CLUSTERING

Diajukan Untuk Memenuhi Salah Satu Tugas Mata Kuliah Analisis Multivariat

Disusun oleh:

Adinda Khalil. A (055851)

Chandra Aji (055452)

Egi Iriawan (055641)

Ratih Rahmawati (055495)

Rohimah (055608)

Jurusan Pendidikan Matematika

Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

Universitas Pendidikan Indonesia

2009

KATA PENGANTAR

Assalamualaikum, Wr.Wb

Segala puji bagi Allah SWT yang telah memberikan rahmat, ridho serta kasih sayangnya terhadap umat-Nya sehingga makalah yang berjudul "***CLUSTERING***" dapat terselesaikan tepat pada waktunya.

Makalah ini disusun sebagai salah satu tugas untuk mata kuliah Metode Statistika Multivariat. Penulis menyadari betul bahwa masih banyak terdapat kekurangan dalam bentuk penulisan makalah ini. Untuk itu adanya saran dan pendapat serta masukan-masukan yang membangun demi perbaikan makalah ini sangat penulis harapkan.

Pada kesempatan ini penulis mengucapkan terima kasih kepada Bapak Drs. Jarnawi M.kes yang telah membantu dan mendukung dalam pembuatan makalah ini.

Akhir kata, penulis berharap kiranya makalah ini dapat bermanfaat bagi perkembangan Ilmu Pengetahuan Matematika khusunya bidang Statistika sekarang dan pada masa yang akan datang.

Bandung, Juni 2009

Penulis

BAB I

PENDAHULUAN

1.1 Latar Belakang

Ketaksempurnaan, penyelidikan langkah-langkah sering membantu dalam pengertian hubungan multivariat kompleks. Untuk contoh, melalui buku ini kita tegaskan nilainya dari plot-plot data. Dibagian ini, akan didiskusikan beberapa teknik grafik tambahan dan diusulkan aturan langkah per langkah (algoritma) untuk pengelompokkan objek-objek (variabel-variabel atau bentuk-bentuk). Pencarian data untuk suatu struktur pada pengelompokan dasar adalah suatu teknik penyelidikan yang penting. Pengelompokkan-pengelompokkan dapat menentukan suatu makna-makna informal untuk penaksiran secara dimensi, pengidentifikasian pencilan, dan penyaranan dalam menarik hubungan pemusatan hipotesis.

Pengelompokkan (*grouping*) atau *clustering* berbeda dari metode pengklasifikasian yang didiskusikan pada bab sebelumnya. Pengklasifikasian menyinggung pada jumlah kelompok yang diketahui; dan secara operasionalnya objek yang memberikan satu pengamatan baru dari beberapa kelompok. Analisis *cluster* merupakan suatu teknik yang lebih sederhana bukan dalam asumsinya yang memusatkan jumlah kelompok-kelompok atau struktur kelompok. Pengelompokkan dilakukan pada kesamaan dasar atau jarak (ketaksamaan). Masukan-masukan yang dibutuhkan merupakan kesamaan ukuran atau data-data dari kesamaan-kesamaan yang dapat dihitung.

Penerapan praktis paling banyak pada analisis *cluster*, penyelidik cukup mengetahui masalah untuk membedakan pengelompokkan “baik” dan pengelompokkan “buruk”. Objek dasar dalam analisis *cluster* adalah untuk menemukan pengelompokkan dasar pada bentuk-bentuknya (variabel-variabel).

Dalam metode *clustering* terdapat metode yang digunakan yaitu metode *clustering* hirarki. Dalam metode ini, dilakukan *single cluster* dengan menggunakan prosedur *agglomerative dan divisive* yang dapat digambarkan dalam diagram dua dimensi yang dinamakan dendogram. Ini akan lebih fokus pada prosedur hirarki *agglomerative* dan bagiannya yaitu metode Linkage. Akan digunakan yaitu *single linkage* (jarak minimum atau tetangga terdekat), *complete linkage* (jarak maksimum atau tetangga terjauh), serta *average linkage* (jarak rata-rata).

Dalam *clustering* akan dilakukan *multidimensional scaling* suatu teknik pengurangan dimensi selain itu, juga akan dijelaskan pengambaran data-data dan representasinya.

1.2 Rumusan Masalah

Dalam uraian diatas maka dapat dibentuk rumusan masalah sebagai berikut:

- Bagaimana melakukan pengelompokkan data dengan menggunakan metode *clustering*?

1.3 Tujuan dan Maksud

Dari rumusan masalah di atas maka tujuan dan maksud dari presentasi ini adalah sebagai berikut:

- Memberikan penjelasan bagaimana menggelompokkan data dengan menggunakan metode *clustering*.

BAB II

ISI

2.1 Pendahuluan

Analisis *cluster* merupakan suatu teknik yang lebih sederhana bukan dalam asumsinya yang memusatkan jumlah kelompok-kelompok atau struktur kelompok. Pengelompokkan setuju pada kesamaan dasar atau jarak (ketaksamaan). Masukan-masukan yang dibutuhkan merupakan kesamaan ukuran atau data-data dari kesamaan-kesamaan yang dapat dihitung.

Untuk menggambarkan sifat yang sulit dalam pendefinisian suatu pengelompokkan dasar, misalnya pengurutan 16 kartu dalam permainan kartu biasa ke dalam *cluster* dari kesamaan objek-objek. Beberapa pengelompokkan digambarkan dalam gambar 12.1, ini dengan jelas bahwa maksud pembagian-pembagian tergantung pada pendefinisian kesamaan.

Untuk permainan kartu contohnya, terdapat satu cara membentuk suatu kelompok tunggal pada 16 kartu; terdapat 32.767 cara untuk membagi kartu ke dalam dua kelompok (bermacam-macam ukuran); terdapat 7.141.686 cara untuk mengurutkan kartu-kartu ke dalam tiga kelompok (bermacam-macam ukuran) dan seterusnya. Dengan jelas, batasan waktu membuat ini tidak mungkin untuk mennetukan pengelompokkan terbaik pada kesamaan objek-objek dari suatu daftar dari semua struktur yang mungkin. Meskipun komputer-komputer besar dengan mudah meliputi jumlah kasus yang besar. Jadi satu kasus menyelesaikan

pencarian algoritma yang baik, tetapi tidak memenuhi yang terbaik dalam pengelompokkan.

Kembali lagi, pertama harus dikembangkan suatu ukuran kuantitatif untuk assosiasi (kesamaan) ukuran antara objek-objek. Bagian 12.2 memberikan suatu pendiskusian pada kesamaan ukuran. Setelah bagian 12.2 dideskripsikan sedikitnya dari beberapa algoritma umum untuk pengurutan objek-objek ke dalam kelompok-kelompok.

Meskipun tanpa notasi yang tepat pada suatu pengelompokkan biasa, sering digunakan objek *cluster* dalam dua atau tiga dimensi *scatter plot*, memiliki keuntungan pada kemampuan pemikiran untuk mengelompokkan objek-objek yang sama dan untuk memilih pengamatan-pengamatan terpencil, langkah grafik secara umum baru-baru ini dikembangkan untuk penggambaran dimensi tingkat tinggi pengamatan-pengamatan dalam dua dimensi. Beberapa dari teknik langkahnya diberikan dalam bagian 12.5 dan 12.6.

2.2 Kesamaan Ukuran (*Similarity measures*)

Banyak usaha-usaha untuk langkah suatu struktur kelompok yang cukup sederhana dari suatu kumpulan data kompleks yang perlu suatu ukuran pada “pendekatan” atau “kesamaan”. Di sana sering terdapat ide bagus pada kesubjektifan termasuk dalam pemilihan dari suatu kesamaan ukuran. Anggapan-anggapan penting termasuk sifat dari variabel-variabelnya (diskrit, kontinu, biner) atau skala-skala pada pengukuran (nominal, ordinal, interval, rasio) dan subjek masalah keilmuan.

Karena bentuk-bentuk (satuan-satuan atau kasus-kasus) di *cluster*, didekatkan biasanya yang diindikasikan dengan beberapa urutan pada jarak. Di lain pihak, variabel-variabel biasanya dikelompokkan berdasarkan koefisien korelasi atau seperti ukuran assosiasi.

Jarak-jarak dan kesamaan koefisien-koefisien untuk pasangan bentuk-bentuk

Didiskusikan notasi jarak pada bab I, bagian 1.4, mengulang kembali jarak Euclid (garis lurus) antara dua pengamatan p-dimensi (bentuk-bentuk) $x = \left[x_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{p}\right]^t$ dan $y = \left[y_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{p}\right]^t$ adalah, dari (1-12)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^{\boldsymbol{t}}(\boldsymbol{x} - \boldsymbol{y})} \tag{12-1}$$

Jarak secara statistiknya antara dua pengamatan yang sama yaitu bentuknya, (lihat (1-22))

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^{\boldsymbol{t}}\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})} \tag{12-2}$$

Biasanya, $\boldsymbol{A} = \boldsymbol{S}^{-1}$ di mana *S* memuat variansi-kovariansi sampel. Bagaimana pun, tanpa ilmu sebelumnya pada perbedaan kelompok-kelompok, terdapat kuantitas sampel yang tak dapat dihitung. Untuk alasan ini jarak Euclid sering dilebihkan untuk *clustering*.

Ukuran jarak lainnya adalah metrik Minkowski (*Minkowski Metric*)

$$d(x, y) = \left[\sum_{i=1}^{p} [x_i - y_i]^m\right]^{1/m} \tag{12-3}$$

Untuk m = 1, *d(x,y)* mengukur jarak "*city-block*" antara dua titik dalam p-dimensi; untuk m = 2 , *d(x,y)* menjadi jarak Euclid. Umumnya, bermacam-macam mengubah bobotnya yang diketahui perbedaan lebih besar dan lebih kecil.

Di mana pun mungkin, ini dapat menjadi alat untuk menggunakan jarak "sesungguhnya", ini adalah jarak yang memenuhi sifat jarak pada (1-25) untuk objek *clustering*. Dilain pihak, banyak algoritma *clustering* akan menerima secara subjektif yang diberikan jumlah jarak yang mungkin tidak memenuhi, untuk contoh ketaksamaan segitiga.

Contoh 12.1: tabel 12.1 memberikan jarak Euclid antar pasangan pada 22 kegunaan perusahaan publik U.S yang berdasarkan pada datanya dalam tabel 12.5 setelah ini distandarisasikan.

Karena ukuran matriksnya besar, ini sulit untuk memvisualisasikan pilihan perusahaan-perusahaan yang mendekati bersama-sama (sama). Bagaimanapun, metode grafiknya dari *shading* mmberikan untuk penemuan *cluster* pada perusahaan-perusahaan yang sama secara mudah dan cepat.

