel berikut.

	Jakarta
(Bgr, Bks, Dpk, Tang)	50

Tabel 7

Kemudian penggabungan terakhir adalah Jakarta dengan (Bogor, Bekasi, Depok, Tangerang) pada jarak penggabungan 50. Dengan begitu, dapat diperoleh pengelompokkan Kota-kota Jabodetabek dengan menggunakan metode *single linkage* adalah sebagai berikut.

Tahap	Jarak Penggabungan	Yang digabung		Banyak Kelompok	Kelompok
		cluster1	cluster2		
0	-	-	-	5	(Bgr) (Bks) (Jak) (Tan) (Dpk)
1	5	Bgr	Bks	4	(Bgr, Bks) (Dpk) (Jak) (Tan)
2	13	Bgr, Bks	Dpk	3	(Bgr, Bks, Dpk) (Jak) (Tan)
3	26	Bgr, Bks, Dpk	Tan	2	(Bgr, Bks, Dpk, Tan) (Jak)
4	50	Bgr, Bks, Dpk, Tan	Jak	1	(Bgr. Bks, Tan, Dpk, Jak)

Tabel 8

Berdasarkan kriteria loncatan, jarak penggabungan terbesar adalah jarak dari 26 ke 54. Oleh karena itu, dapat diketahui banyaknya kelompok adalah 2 yaitu (Bogor, Bekasi, Depok, tangerang) (jakarta).

Selain loncatan jarak penggabungan terbesar, banyaknya kelompok dapat juga ditentukan dengan beberapa kriteria, yaitu :

a) Maksimum nisbah (*ratio)* keragaman data antar kelompok dengan keragaman data di dalam kelompok. Statistik uji ini dapat dihitung melalui statistik uji F dalam *oneway anova* atau statistik uji Wilk dalam *oneway Manova*.

b) Maksimum statistik Hartigan(1975) :

$$H(k) = \left\{ \frac{W(k)}{W(k+1)} - 1 \right\} / (n - k - 1)$$

c) Maksimum rata-rata statistik *silhoutte* yang diajukan oleh Kaufman dan Rousseuw (1990)

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

Dengan a(i) adalah rata-rata jarak observasi ke-i dengan observasi yang lain dalam *cluster* yang sama dan

b(i) adalah rata- rata jarak observasi ke-i dengan *cluster* terdekat. Statistik ini disajikan oleh program SPLUS

Untuk pengelompokan variabel, banyaknya kelompok dapat ditentukan dengan menggunakan kriteria banyaknya nilai eigen yang lebih besar dari satu dari matriks korelasi.

IV. APLIKASI

Konsep dari metode ini cukup luas sehingga dapat diimplementasikan dan diaplikasikan dalam berbagai bidang seperti bidang ilmu biologi, kedokteran, penelitian pasar, pendididikan, analisis jaringan sosial, perangkat lunak, segmentasi gambar, optimasi peta *slippy*, pengelempokan produk, analisi kejahatan

a) Bidang ilmu Biologi

Penerapannya dalam ilmu biologi cukup banyak. *Pertama* adalah *imaging*. Dalam *imaging*, data clustering dapat mengambil bentuk yang berbeda berdasarkan dimensi data. Sebagai contoh, EM SOCR yang menunjukkan bagaimana mendapatkan titik, wilayah atau klasifikasi volume. *Kedua*, *clustering* digunakan untuk menggambarkan dan membuat perbandingan spasial dan temporal pada kumpulan organisme di lingkungan heterogen. *Ketiga*, dalam *transciptomik* digunakan untuk membangun gen dengan pola ekspresi terkait. *Keempat*, algoritma *clustering* dapat digunakan secara otomatis untuk menetapkan *genotype*.

b) Bidang kedokteran

Dalam pencitraan medis, *cluster analysis* dapat digunakan untuk membedakan berbagai jenis jaringan dan darah dalam gambar 3 dimensi. Contohnya adalah alat medis scan PET.

c) Bidang penelitian pasar

Cluster analysis banyak digunakan dalam riset pasar ketika bekerja dengan data *multivariate* dari survei dan planel uji. Peneliti pasar menggunakan analisis ini untuk pengelompokkan penduduk dari konsumen ke dalam segmen pasar dan untuk lebih memahami hubungan antara berbagai kelompok konsumen.

d) Bidang pendidikan

dalam analisis penelitian pendidikan, penggunaan data bisa untuk siswa, orang tua, jenis kelamin, atau skor tes. *Cluster analysis* dapat digunakan untuk eksplorasi data, dan pengujian hipotesis eksplorasi. Biasanya data digunakan ketika ada informasi mengenai sekolah atau siswa yang akan dikelomokkan secara bersama-sama.

e) Analisis jaringan sosial

Dalam studi jaringan sosial,*clustering* dapat digunakan untuk mengenali tiap individu orang dalam sekelempok besar orang.

f) Perangkat lunak

Clustering berguna dalam evolusi perangkat lunak karena membantu mengurangi sifat warisan dalam kode.

g) Segmentasi gambar

Clustering dapat digunakan untuk membagi sebuah gambar digital ke daerah yang berbeda untuk deteksi perbatasan dan pengenalan objek.

h) Optimasi peta *slippy*

Pada foto peta *flickr* dan peta situs lainnya digunakan *clustering* untuk mengurangi jumlah penanda pada peta. Hal ini dapat mengurangi jumlah kekacauan visual.

i) Pengelompokkan produk

Clustering digunakan untuk kelompok semua barang belanja yang tersedia di web menjadi serangkaian produk. Seperti contoh, *item* di eBay.

j) Analisis kejahatan

Cluster analysis dapat digunakan untuk mengidentifikasi daerah-daerah insiden besar terjadi ataupun kejahatan-kejahatn tertentu

V. KESIMPULAN

teori *cluster analysis* memiliki berbagai macam metode dengan cakupan yang luas sehingga terdapat banyak aplikasi dan implementasinya pada kehidupan sehari-hari . Pengimplementasian formula pada contoh permasalahan pun tidak terlalu sulit, hanya saja memerlukan ketelitian dalam perhitungan.

Cluster analysis juga memiliki kelebihan dan kekurangan. Keuntungan dari penggunaan metode ini adalah metode ini baik untuk meninjau pendataan dengan cepat, terutama jika benda tersebut diklasifikasikan ke dalam banyak kelompok. Namun kelemahannya adalah pada metode k-means diperlukan beberapa analisis sebelum jumlah cluster dapat ditentukan. Hal ini dapat sangat sensitive terhadap pilihan pusat awal *cluster*. Dalam beberapa tahun ini, telah banyak upaya dalam meningkatkan kinerja algoritma sehingga penggunaan metode ini menjadi efektif dan efisien.

VI. REFERENSI

1. http://en.wikipedia.org/wiki/Cluster_analysis
   waktu akses 15 Desember 2010

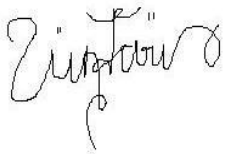
2. http://statistikaterapan.files.wordpress.com/2008/10/analisis-kelompok.doc
   waktu akses 15 Desember 2010

3. http://winnerstatistik.blogspot.com/.../analisis-gerombolcluster-analysis.html
   waktu akses 15 Desember 2010

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 29 April 2010



SUNDARI MEGA PURNAMASARI
18209007