Jarak pertama disusun ke dalam kelas-kelas umum (jelasnya, 15 atau lebih sedikit) yang berdasarkan pada besar atau jaraknya. Selanjutnya semua jarak antar suatu kelas yan diketahui diganti dengan suatu simbol yang umum dengan suatu perbedaan khusus. Simbol-simbol yang mengkorespondensikan untuk menutupi (*patches*) dari "*dark shading*".

Dari gambar 12.2 dilihat bahwa bentuk perusahaan 1, 18, 19 dan 14 sebuah kelompok; bentuk perusahaan 22, 10, 13, 20 dan 4 sebuah kelompok; bentuk perusahaan 9 dan 3 sebuah kelompok; bentuk perusahaan 3 dan 6 sebuagh

kelompok dan seterusnya. Kelompok (9, 3) dan (3, 6) saling melengkapi, begitu pula kelompok lain dalam diagramnya, perusahaan-perusahaan 11, 5 dan 17 kelihatan berdiri sendiri.

Karena bentuk-bentuknya tak dapat direpresentasikan secara berarti pengukuran p-dimensi, pasangan-pasangan pada bentuk-bentuk sering dibandingkan pada basisnya dari kemunculan atau takkemunculan pada karakteristik-karakteristik khususnya. Bentuk-bentuk yang sama lebih mempunyai karakteristik-karakteristik pada umumnya daripada bentuk-bentuk ketaksamaan. Kemunculan atau ketakmunculan dari suatu karakteristik dapat digambarkan secara matematik dengan pengenalan suatu variabel biner (*binary variable*), yang mengasumsikan nilai 1 jika karakteristiknya muncul dan nilai 0 jika karakteristiknya tak muncul. Untuk p = 5 variabel biner, untuk lebih jelasnya, nilai "*score*" variabelnya untuk dua bentuk i dan k mungkin disusun sebagai berikut,

	Variabel				
	1	2	3	4	5
Bentuk i	1	0	0	1	1
Bentuk k	1	1	0	1	0

Dalam kasus ini terdapat dua yang cocok dengan 1-1, satu yang cocok dengan 0-0 dan tidak cocok.

Misalkan x_{ij} nilainya menjadi (1 atau 0) dari variabel biner ke-j pada bentuk ke-i dan x_{kj} nilainya menjadi (1 atau 0) dari variabel ke-j pada bentuk ke-k, j = 1, 2, …, p. Konsekuensinya,

$$\left(x_{ij} - x_{kj}\right)^2 = \begin{cases} 0 & jika\ x_{ij} = x_{kj} = 1\ atau\ x_{ij} = x_{kj} = 0 \\ 1 & jika\ x_{ij} \neq x_{kj} \end{cases} \qquad (12\text{-}4)$$

Dan jarak kuadrat Euclid, $\sum_{j=1}^{p}\left(x_{ij} - x_{kj}\right)^2$ memberikan suatu perhitungan pada jumlah dari ketakcocokan. Suatu jarak besar mengkorespondensikan banyaknya ketakcocokan, ini berarti, bentuk-bentuk ketaksamaan. Dari pemaparan di atas, jarak kuadrat antara bentuk i dan k menjadi,

$$\sum_{j=1}^{k}\left(x_{ij} - x_{kj}\right)^2 = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 = 2$$

Meskipun suatu jarak berdasarkan pada (12-4) mungkin digunakan untuk ukuran yag sama, ini mendapatkan dari pembobotan yang sama 1-1 dan 0-0. Dalam beberapa kasus kecocokan 1-1 mengindikasikan lebih kuat dari kesamaan daripada 0-0. Untuk lebih jelasnya, ketika pengelompokkan orang-orang, keterangan bahwa dua orang keduanya membaca Yunani kuno lebih kuat keterangannya pada kesamaan daripada ketakmunculan pada kemampuan ini. Jadi ini mungkin beralasan untuk tak menghitung kecocokan 0-0 atau meskipun diabaikan secara kelengkapannya. Penyediaan untuk perbedaan perlakuan pada 1-1 dan 0-0, maksud umum untuk pendefinisian kesamaan koefisien yang diusulkan.

Untuk memperkenalkan maksud ini, misalkan disusun jumlah dari kecocokan dan takkecocokan untuk bentuk i dan k dalam bentuk tabel kontingensi berikut,

		Bentuk k		Total
		1	0	
Bentuk i	1	a	b	a + b
	0	c	d	c + d
Total		a + c	b+ d	p = a + b + c + d

(12-5)

Dalam tabel ini, a mempresentasikan jumlah 1-1, b adalah jumlah 1-0 dan seterusnya. Diketahui lima pasangan pada keluaran (*outcomes*) biner di atas, a = 2 dan b = c = d = 1. Tabel 12.2 memberikan kesamaan koefisien umum yang didefinisikan dalam bentuk-bentuk pada jumlah dalam (12-5). Sebuah alasan pemikiran yang diikuti beberapa definisi.

	koefisien	Pemikiran
1.	$\frac{a+d}{p}$	Bobot yang sama untuk 1-1 dan 0-0.
2.	$\frac{2(a+d)}{2(a+d)+b+c}$	Bobot *double* untuk 1-1 dan 0-0.
3.	$\frac{a+d}{a+d+2(b+c)}$	Bobot *double* untuk ketakcocokan.
4.	$\frac{a}{p}$	0-0 bukan dalam pembilang (*numerator*).
5.	$\frac{a}{a+b+c}$	0-0 bukan dalam pembilang (*numerator*)

		atau persamaan (*denominator*). (0-0 diperlakukan sebagai irrelevan).
6.	$\frac{2a}{2a+b+c}$	0-0 bukan dalam pembilang (*numerator*) atau persamaan (*denominator*). Bobot *double* untuk 1-1.
7.	$\frac{a}{a+2(b+c)}$	0-0 bukan dalam pembilang (*numerator*) atau persamaan (*denominator*). Bobot *double* untuk pasangan ketakcocokan.
8.	$\frac{a}{b+c}$	Rasio kecocokan untuk ketakcocokan dengan 0-0 dikeluarkan.

Koefisien 1, 2 dan 3 dalam tabel 12.2 memperoleh suatu hubungan monotonik "*monotonic*". Misalkan koefisien 1 dihitung untuk dua tabel kontingensi, tabel I dan tabel II. Maka jika

$$\frac{(a_I + d_I)}{p} \geq \frac{(a_{II} + d_{II})}{p}$$

dan juga

$$\frac{2(a_I + d_I)}{[2(a_I + d_I) + b_I + c_I]} \geq \frac{2(a_{II} + d_{II})}{[2(a_{II} + d_{II}) + b_{II} + c_{II}]}$$

Dan koefisien 3 paling tidak akan menjadi besar untuk tabel I seperti untuk tabel II. Koefisien 5, 6 dan 7 (tabel 12.2) juga menyimpan urutan kerelatifannya (lihat latihan 12.4).

Monotonitas "*monotonicity*" penting karena beberapa langkah *clustering* tak berpengaruh jika definisinya pada kesamaan diubah dalam suatu cara bahwa

lwmbaran pengurutan kerelatifannya pada kesamaan tak berubah. Langkah secara hirarki hubungan tunggal dan lengkap didiskusikan dalam bagian 12.3. Untuk metode-metodenya beberapa pilihan pada koefisien 1, 2 dan 3 (dalam tabel 12.2) langkah pengelompokkan yang sama. Dengan cara yan sama, beberapa pilihan pada koefisien-koefisien 5, 6, dan 7 hasil pengelompokkan identikal.

Contoh 12.2: Misalkan lima individu mempunyai karakteristik-karakteristik sebagai berikut,

	Tinggi (inci)	Berat (lb)	Warna mata	Warna rambut	*handedness*	Jenis kelamin
Individu 1	68	140	Hijau	Pirang	Kanan	Wanita
Individu 2	73	185	Coklat	Coklat	Kanan	Pria
Individu 3	67	165	Biru	Pirang	Kanan	Pria
Individu 4	64	120	Coklat	Coklat	Kanan	Wanita
Individu 5	76	210	Coklat	Coklat	Kiri	Pria

Didefinisikan enam variabel biner $X_1, X_2, X_3, X_4, X_5, X_6$ sebagai

$$X_1 = \begin{matrix} 1; tinggi \geq 72\ inci \\ 0; tinggi < 72\ inci \end{matrix} \qquad X_6 = \begin{matrix} 1; wanita \\ 0; pria \end{matrix}$$

$$X_2 = \begin{matrix} 1; berat \geq 150\ lb \\ 0; berat < 150\ lb \end{matrix}$$

$$X_3 = \begin{matrix} 1; mata\ coklat \\ 0; yang\ lainnya \end{matrix}$$

$$X_4 = \begin{matrix} 1; rambut\ pirang \\ 0; yang\ lainnya \end{matrix}$$

$$X_5 = \begin{matrix} 1; tangan\ kanan \\ 0; tangan\ kiri \end{matrix}$$

Nilai-nilai untuk individu 1 dan 2 pada p = 6 variabel biner adalah

		X_1	X_2	X_3	X_4	X_5	X_6
Individu	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

Dan jumlah kecocokan dan ketakcocokan diindikasikan dalam susunan dua cara,

		Individu 2		
		1	0	**Total**
Individu 1	1	1	2	3
	0	3	0	3
	Total	4	2	6

Kesamaan koefisien 1, yang memberikan bobot yang sama untuk kecocokan, dihitung

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}$$

Selanjutnya dengan kesamaan koefisien 1, dihitung sisa jumlah kesamaan untuk pasangan individu. Ditampilkan dalam matriks simetris berukuran 5 x 5,

Individu

individu	1	2	3	4	5
1	1				
2	1/6	1			
3	4/6	3/6	1		
4	4/6	3/6	2/6	1	
5	0	5/6	2/6	2/6	1

Berdasarkan pada besar atau jarak dari koefisiennya, dapat disimpulkan individu 2 dan 5 paling sama (serupa) dan individu 1 dan 5 paling sedikit sama. Beberapa pasangan berada antara keekstrimannya. Jika dibagi individu-individu ke dalam 2 sub kelompok yang sama relatif pada basisnya dari kesamaan jumlahnya, memungkinkan membentuk sub kelompoknya (1 3 4) dan (2 5).

Catatan bahwa $x_3 = 0$ memenuhi ketakmunculan secara kasat mata jadi, dua orang mempunyai pandangan yang berbeda, akan hasil 0-0. Konsekuensinya, ini mungkin tidak tepat untuk menggunakan kesamaan koefisien 1, 2 atau 3 karena koefisien-koefisiennya memberikan bobot yang sama unutk 1-1 dan 0-0.

Dideskripsikan konstruksi dari jarak dan kesamaannya. Ini selalu mungkin untuk mengkontruksikan kesamaan dari jarak. Untuk contoh, himpunan

$$\tilde{s}_{ik} = \frac{1}{1+d_{ik}} \qquad (12\text{-}6)$$

Di mana $0 < \tilde{s}_{ik} \leq 1$ adalah kesamaan antara bentuk i dan k dan d_{ik} mengkorespondensikan jarak.

Bagaimanapun, jarak-jarak harus memenuhi (1-25) tidak dapat selalu dikonstruksikan dari kesamaan-kesamaan. Gower [10, 11] telah menunjukkan, ini dapat berlaku jika matriks dari kesamaan-kesamaannya definit tak negatif, dengan keadaan definit tak negatif dan dengan skala kesamaan maksimum sedemikian hingga $\tilde{s}_{ii} = 1$.

$$d_{ik} = \sqrt{2(1 - \tilde{s}_{ik})} \qquad (12\text{-}7)$$

mempunyai sifat jarak.

Kesamaan dan Assosiasi Ukuran untuk Pasangan-Pasangan pada Variabel-variabel

Akan didiskusikan kesamaan ukuran untuk bentuk-bentuk yang di atas. Dalam beberapa penerapan, variabel-variabel yang harus dikelompokkan daripada bentuk-bentuknya. Kesamaan ukuran untuk variabel-variabel sering mengambil bentuk-bentuknya dari koefisien korelasi sampel. Selanjutnya, dalam beberapa penerapan *clustering*, korelasi-korelasi negatif diganti dengan memutlakkan nilainya.

Karena variabel-variabel biner, datanya dapat disusun kembali dalam bentuk suatu tabel kontingensi. Bagaimanapun, variabel-variabelnya, daripada bentuk-bentuknya, menggambarkan kategori-kategorinya. Untuk setiap pasangan pada variabel-variabel, terdapat n bentuk yang dikategorikan dalam tabel, dengan pengkodean yang biasa 0 dan 1, tabelnya menjadi sebagai berikut,

		Variabel k		Total
		1	0	
Variabel i	1	*a*	*b*	*a* + *b*
	0	*c*	*d*	*c* + *d*
Total		*a* + *c*	*b*+ *d*	*p* = *a* + *b* + *c* + *d*

(12-8)

Untuk lebih jelasnya variabel i sama dengan 1 dan variabel k sama dengan 0 untuk b pada n bentuk.

Perhitungan hasil korelasi momen yang biasa diterapkan ke variabel biner dalam tabel kontingensinya pada (12-8) memberikan (lihat latihan 12.3),

$$r = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{1/2}}$$

(12-9)

Bilangan ini dapat diambil sebagai suatu ukuran dari kesamaan antara dua variabel.

Koefisien korelasi dalam (12-9) direlasikan ke *chi*-kuadrat statistik $\left(r^2 = {\chi^2}/{n}\right)$ untuk pengujian kebebasan dari kategori dua variabel. Untuk n yang sudah ditetapkan, besarnya suatu kesamaan (atau korelasi) konsisten dengan ketidakbebasan.

Diketahui dalam tabel (12-8), ukuran dari assosiasi (atau kesamaan) secara tepat menganalogikan satu daftar dalam tabel 12.2 yang dapat dikembangkan. Hanya mengubah yang diperlukan yaitu pensubstitusian pada n (jumlah bentuk) dari p (jumlah variabel).

2.3 Hierarchical Clustering Methods (Metode Pengelompokan Hierarki)

Tidak semua kemungkinan dalam pengelompokan (*clustering*) dapat diselidiki secara keseluruhan, meski dengan media penghitung tercepat dan terbesar. Oleh karena itu, berbagai variasi dari algoritma *clustering* muncul sehingga dapat menemukan kelompok yang cocok tanpa menyelidiki semua bentuk yang mungkin.

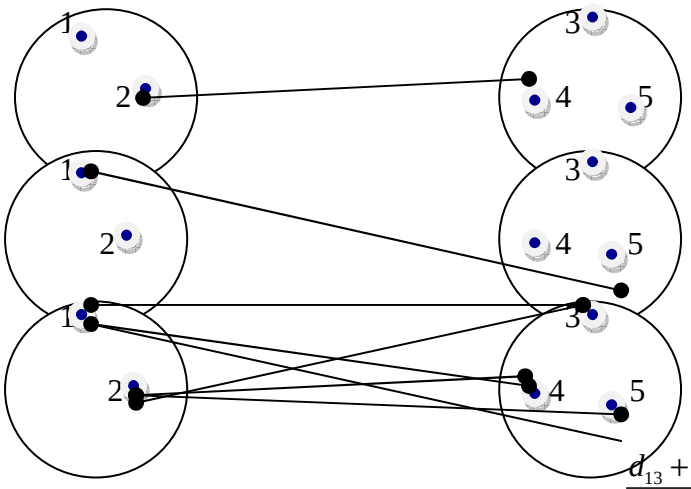
Teknik *hierarchical clusterin)* yang dapat digunakan antara lain deret gabungan yang berturut-turut (*series of successive mergers*) dan deret bagian yang berturut-turut (*series of successive divisions*). Metode hirarki aglomeratif berawal dari objek individual. Dengan demikian akan terdapat proses awal sebanyak objek *cluster* (kelompok). Objek-objek yang paling banyak memiliki kesamaan adalah yang pertama dikelompokkan, dan ini sebagai grup awal. Akan tetapi, seiring berkurangnya kesamaan di antara objek-objeknya, maka semua subgroup tergabung dalam suatu kelompok tunggal *single cluster*.

Metode hirarki yang terbagi (*divisive hierarchical methods*) bekerja pada arah yang berlawanan. Objek-objek dalam grup tunggal awal terbagi menjadi dua subgrup dimana objek-objek pada satu subgroup terletak jauh dari objek-objek pada subgroup yang lain. Kedua subgroup ini kemudian dibagi atas subgroup-subgrup yang tidak sama. Proses ini berlanjut hingga terdapat banyak subgroup sebanyak objek, yakni hingga setiap objek membentuk sebuah grup.

Hasil dari kedua metode (*agglomerative dan divisive*) dapat digambarkan dalam diagram dua dimensi yang dinamakan dendogram. Dendogram

mengilustrasikan penggabungan ataupun pembagian yang telah dibuat pada proses *successive* (berturut-turut).

Pada bagian ini akan lebih fokus pada prosedur hirarki agglomerative dan bagiannya yaitu metode Linkage. Metode Linkage cocok untuk item *clustering*, sebagaimana variabel. Namun hal ini tidak untuk semua prosedur hirarki agglomerative. Harus diperhatikan beberapa kemungkinan yaitu *single linkage* (jarak minimum atau tetangga terdekat), *complete linkage* (jarak maksimum atau tetangga terjauh), serta *average linkage* (jarak rata-rata). Gabungan dari kelompok-kelompok dengan tiga kriteria linkage diilustrasikan sebagai berikut:



Jarak Kelompok

d_{24}

d_{15}

$$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$$

Dari gambar di atas dapat dilihat bahwa hasil *single linkage* ketika grup tergabung berdasarkan jarak antara anggota-anggota yang terdekat. *Complete linkage* terjadi ketika grup tergabung berdasarkan jarak antar anggotanya yang palin berjauhan. Sedangkan untuk *average linkage*, grup tergabung berdasarkan jarak rata-rata antara pasangan anggota-anggotanya dalam maing-masing himpunan.

Berikut adalah langkah-langkah dalam algoritma pengelompokan hirarki agglomeratif (*agglomerative hierarchical clustering algorithm*) untuk mengelompokkan N objek (bagian atau variabel):

1. Dimulai dengan N kelopmpok, masing-masing mengandung kesatuan yang tunggal dan matriks simetris N x N dari jarak (kesamaan), $D = \{d_{ik}\}$
2. Dicari matriks jarak untuk pasangan kelompok terdekat (yang paling banyak kesamaan). Dimisalkan jarak antara kelompok U dan V yang paling sama dinotasikan dengan d_{uv}.
3. Gabungkan kelompok U dan V. Gabungan tersebut dinotasikan dengan (UV). Letakkan objek pada matriks jarak dengan:
   a. menghapus baris dan kolom yang berkorespondensi dengan kelompok U dan V
   b. menambahkan baris dan kolom yang terdapat jarak antara kelompok (UV) dan kelompok yang tertinggal.
4. Ulangi langkah 2 dan 3 sebanyak N-1 kali. (Semua objek akan berada pada single cluster saat olgoritma terakhir). Catat identitas dari cluster yang tergabung dan levelnya (jarak atau kesamaannya) dimana gabungannya ditempatkan.

(12-10)

Single Linkage

Input pada algoritma *single linkage* dapat berupa jarak atau kesamaan antara pasangan-pasangan objek. Grup dibentuk dari kesatauan individu dengan

menggabungkan tetangga terdekatnya, dimana kata “tetangga terdekat” mengandung arti jarak terkecil atau kesamaan terbesar (terbanyak).

Sebagai langkah awal kita harus menemukan jarak terkecil pada $D = \{d_{ik}\}$ dan menggabungkan objek-objek yang saling berkorespondensi, katakanlah U dan V, untuk mendapatkan kelompok (UV). Untuk langkah ketiga pada algoritma umum (12-10), jarak antara di antara (UV) dan kelompok yang lainnya, katakanlah W, dihitung dengan cara

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$$

Di sini, nilai d_{UW} dan d_{VW} adalah jarak antara tetangga terdekat dari kelompok U dan W serta kelompok V dan W, begitupun sebaliknya .

Hasil dari pengelompokan *single linkage* dapat digambarkan secara grafis melalui dendogram atau diagram pohon. Cabang-cabang pada pohon melambangkan kelompok (*clusters*). Cabang-cabang tersebut tergabung pada poros *node* (simpul) yang posisinya sepanjang jarak (atau kesamaan) yang menunjukkan level dimana gabungan terjadi.

Dendogram untuk beberapa kasus spesifik diilustrasikan pada contoh-contoh sebagai berikut:

Contoh 1

Untuk mengilustrasikan algoritma single linkage, kita misalkan jarak antara pasangan dari lima objek diduga sebagai berikut:

$$D = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

Perlakukan setiap objek sebagai kelompok (cluster), pengelompokan (clustering) dimulai dengan menggabungkan dua item terdekat. Sehingga

$$\min_{i,k}(d_{ik}) = d_{53} = 2$$

Objek 5 dan 3 digabungkan untuk membentuk kelompok (35). Alat untuk level selanjutnya dalam pengelompokan ini adalah dibutuhkan jarak antara kelompok (35) dan objek sisa, 1, 2, 3 dan 4. Jarak tetangga terdekat adalah

$$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3,11\} = 3$$
$$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7,10\} = 7$$
$$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9,8\} = 8$$

Hapus baris dan kolom dari D yang bekorespondensi dengan objek # dan 5 dan tambahkan baris dan kolom untuk kelompok (35), maka diperoleh matriks jarak yang baru berikut

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Jarak terkecil antara pasangan-pasangan cluster (kelompok) sekarang adalah $d_{(35)1} = 3$ dan gabungkan kelompok (1) dengan kelompok (35) untuk mendapatkan kelompok berikutnya. Kemudian dihitung

$$d_{(135)2} = \min\left\{d_{(35)2}, d_{12}\right\} = \min\{7, 9\} = 7$$
$$d_{(135)4} = \min\left\{d_{(35)4}, d_{14}\right\} = \min\{8, 6\} = 6$$

Matriks jarak untuk pengelompokan pada level selanjutnya adalah

$$\begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

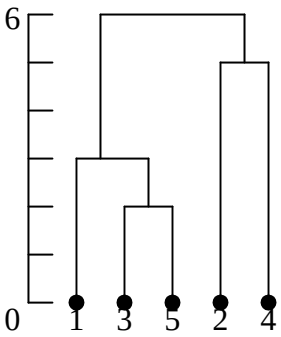
Jarak minimum tetangga terdekat antara pasangan-pasangan kelompok adalah $d_{42} = 5$, dan kemudian gabungkan objek 4 dan 2 untuk mendapatkan kelompok (24).

Pada titik ini diperoleh dua kelompok yang berbeda, (135) dan (24). Jarak tetangga terdekatnya adalah $d_{(135)(24)} = \min\left\{d_{(135)2}, d_{(135)4}\right\} = \min\{7, 6\} = 6$

Maka matriks jarak terakhir yang diperoleh adalah

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ \boxed{6} & 0 \end{bmatrix} \end{array}$$

Akibatnya, kelompok (135) dan (24) tergabung untuk membentuk single cluster (kelompok tunggal) dari kelima objek, (12345), dimana jarak tetangga terdekatnya adalah 6.



Gambar 12.4

Dendogram di atas menggambarkan pengelompokan hirarki (hierarchical clustering) telah disimpulkan. Pengelompokan, dan level jarak yang terjadi, diiliustrasikan melalui dendogram tersebut.

Contoh 2

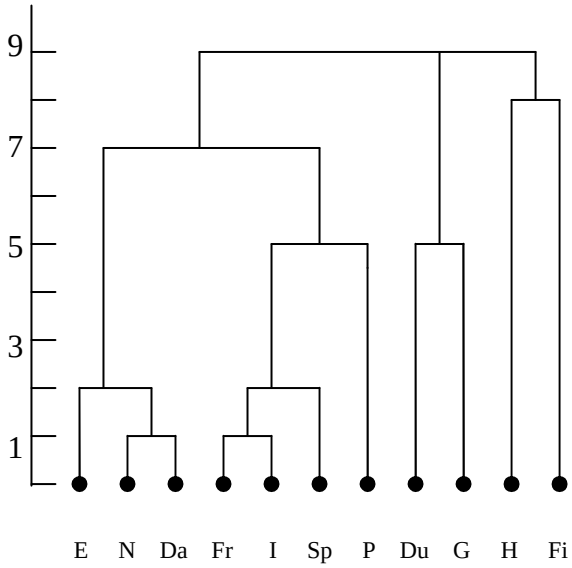
Misalkan barisan persetujuan pada tabel 12.4 menunjukkan kedekatan antara nomor 1-10 dalam 11 bahasa. Untuk mengembangkan matriks jaraknya, kita mendasarkan persetujuan dari gambar persetujuan yang sempurna dari 10, dimana setiap bahasa memiliki karakteristik masing-masing. Jarak selanjutnya adalah sebagai berikut:

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	0										
N	2	0									
Da	2	1	0								
Du	7	5	6	0							
G	6	4	5	5	0						
Fr	6	6	6	9	7	0					
Sp	6	6	5	9	7	2	0				
I	6	6	5	9	7	1	1	0			
P	7	7	6	10	8	5	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
Fi	9	9	9	9	9	9	9	9	9	8	0

Langkah pertama adalah mencari jarak minimum antara pasangan bahasa (kelompok). Jarak minimum adalah 1, terjadi antara bahasa Denmark dan Jerman, Italia dan Perancis, serta Italia dan Spanyol. Penomoran bahasa dimana hal ini muncul melintasi puncak barisan, diperoleh

$d_{32} = 1;$ $d_{86} = 1;$ dan $d_{87} = 1$

Dengan $d_{76} = 2$ maka yang dapat digabungkan hanya kelompok 8 dan 7 atau kelompok 8 dan 7. Sedangkan kelompok 6, 7, dan 8 pada level 1 tidak dapat digabungkan. Pertama, dipilih untuk menggabungkan 8 dan 6, kemudian mengentri matriks jarak dan menggabungkan 2 dan 3 untuk memperoleh kelompok (68) dan (23). Penghitungan di atas menghasilkan dendogram sebagai berikut:



Gambar 12.5

Dari dendogram dapat dilihat bahwa bahasa Norwegia dan Denmark dan juga Perancis dan Italia, tergabung berdasarkan jarak minimum (kesamaan maksimum). Ketika kemungkinan jarak meningkat, bahasa Inggris ditambahkan

ke grup Norwegia-Denmark dan Spanyol tergabung dengan grup Perancis-Italia.Perhatikan bahwa Hongariadan Finlandia lebih banyak kesamaan diantara keduanya dibanding kelompok bahasa lainnya. Akan tetapi, dua kelompok bahasa ini tidak tergabung sampai jarak di antara tetangga terdekatnya meningkat sepenuhnya. Pada akhirnya, semua kelompok bahasa tergabung dalam single cluster (kelompok tunggal) dengan tetangga terdekat yang terbesar yaitu 9.

Complete Linkage

Prosedur pengelompokan *complete-linkage* hampir sama dengan *single linkage*, dengan satu pengecualian. Pada setiap tingkat, jarak (kesamaan) antar kelompok ditentukan dengan jarak (kesamaan) anatara dua elemen, satu dari setiap kelompok, yakni yang paling jauh. Dengan demikian *complete linkage* menjamin bahwa dalam seluruh item pada kelompok terdapat jarak maksimum (atau kesamaan minimum).

Algoritma aglomeratif umum dimulai dengan menemukan entri (elemen) dalam $D=\{d_{ik}\}$ dan menggabungkan objek yang berkorespondensi, misalkan U dan V, untuk membentuk kelompok (UV). Pada langkah ketiga dalam algoritma umum (12-10), jarak antara (UV) dan kelompok lainnya, misalkan W ditentukan sebagai berikut:

$$d_{(uv)w} = \max\{d_{uw}, d_{vw}\}$$

Dimana d_{uw} dan d_{vw} merupakan jarak terjauh antara anggota kelompok U dan W serta kelompok V dan W, begitupun sebaliknya.

Contoh 3

Misalkan matriks jarak berikut adalah matriks jarak pada Contoh 1. Dalam kasus ini

$$D=\{d_{ik}\}=\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

Pada tingkatan pertama, objek 3 dan 5 tergabung jika diantaranya paling banyak kesamaan. Hal ini menghasilkan kelompok (35). Pada tingkatan kedua, dapat dihitung

$$d_{(35)1}=\max\{d_{31},d_{51}\}=\max\{3,11\}=11$$
$$d_{(35)2}=\max\{d_{32},d_{52}\}=\max\{7,10\}=10$$
$$d_{(35)4}=\max\{d_{34},d_{54}\}=\max\{9,8\}=9$$

dan matriks jarak yang dimodifikasi sebagai berikut:

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & \boxed{5} & 0 \end{bmatrix} \end{array}$$

Penggabungan selanjutnya terjadi antara grup palig sama, 2 dan 4, untuk membentuk kelompok (24). Pada tingkatan ketiga diperoleh

$$d_{(24)(35)}=\max\left\{d_{2(35)},d_{4(35)}\right\}=\max\{10,9\}=10$$
$$d_{(24)1}=\max\{d_{21},d_{41}\}=9$$

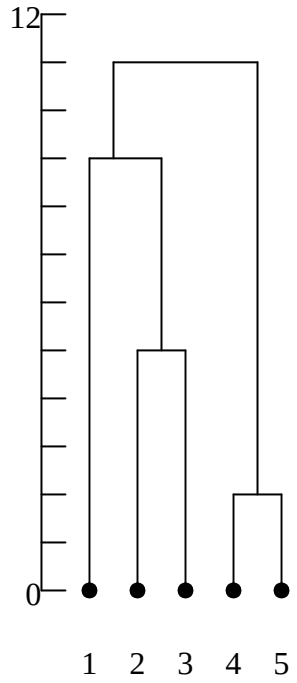
dan matriks jaraknya sebagai berikut:

$$\begin{array}{c} \\ (35) \\ (24) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & \boxed{9} & 0 \end{bmatrix} \end{array}$$

Penggabungan berikutnya membentuk kelompok (124). Pada tingkatan akhir, kelompok (35) dan (124) tergabung dalam kelompok tunggal (single cluster) (12345) pada level

$$d_{(124)(35)} = \max\left\{d_{1(135)}, d_{(24)(35)}\right\} = \max\{11,10\} = 11.$$

Dendogram dari kasus ini adalah sebagai berikut:



Gambar 12.7

Average Linkage

Average Linkage didasarkan pada rata-rata jarak dari seluruh objek pada suatu cluster dengan seluruh objek pada cluster lain. Algoritma yang digunakan dalam Average Linkage hampir sama dengan algoritma agglomerative hierarchical clustering. Kita mulai dengan mencari jarak dari matrik $D=\{d_{ik}\}$

Untuk mencari objek terdekat, sebagai contoh U dan V, objek ini digabung ke dalam bentuk cluster (UV). Untuk tahap ketiga, jarak antara (UV) dan cluster W adalah:

$$d_{(UV)W}=\frac{\sum_i\sum_k d_{ik}}{N_{(UV)}N_W}$$

Dimana d_{ik} adalah jarak antara objek I pada cluster (UV) dan objek k pada cluster W, dan $N_{(UV)}$ dan N_W adalah jumlah dari item-item pada cluster (UV) dan W.

Contoh:

Misalkan kita ambil matrik di contoh 12.4

$$\mathbf{D}=\{d_{ik}\}=\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & (2) & 8 & 0 \end{bmatrix} \end{array}$$

- Pertama kita cari jarak min, yaitu $\min_{i,k}\{d_{ik}\}=d_{53}=2$

- Objek 5 dan 3 di gabung ke bentuk cluter (35). Lalu akan dicari jarak antara cluster (35) terhadap 1, 2, dan 4.

$$d_{(35)1} = \frac{d_{31} + d_{51}}{2} = \frac{3+11}{2} = 7$$
$$d_{(35)2} = \frac{d_{32} + d_{52}}{2} = \frac{7+10}{2} = \frac{17}{2}$$
$$d_{(35)4} = \frac{d_{34} + d_{54}}{2} = \frac{9+8}{2} = \frac{17}{2}$$

- Dengan menghapus baris dan kolom dari matrik korespondensi **D** terhadap objek 3 dan 5 dan dengan menambahkan baris dan kolom untuk cluster (35), kita akan memperoleh matrik baru.

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 7 & 0 & & \\ 17/2 & 9 & 0 & \\ 17/2 & 6 & (5) & 0 \end{bmatrix} \end{array}$$

- Penggabungan berikutnya adalah antara 2 dan 4,

$$d_{(24)(35)} = \frac{d_{2(35)} + d_{4(35)}}{2} = \frac{\frac{17}{2} + \frac{17}{2}}{2} = \frac{17}{2}$$
$$d_{(24)1} = \frac{d_{21} + d_{41}}{2} = \frac{9+6}{2} = \frac{15}{2}$$

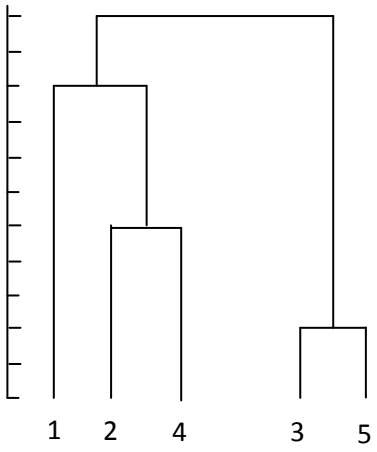
Dan matrik jaraknya

$$\begin{array}{c} \\ (35) \\ (24) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 17/2 & (15/2) & 0 \end{bmatrix} \end{array}$$

- Penggabungan berikutnya menghasilkan cluster (124). Pada tahap terakhir, grup (35) dan (124) akan digabung pada cluster tunggal (12345) dimana

$$d_{(124)(35)} = \frac{d_{1(35)} + d_{(24)(35)}}{2} = \frac{7 + \frac{17}{2}}{2} = \frac{31}{4}$$

- Dendogram nya adalah sebagai berikut:



2.4 Metode Pengelompokkan Nonhierarchical

Tipe Clustering

- Metode pengelompokan pada dasarnya ada dua, yaitu Hierarchical Clustering Method) dan Non Hierarchical Clustering Method).
- Metode pengelompokan hirarki digunakan apabila belum ada informasi jumlah kelompok. Sedangkan metode pengelompokan Non Hirarki bertujuan mengelompokan *n* obyek ke dalam *k* kelompok *(k < n).*

- Salah satu prosedur pengelompokan pada non hirarki adalah dengan menggunakan metode K-Means.

Metode *K-means*

Metode ini merupakan metode pengelompokan yang bertujuan mengelompokan obyek sedemikian hingga jarak tiap-tiap obyek ke pusat kelompok di dalam satu kelompok adalah minimum.

Algoritma *K-Means*

1. Tentukan Jumlah *K* cluster.
2. Cari data yang lebih dekat dengan pusat cluster. Hitung jarak Euclidean masing-masing item dari pusat cluster. Tentukan kembali pusat cluster.
3. Ulangi langkah 2 sampai tidak ada yang berpindah posisi.

Contoh 12.11

Misalkan kita mempunyai dua variable X_1 dan X_2, dan masing-masing terdiri dari item *A, B, C, D*. data nya adalah sebagai berikut.

item	observation	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Objek-objek diatas akan dibagi kedalam $K = 2$ cluster. Dengan Metode $K = 2$-*means* kita akan mempartisi kedalam dua cluster, misalkan (AB) dan (CD), koordinat dari pusat cluster (rata-rata) adalah sebagai berikut:

cluster	koordinat pusat	
	$\overline{x_1}$	$\overline{x_1}$
(*AB)*	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(*CD*)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

Pada tahap kedua, kita menghitung jarak Euclidean masing-masing item dari grup pusat dan kembali menentukan item ke grup terdekat. Jika item dipindahkan dari posisi awal, pusat cluster harus diperbarui sebelum diproses. Jarak kuadratnya adalah sebagai berikut:

$d^2(A, (AB)) = (5\text{-}2)^2 + (3\text{-}2)^2 = 10$

$d^2(A, (CD)) = (5+1)^2 + (3+2)^2 = 61$

karena *A* terdekat terhadap cluster (*AB*) daripada cluster (*CD*), proses berlanjut.

$d^2(B, (AB)) = (-1\text{-}2)^2 + (1\text{-}2)^2 = 10$

$d^2(B, (CD)) = (-1+1)^2 + (1+2)^2 = 9$

akibatnya, *B* kembali ditentukan terhadap cluster (*CD*) sehingga diberikan cluster (*BCD*) dan koordinat pusat yang baru adalah:

cluster	Cordinat pusat	
	$\overline{x_1}$	$\overline{x_1}$
A	5	3
(*BCD*)	-1	-1

Kemudian masing-masing item di cek kembali. Hasil penghitungan jarak kuadrat adalah sebagai berikut:

cluster	Jarak kuadrat terhadap pusat-pusat grup			
	item			
	A	*B*	*C*	*D*
A	0	40	41	89
(*BCD*)	52	4	5	5

masing-masing item telah ditentukan terhadap cluster dengan pusat terdekat dan proses dihentikan. Akhirnya, $K = 2$ cluaster adalah *A* dan (*BCD*).

2.5 Multidimensional Scaling

Teknik *multidimensional scaling* digunakan pada permasalahan berikut : untuk kesamaan(jarak) himpunan obsevasi antara setiap pasangan sebanyak N item, temukan gambaran dari item tersebut dalam dimensi yang sedikit sedemikian sehingga kedekatan antar item “hampir sesuai *(nearly match)* dengan jarak aslinya.

Hal ini sangatlah mungkin untuk menyesuaikan secara tepat urutan jarak asli. Akibatnya, teknik scaling ini mencoba untuk menemukan susunan dalam $q \leq N-1$ dimensi sedemikian sehingga kecocokannya sedekat mungkin. Ukuran numerik kedekatan tersebut dinamakan stress.

Kemungkinan untuk menyusun sebanyak N item dalam dimensi yang rendah dalam suatu koordinat system hanya dengan menggunakan urutan tingkatan dari $N(N-1)/2$ jarak aslinya dan bukan *magnitudes-nya* (besarnya). Ketika informasi ordinal (nomor urutan) digunakan untuk memperoleh gambaran secara geometris, maka prosesnya disebut dengan *nonmetric multidimensional scalling*. Jika *magnitudes* sebenarnya dari jarak asli digunakan untuk memperoleh gambaran dalam q-dimensi, maka prosesnya dinamakan *metric multidimensional scalling*.

Teknik *scaling* ini dibangun oleh Shepard (lihat [18] untuk kilas balik dari pekerjaan pertama), Kruskal [14,15,16] dan lain-lain. Ringkasan sejarah, teori dan aplikasi multidimensional scaling tercakup dalam [22]. Didalam multidimensional scaling selalu menggunakan computer, dan beberapa program computer yang menyediakan untuk tujuan ini.

Algoritma Dasar

Untuk N item, maka terdapat $M = N(N-1)/2$ kesamaan (jarak) antara pasangan item yang berbeda. Jarak ini merupakan data utama. (dalam kasus dimana kesamaannya tidak dapat dengan mudah diukur, contohnya kesamaan antara dua warna, urutan tingkatan dari suatu kesamaan merupakan data utama).

Asumsikan *no ties*, maka kesamaannya dapat disusun dalam urutan yang meningkat sebagai

$$s_{i_1k_1} < s_{i_2k_2} < \cdots < s_{1_M\,k_M} \qquad (12\text{-}15)$$

Disini $s_{i_1k_1}$ adalah M kesamaan terkecil. Sedangkan subscript i_1k_1 menunjukan pasangan item yang paling sedikit sama ; yaitu item dengan rank 1 dalam urutan kesamaan. Begitupun dengan *subscript* yang lain. Misalkan kita ingin menemukan susunan dalam q-dimensi dari N item sedemikian sehingga jarak, $d_{ik}^{(q)}$, antar pasangan sesuai dengan urutan dalam persamaan (12-15). Jika jaraknya dibuat dalam cara yang berkorespondensi dengan persamaan (12-15), maka kesesuaian yang sempurna terjadi ketika

$$d_{i_1k_1}^{(q)} > d_{i_2k_2}^{(q)} > \cdots > d_{i_Mk_M}^{(q)} \qquad (12\text{-}16)$$

Yakni, urutan menurun dari jarak dalam q-dimensi secara tepat menganalogikan dengan susunan yang meningkat dari kesamaan awal. Sepanjang urutan dalam persamaan (12-16) dipertahankan, magnitude (besar) tidaklah penting.

Untuk nilai q yang diberikan, tidaklah mungkin untuk menemukan susunan titik-titik yang jarak pasangannya dihubungkan secara monoton dengan kesamaan aslinya. Kruskal (14) mengemukakan ukuran kedekatan (stress) yang didefinisikan sebagai :

$$\text{Stress (q)} = \left\{ \frac{\sum\sum_{i<k} \left(d_{ik}^{(q)} - \hat{d}_{ik}^{q}\right)^2}{\sum\sum_{i<k} \left[d_{ik}^{(q)}\right]^2} \right\}^{1/2} \qquad (12\text{-}17)$$

$\hat{d}_{ik}^{q}$ dalam rumus di atas adalah jumlah yang tidak diketahui untuk memenuhi persamaan (12-16); yaitu kesamaan yang dihubungkan secara monoton. $\hat{d}_{ik}^{q}$ bukanlah jarak dalam pengertian ini yaitu mereka yang memenuhi sifat-sifat jarak yang umum pada (1-25). Mereka hanya sejumlah keterangan (reference) yang digunakan untuk menilai ketidakmonotonan dari observasi $d_{ik}^{(q)}$.

Gagasan untuk menemukan gambaran item sebagai titik-titik dalam q-dimensi sedemikian sehingga nilai stress (kedekatan) sekecil mungkin. Kruskal (14) mengemukakan penafsiran secara informal menurut garis pedoman berikut :

Stress	Goodness of fit
20 %	Tidak baik
10 %	Kurang
5 %	Baik
2.5 %	Baik sekali
0 %	Sempurna

Goodness of fit mengacu kepada hubungan kemonotonan antara kesamaan dan jarak akhir.

Telah kita nyatakan bahwa ukuran stress sebagai suatu fungsi q, jumlah dimensi untuk penggambaran secara geometri. Untuk setiap q, susunan yang menghasilkan stress minimum dapat diperoleh. Karena q akan meningkatkan stress minimum dalam *rounding error*, meningkatkan dan akan sama dengan nol untuk q = N-1. pertama-tama untuk q = 1, plot jumlah dari stress (q) melawan q

dapat dikonstruksi. Dari nilai q ini kita memilih dimensi yang paling baik yaitu kita mencari *"siku (elbow)"* dalam plot dimensi stress.

Algoritma *multidimensional scaling* dapat diringkas melalui tiga tahapan :

1. Untuk N item, maka $M = N(N-1)/2$ kesamaan (jarak) antara pasangan-pasangan itemnya. Susun kesamaan(jarak) seperti dalam persamaan (12-15). (Jarak disusun dari yang terbesar hingga yang terkecil. Jika kesamaannya (jarak) tidak dapat dihitung, maka susunan rank harus ditentukan.)
2. dengan menggunakan susunan percobaan dalam q-dimensi, tentukan jarak antar item, $d_{ik}^{(q)}$ dan jumlah $\hat{d}_{ik}^{(q)}$ yang kemudian memenuhi persamaan (12-16) dan minimumkan *stress* dalam persamaan (12-17). ($\hat{d}_{ik}^{(q)}$ biasanya ditentukan dengan menggunakan program komputer menggunakan metode regresi yang dirancang untuk menghasilkan jarak monoton yang "fitted".
3. Dengan menggunakan $\hat{d}_{ik}^{(q)}$, titik-titik dipindahkan untuk memperoleh susunan yang baru. (untuk q tetap, susunan yang baru ditentukan oleh fungsi umum prosedur minimisasi yang diterapkan pada *stress*. Dalam konteks ini stress dianggap sebagai fungsi dari koordinat N x q dari N item.) susunan yang baru akan memiliki $d_{ik}^{(q)}$ dan $\hat{d}_{ik}^{(q)}$ yang baru, dan stress yang lebih kecil dari sebelumnya. Proses tersebut diulang sampai diperoleh stress minimum terbaik.
4. Plot stress minimum dan pilih jumlah dimensi q^* terbaik. Kita telah mengasumsikan nilai kesamaan awal adalah simetri $(s_{ik} = s_{ki})$, maka

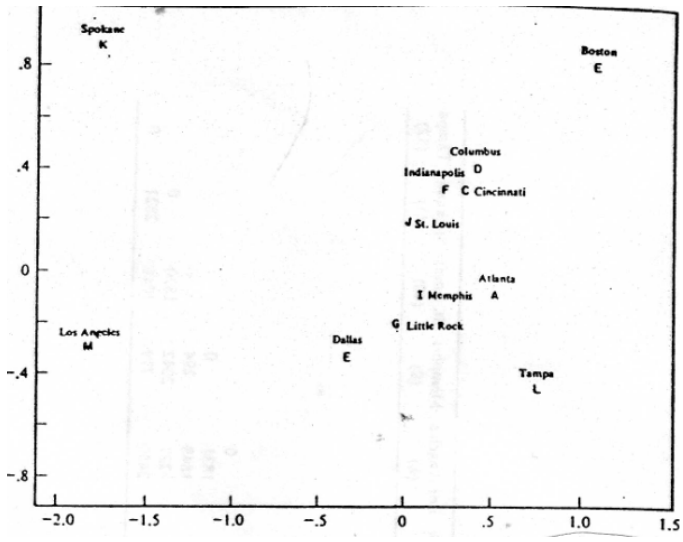
no ties, dan tidak ada observasi yang hilang. Kruskal menyarankan suatu metode untuk menangani ketidaksimetrian ini, ties, dan observasi hilang. Lagi pula sekarang terdapat program komputer yang dapat menangani tidak hanya jarak euclid tetapi juga jarak Minkowski. [lihat (12 - 3)].

Contoh berikut merupakan ilustrasi dari *multidimensional scaling* dengan jarak sebagai ukuran kesamaan awal.

Contoh 12.13

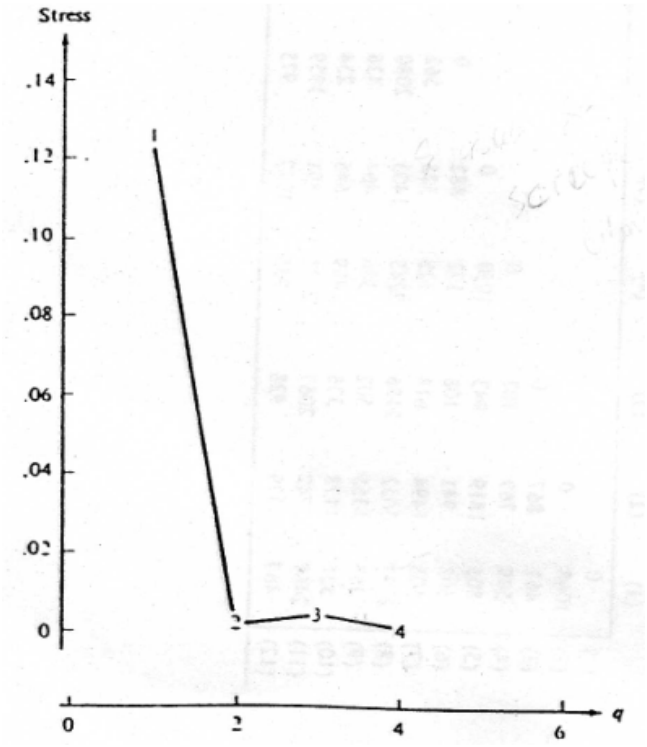
Tabel 12.7 memperlihatkan jarak antara pasangan kota-kota terpilih di Amerika.

Karena kota-kota tersebut tentu saja terletak dalam jarak dua dimensi. Perhatikan jika jarak pada tabel 12.7 diurut dari yang terbesar hingga yang terkecil yaitu yang paling sedikit sama hingga yang paling banyak kesamaannya, maka posisi pertama ditempati oleh $d_{Boston,LA.} = 3052$.



Gambar 12.13

Gambaran geometris dari kota-kota yang dihasilkan oleh *multidimensional scaling*



Gambar 12.14

Fungsi *stress* jarak antar kota pada perusahaan penerbangan

Plot multidimensional scaling untuk q = 2 dimensi ditunjukkan dalam gambar 12.13. sumbu yang terletak sepanjang scatterplot principal components sampel.Plot dari stress(q) melawan q ditunjukan dalam gambar 12.14. karena stress(1)x100% = 12%, suatu gambaran kota-kota dalam satu dimensi (sepanjang sumbu tunggal) kurang pantas. Siku (elbow) pada fungsi stress trjadi pada q = 2. disini stress(2) x 100 % = 0.08% dan dilihat dari tabel ”Goodness of fit” nya hampir sempurna.

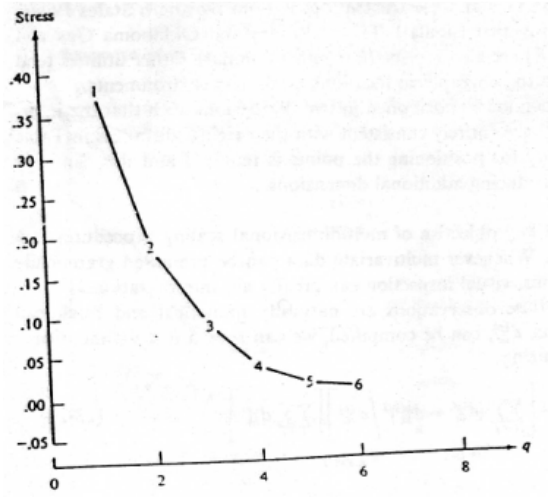
Plot pada gambar 12.14 menunjukkan q = 2 adalah pilihan terbaik untuk dimensi. Perhatikan sesungguhnya untuk nilai stress meningkat untuk q = 3. ini merupakan

keanehan yang dapat terjadi untuk nilai stress yang sangat kecil karena kesulitan untuk pencarian prosedur numerik yang digunakan untuk meletakan stress minimum.

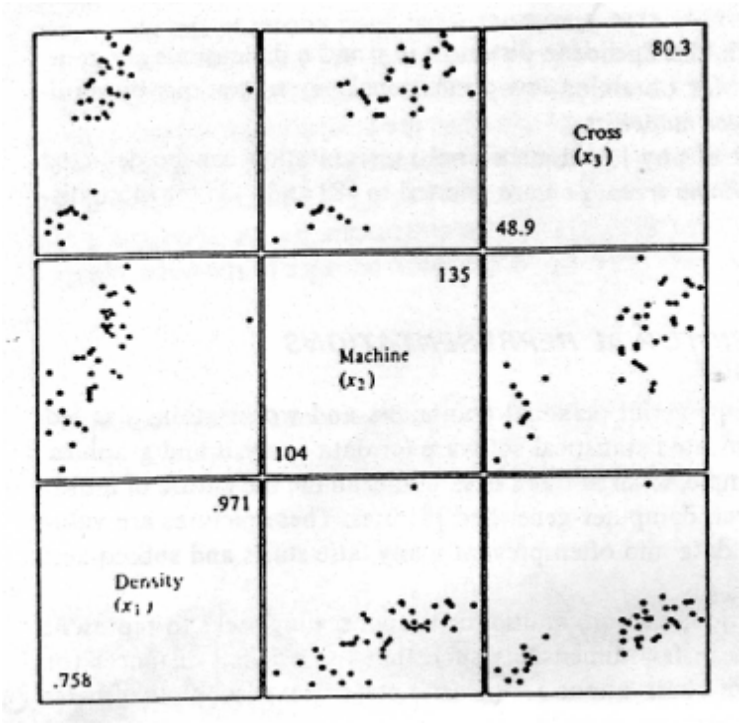
Contoh 12.14

Misalkan untuk menggambarkan 22 perusahaan keperluan umum yang telah didiskusikan pada contoh 12.8 sebagai titik-titik dalam dimensi kecil. Ukuran dis(similarrities) antara pasangan perusahaan merupakan jarak euclid yang terdaftar dalam tabel 12.1. multidimensional scaling dalam q = 1, 2, 3, ...,6 dimensi dihasilkan fungsi stress dalam gambar 12.15 di bawah ini.

Dalam gambar tersebut terlihat tidak adanya siku (elbow) yang mencolok . nilai stressnya adalah kurang lebih 5 % disekitar q = 4. sebuah penggambaran yang baik dalam 4 dimensi dari suatu keperluan dapat dicapai akan tetapi sulit untuk ditunjukkan. Kita menunjukkan plot suatu keperluan susunan diperoleh dalam q = 2 dimensi dalam gambar 12.16. sumbu yang terletak sepanjang komponen utama sampe dari *scatter* akhir.



Gambar 12.15



Gambar 12.16

Meskipun stress untuk dua dimensi cukup tinggi (stress(2) x 100 % = 19 5), jarak antar perusahaan dalam gambar 12.16 konsisten dengan hasil pengelompokan dihadirkan dalam pembahasan sebelumnya. Sebagai contoh keperluan bagian barat tengah, Commonwealth Edison, Wisconsin Electric Power (WEPCO), Madison Gas and Electric (MG &E), dan Northen State Power (NSP) berdekatan. Keperluan texas dan Oklahoma gas dan Electric (Ok. G & E) juga sangat berdekatan. Keperluan lainya cenderung kepada grup yang berdasarkan pada lokasi geografi atau lingkungan yang sama.

Keperluan tidak dapat diposisikan dalam dua dimensi sedemikian sehingga jarak antar keperluan $d_{ik}^{(2)}$ secara keseluruhan konsisten dengan jarak asli pada tabel 12.1 kefleksibelan untuk memposisikan titik-titik diperlukan dan hal ini hanya dapat diperoleh dengan memperkenalkan dimensi tambahan.

Untuk meringkaskan , sasaran utama dalam prosedur multidimensional scaling adalah sebuah gambar dalam dimensi yang rendah. Sewaktu-waktu data multivariat dapat digambarkan scara grafik dalam dua atau tiga dimensi, inspeksi visual sangat dapat membantu interpretasi.

Ketika observasi multivariat merupakan data numerik, dan jarak euclid dalam p-dimensi, $d_{ik}^{(p)}$ dapat dihitung, kita dapat mencari gambaran q < p dimensi dengan meminimumkan

$$E = \left[\sum\sum_{i<k}\left(d_{ik}^{(p)} - d_{ik}^{(q)}\right)^2\right] \Big/ \left[\sum\sum_{i<k} d_{ik}^{(p)}\right]^{-1} \qquad (12\text{-}20)$$

Dalam pendekatan ini, jarak euclid dalam dimensi p dan q dibandingkan secara langsung. Teknik-teknik untuk mendapatkan dimensi yang endah denganmeminimumkan E disebut *nonlinear mapping* (pemetaan tidak linear).

Goodness of fit akhir dari gambaran dimensi yang rendah dapat diperoleh scara grafik dengan *spanning tree* minimal . untuk lebih lanjut pembahasan topik ini dapat dilihat pada (8) dan (13).

2.6 Tampilan-tampilan Data dan Penyajian-penyajian gambar

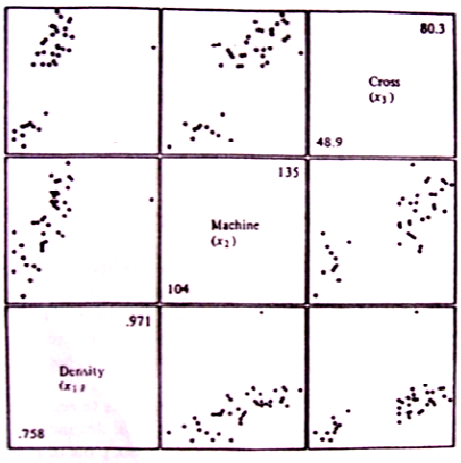
Seperti yang telah kita lihat pada bagian sebelumnya, multidimensional scaling mencoba untuk menggambarkan observasi dalam p-dimensi menjadi observasi dengan sedikit dimensi sedemikian sehingga jarak asli antara pasangan observasi dipertahankan. Secara umum jika obsrvasi multidimensional dapat digambarkan dalam dua dimensi, maka outlier, keterhubungan, pengelompokan yang dapat dibedakan kerap kali dapat dilihat oleh mata. Kita akan

mendiskusikan dan mengilustrasikan beberapa metode untuk memperlihatkan data multivariat dalam dua dimensi.

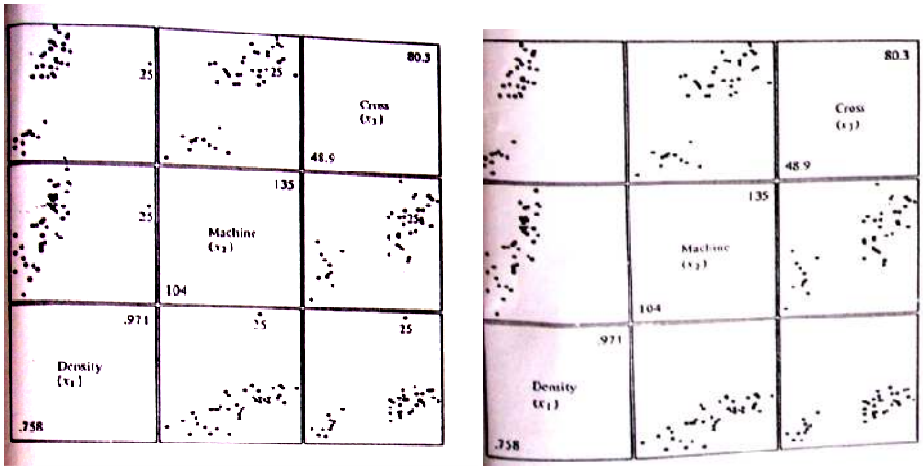
Hubungan Perkalian Scatterplot Dua Dimensi

Contoh 12.15

Untuk mengilustrasikan keterhubungan scatterplot dua dimensi, kita mengacu pada data kualitas kertas dalam tabel 1.2. data ini menggambarkan ukuran variabel X_1 = kepadatan, X_2 = daya regang dalam machine direction X_3 = daya regang dalam cross-direction. Gambar 12.17 menunjukkan scaterplot dua dimensi untuk pasangan variabel-variabel ini yang disusun sebagai array 3 x 3. sebagai contoh, gambar pada sudut sebelah kiri atas pada gambar merupakan scatterplot dari pasangan (x_1, x_3) . yaitu nilai x_1 diplot sepanjang sumbu horizontal dan nilai x_3 diplot sepanjang sumbu vertikal. Sedangkan scaterplot pada sudut sebelah kanan bawah dari gambar merupakan observasi (x_3, x_1). Dengan kata lain sumbu-sumbunya berkebalikan. Perhatikan variabel-variabel dan rentang tiga digitnya ditunjukkan dalam kotak sepanjang diagonal SW-NE.



Operasi pemilihan outlier tertentudalam scatterplot (x_1, x_3) dari gambar 12.17 menghasilkan 12.18 (a), dimana outlier ditandai sebagai specimen 25 dan titik data yang sama disorot dalam scatterplot lain. Specimen 25 juga terlihat sebagai outlier dalam scatterplot (x_1, x_2) tetapi bukan pada scatterplot (x_2, x_3). Operasi penghapusan specimen ini mengantarkan pada scatterplot pada gambar 12.18(b).



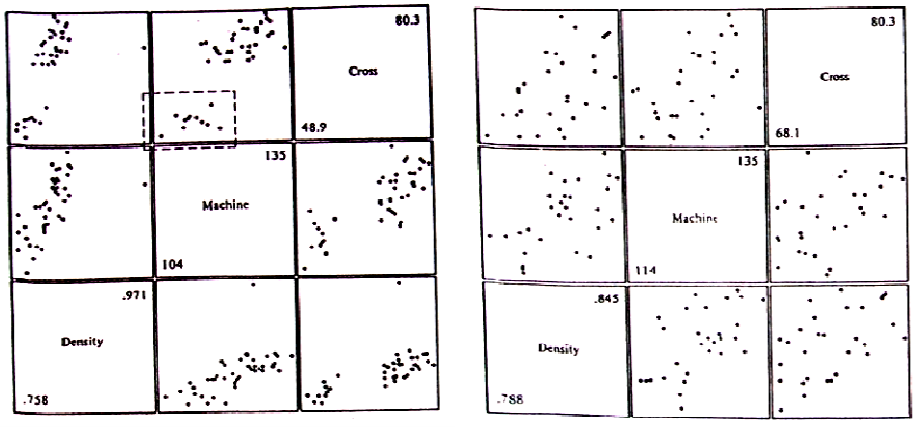
(a) (b)

Dari gambar 12.17, kita dapat lihat bahwa beberapa titik pada contoh tersebut scatterplot (x_2, x_3) terlihat terhubung dengan scatterplot lain. Pemilihan titik-titik ini menggunakan bujur sangkar (lihat halaman 612), menyoroti titik terpilih pada semua scatterplot dan dilihat pad gambar 12.19(a).lagipula pengecekan specimen (contoh) 16-21, 34, dan 38-41 sesungguhnya adalah contoh dari gulungan kertas yang lebih lama yang termasuk dalam urutan yang memiliki cukup lapisan dalam kardus yang diproduksi.

Pengoperasian poin-poin penyorotan yang sesuai dengan suatu cakupan yang terpilih salah satu dari variabel-variabel disebut *Brushing*. *Brushing* bisa mulai dengan suatu persegi panjang, seperti di Gambar 12.19 (a), akan tetapi proses

brushing tersebut bisa dipindah ke penetapan suatu urutan dari poin-poin yang digarisbawahi. Proses itu dapat dihentikan pada setiap waktu untuk menetapkan suatu *snapshot* dari situasi yang ada.

Scatterplots seperti itu berada dalam contoh 12.15 adalah bantuan-bantuan sangat bermanfaat di dalam analisis data. Teknik grafis baru penting yang lain adalah dengan menggunakan perangkat lunak. Hal ini bisa dilakukan secara dinamis dan secara terus-menerus sampai data yang informatif dan bersaingan diperoleh.



(a) (b)

Suatu strategi untuk analisa penyelidikan multivariate grafis dalam garis, yang termotivasi oleh kebutuhan akan suatu prosedur yang rutin untuk mencari-cari struktur di data multivariat, disampaikan dalam contoh berikut.

Contoh 12.16

Empat pengukuran yang berbeda dari kekakuan kayu diberikan dalam Table 4.3. Di Dalam Contoh 4.13, kita mengenali spesimen (papan) 16 dan mungkin spesimen (papan) 9 sebagai pengamatan-pengamatan yang tidak biasa. Gambar 12.20 (a), (b). dan (c) berisi perspektif-perspektif dari data kekakuan di dalam x_1, x_2, x_3 ruang. Pandangan-pandangan ini diperoleh oleh secara terus menerus

berputar dan memutar tiga koordinat dimensional. Memutar koordinat membiarkan satu dan lainnya untuk mendapat suatu pemahaman yang lebih baik tentang tiga aspek dimensional dari data. Gambar 12.20 (d) adalah gambar dari data kekakuan di x_2, x_3, x_4 ruang. Kenali bahwa Gambar 12.20 (a) dan (d) secara visual mengkonfirmasikan spesimen-spesimen 9 dan 16 seperti pencilan. Spesimen 9 sangat besar di dalam ketiga koordinat tersebut. Perputaran yang berlawanan arah jarum jam seperti perputaran di dalam Gambar 12.20(a) hasilkan Gambar 12.20 (b), dan kedua pengamatan-pengamatan yang tidak biasa disembunyikan di dalam pandangan ini. Suatu penjabaran lebih lanjut x_2, x_3 memberi Gambar 12.20 (c); salah satu pencilan (16) kini tersembunyi.

Kita sekarang berpindah kepada tiga penyajian-penyajian bergambar yang populer data multivariat dalam dua dimensi yaitu *stars*, *Andrews plot*, dan *Chernoff faces*.

Stars

Umpamakan masing-masing unit data terdiri dari pengamatan-pengamatan tidak negatif di $p \geq 2$ variabel. Dalam dua dimensi, kita dapat membangun lingkaran-lingkaran dari suatu radius yang ditetapkan (menjadi acuan) dengan sinar yang sama yang berasal dari pusat dari lingkaran. Panjang-panjang dari sinar menunjukkan nilai-nilai dari variabel-variabel. Akhir dari sinar itu dapat dihubungkan dengan garis lurus untuk membentuk suatu bintang. Masing-masing bintang menunjukkan suatu pengamatan multivariate dan bintang-bintang dapat dikelompokkan menurut persamaan.

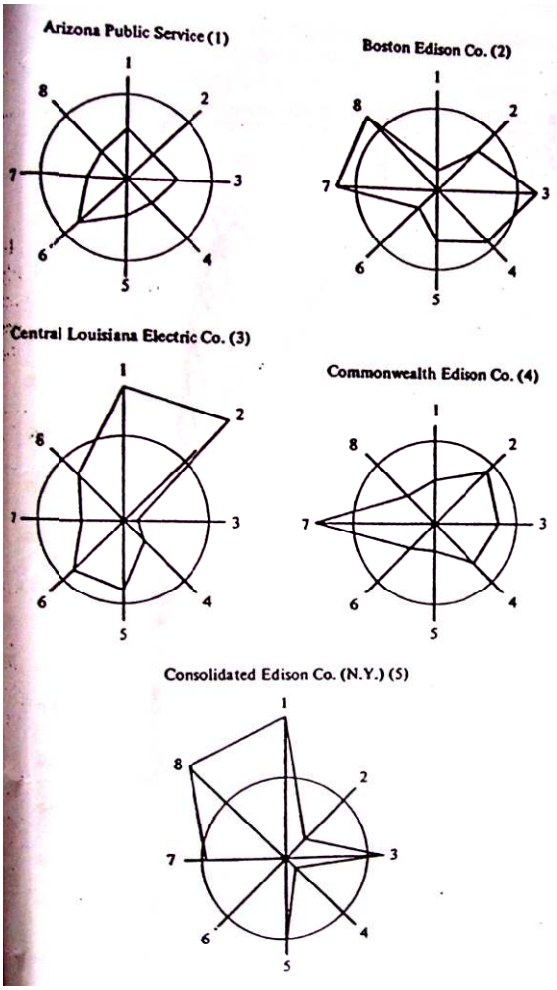
Metode *stars* sering sangat membantu. Ketika akan membuat bintang-bintang, sebaiknya untuk menstandardisasi hasil pengamatan-pengamatan. Dalam hal ini

mungkin sebagian dari hasil pengamatan itu biasanya negatif. Pengamatan-pengamatan itu kemudian bisa ditampilkan kembali setelah distandardisasi sehingga pusat dari lingkaran menunjukkan nilai pengamatan paling kecil dari seluruh data.

Andrews Plot

Andrews sudah mengusulkan bahwa suatu vektor dimensional dari p pengukuran-pengukuran $[x_1, x_2, \ldots, x_p]$ diwakili oleh Deret Fourier yang terbatas

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots \quad -\pi \leq t \leq \pi \text{(12-21)}$$



Lalu, pengukuran-pengukuran dijadikan koefisien-koefisien dalam suatu grafik merupakan suatu fungsi periodik. Sebagai contoh, pengamatan 4-dimensional [6, 3, -1,2]' dikonversi menjadi fungsi

$$f(t) = \frac{6}{\sqrt{2}} + 3\sin t - \cos t + 2\sin 2t, \qquad -\pi \leq t \leq \pi$$

dan plot sebagai suatu fungsi t.

Plot dari Penyajian-penyajian deret Fourier dari pengamatan multivariat akan kurva-kurva yang kemudian bisa secara visual dikelompokkan. Andrews plots dilakukan dengan menukar koordinat-koordinat (koefisien-koefisien). Sebagai konsekwensinya yaitu mencoba bermacam-macam tampilan sebelum memutuskan satu-satunya yang terbaik untuk suatu data yang diberikan.

Pengalaman sudah menunjukkan bahwa data itu harus distandardisasi sebelum membentuk Deret Fourier. Lebih dari itu, jika banyaknya materi melembutkan kepada besar, Andrews plot menjadi sulit. Banyaknya Andrews membengkok yang dilapiskan di grafik perlu mungkin dibatasi sebanyak lima atau enam.

Contoh 12.18

Perwakilan pengamatan-pengamatan 22 utilitas publik menurut (12.21) di dalam Gambar 12.22. Kelompok perusahaan yang serupa kebanyakan sulit untuk di lihat. Termotivasi oleh matriks jarak di dalam Gambar 12.2 (lihat Contoh 12.1), kita memplot kelompok terdiri dari perusahaan (4,10,13,20,22). Hasil itu ditunjukkan di dalam Gambar 12.23. Catat bahwa perusahaan 22 (Virginia Electric dan Power Company) terlihat mempunyai bit yang berbeda dari istirahat dan plot *Andrews* konsisten dengan algoritma pengelompokan rata-rata keterhubungan hirarkis pada ilustrasi 12.10 (lihat Gambar 12.11).

Chernoff faces

Orang-orang bereaksi dengan muka. Chernoff menggambarkan pengamatan-pengamatan dimensional p sebagai suatu muka dimensional dengan karakteristik-karakteristik bentuk muka, lengkungan mulut, panjang hidung, ukuran mata, posisi pupil, dan sebagainya ditentukan oleh nilai pengukuran-pengukuran dari variabel-variabel di p.

Seperti mula-mula merancang, *Chernoff faces* mampu menangani sampai dengan 18 variabel. Tugas dari variabel-variabel kepada fitur fasial dilaksanakan oleh eksperimen dan aneka pilihan yang berbeda menghasilkan hasil-hasil yang berbeda. Beberapa perkataan berulang-ulang adalah biasanya perlu sebelum penyajian-penyajian yang memuaskan dicapai. Jika penyelidik itu adalah [secara] wajar pasti dua atau tiga variabel terutama bertanggung jawab untuk seikat-seikat yang pembeda, variabel-variabel ini dapat dihubungkan dengan karakteristik-karakteristik fasial yang terkemuka. Menghubungkan satu "yang penting" variabel dengan suatu karakteristik seperti panjangnya hidung, dibanding suatu lebih sedikit karakteristik yang terkemuka seperti posisi murid, mengizinkan[membiarkan satu untuk memilih pengelompokan-pengelompokan lebih siap.

Seperti *Andrews plots*, *Chernoff faces* bermanfaat karena membuktikan (1) satu pengelompokan awal yang diusulkan oleh pengetahuan pokok dan intuisi atau (2) pengelompokan akhir yang dihasilkan oleh algoritma cluster.

Contoh 12.19

Dengan menggunakan data dalam table 12.5, perusahaan fasilitas umum menggunakan Chernoff faces. Kita mengikuti aturan berikut.

	Variabel	Karakteristik wajah
X_1	Fixed charge coverage.	Tinggi setengah muka
X_2	Rate of return on capital.	Lebar muka
X_3	Cost per KW capacity in place.	Posisi pusat mulut
X_4	Annual load factor.	Kemiringan mulut
X_5	Peak KV/H demand growth from 1974.	Keeksentrikan mata
X_6	Sales (KWH use per year).	Separuh panjang mata
X_7	Percent nuclear.	Kelengkungan mulut
X_8	Total fuel costs (cents per KWH).	Panjang Hidung

Membangun *Chernoff faces* adalah suatu tugas itu harus dilakukan dengan bantuan komputer. Data itu biasanya distandardisasi di dalam program komputer sebagai bagian dari proses untuk menentukan lokasi-lokasi, ukuran-ukuran, dan orientasi-orientasi karakteristik-karakteristik yang fasial. Dengan beberapa pelatihan, *Chernoff faces* bisa merupakan suatu cara yang efektif untuk komunikasi;kan persamaan atau perbedaan-perbedaan.

Kesimpulan Akhir

Ada beberapa cara untuk menggambarkan data multivariat dalam dua dimensi. Kita sudah menggambarkan beberapa diantaranya.

Efektivitas dari *Stars, Andrews plots*, dan *Chernoff faces* disatukan. Kadang-kadang gambar tersebut dapat lebih informatif; bagaimanapun, lebih sering daripada tidak, mereka tidak akan menghilangkan ciri tiap kelompok.

BAB III

KESIMPULAN DAN SARAN

3.1 Kesimpulan

- Analisis cluster dilakukan untuk mengelompokan objek-objek yang memiliki kemiripan (homogen). Berdasarkan karakteristik yang dimiliki, dengan analisis cluster sekelompok objek dapat dikelompokkan.
- Metode pengelompokan pada dasarnya ada dua, yaitu pengelompokan hirarki (HierarchicalClustering Method) dan pengelompokan non hirarki Non Hierarchical Clustering Method).
- Metode pengelompokan hirarki digunakan apabila belum ada informasi jumlah kelompok. Sedangkan metode pengelompokan non hirarki bertujuan mengelompokan n obyek ke dalam k kelompok ($k < n$).
- Salah satu prosedur pengelompokan pada non hirarki adalah dengan menggunakan metode K-Means. Metode ini merupakan metode pengelompokan yang bertujuan mengelompokan obyek sedemikian hingga jarak tiap-tiap obyek ke pusat kelompok di dalam satu kelompok adalah minimum.

3.1 Saran

Terdapat beberapa algoritma cluster yang dapat digunakan untuk mengelompokkan objek-objek, baik itu dengan pengelompokan hirarki ataupun pengelompokan non hirarki. Namun yang perlu diperhatikan adalah stabilitas dari

solusi yang diperoleh, oleh karena itu perlu di cek kembali setiap algoritma cluster tersebut baik sebelum atau sesudah pengelompokkan.