TELAAHAN KONSEP JARAK PADA ANALISIS GEROMBOL

Hazmira Yozza
(Jurusan Matematika Universitas Andalas)

email : hazmirayozza@gmail.com

ABSTRAK

Analisis Gerombol adalah suatu teknik analisis peubah ganda yang digunakan untuk mengelompokkan objek-objek ke dalam beberapa gerombol berdasarkan kemiripan/ketakmiripan antar objek. Salah satu ukuran ketakmiripan yang biasa digunakan adalah ukuran jarak. Ukuran jarak yang berbeda akan menghasilkan penggerombolan yang berbeda. Tulisan ini akan membahas perbandingan beberapa konsep jarak yang biasa digunakan dalam analisis gerombol untuk data biner. Perbandingan konsep jarak akan dilakukan dengan menggunakan metode simulasi dan didasarkan pada persentase salah klasifikasi yang dihasilkan. Dari simulasi yang dilakukan disimpulkan bahwa secara umum, beragam konsep jarak untuk data biner ini memberikan persentase salah klasifikasi yang berbeda. Secara umum, konsep jarak penyesuaian sederhana, Hamann, Rogers&Tanimoto dan Sokal&Sneath 1 memberikan nikai salah klasifikasi yang relatif sama, demikian juga dengan konsep jarak Dece, Jaccard dan Sokal&Sneath 2. Konsep jarak yang memberikan salah klasifikasi terbesar adalah konsep jarak Russel & Rao.

Kata Kunci : analisis gerombol, konsep jarak, persentase salah klasifikasi, simulasi

1. PENDAHULUAN

Analisis Gerombol adalah suatu teknik analisis peubah ganda yang digunakan untuk mengelompokkan objek-objek ke dalam beberapa gerombol berdasarkan peubah-peubah yang diamati terhadap objek tersebut. Diharapkan, objek-objek yang mirip akan berada pada gerombol yang sama sedangkan objek tidak mirip akan masuk ke dalam gerombol yang berbeda.

Penerapan dari analisis gerombol ini sangat luas. Analisis ini dijumpai di beragai bidang aplikasi, seperti bidang pertanian, psikologi, pengembangan wilayah dan lain-lain. Dari segi data, analisis ini dapat digunakan bukan hanya untuk data kuantitatif, namun juga untuk data yang bersifat kualitatif, salah satunya adalah untuk data biner. Sebagai contoh, analisis ini dapat digunakan untuk mengelompokkan produk sabun berdasarkan ada atau tidaknya karakteristik tertentu pada produk sabun tersebut.

Terdapat banyak teknik yang dapat dilakukan untuk menggerombolkan objek. Secara umum, teknik-teknik tersebut dapat dikelompokkan menjadi teknik penggerombolan tak berhierarkhi dan teknik penggerombolan berhierarkhi, baik yang bersifat penggabungan (*agglomerative*) maupun pemisahan (*divisive*). Pada teknik yang bersifat penggabungan, pertama-tama dibentuk kelompok sebanyak objek yang ada, dimana setiap objek berada dalam kelompok-kelompok yang terpisah. Kelompok yang terdekat secara bertahap digabung sampai akhirnya semua objek berada dalam satu kelompok. Tahapan penggerombolan tersebut digambarkan dalam sebuah *dendogram*. Metode analisis gerombol selengkapnya dapat dilihat pada Andenberg (1973).

Langkah awal dalam suatu analisis gerombol berhierarkhi penggabungan adalah menentukan ukuran kemiripan atau ketakmiripan antar objek. Ukuran ini mengukur derajat kemiripan antara dua objek. Jarak adalah ukuran yang banyak digunakan untuk mengukur ketakmiripan antara dua objek. Ukuran inilah yang merupakan dasar dalam melakukan penggerombolan. Untuk data hasil pengukuran, jarak, biasanya digunakan ukuran jarak sebagai ukuran ketakmiripan, sedangkan bila data yang digunakan adalah data kualitatif, biasanya digunakan ukuran kemiripan/kesamaan.

Terdapat banyak ukuran jarak yang biasa digunakan. Untuk data kuantitatif, dapat digunakan jarak Minskowski, jarak City-Block, jarak Euclid, jarak Mahalanobis dan lain-lain. Untuk data biner, ukuran jarak biasanya diperoleh dengan terlebih dahulu membentuk table tabel kontingensi 2 x 2 untuk setiap pasangan objek ke-i dan ke-j, sebagai berikut.

Objek ke-j	Objek ke-i	
	1	0
1	*A*	*b*
0	*C*	*d*

Keterangan : *a* : banyaknya peubah yang pada objek ke-i dan ke-j sama-sama bernilai 1;
b : banyaknya peubah yang pada objek-i bernilai 0 sedangkan pada objek-j bernilai 1;
c : banyaknya peubah yang pada objek ke-i bernilai 1 sedangkan pada objek ke-j bernilai 0;
d : banyaknya peubah yang pada objek ke-i dan ke-j sama-sama bernilai 0

Dari tabel tersebut dapat diperoleh berbagai ukuran jarak, seperti koefisien simple matching, dice, Hamann, Jaccard, Rogers & Tanimoto, Russel & Rao, Sokal & Sneath 1 - 5 dan lain-lain.

Penelitian ini dilakukan untuk membandingkan ukuran-ukuran jarak yang biasa digunakan pada analisis gerombol yang dilakukan terhadap data biner. Pembandingan ukuran jarak tersebut dilakukan pada tiga metode ukuran perbaikan jawak, yaitu :metode pautan rataan antar kelompok (*average linkage between group*), metode pautan rataan dalam kelompok (*average linkage within group*) dan metode pautan lengkap (*complete linkage*).

2. DATA DAN METODE

Penelitian ini dilakukan dengan menggunakan metode simulasi. Data yang digunakan dibangkitkan dengan menggunakan paket program Minitab v.15. Data yang dibangkitkan berukuran n = 48, banyak peubah p = 20 dan banyak kelompok k = 2, 3 atau 4. Pembangkitan data dilakukan dari sebaran Bernoulli seperti pada tabel berikut.

Tabel 1. Struktur Data

k	Kel	Peubah 1-5	Peubah 6-10	Peubah 11-15	Peubah 16-20
2	1	Bernoulli(0.2)	Bernoulli(0.4)	Bernoulli(0.6)	Bernoulli(0.8)
	2	Bernoulli(0.6)	Bernoulli(0.8)	Bernoulli(0.2)	Bernoulli(0.4)
3	1	Bernoulli(0.2)	Bernoulli(0.2)	Bernoulli(0.5)	Bernoulli(0.8)
	2	Bernoulli(0.5)	Bernoulli(0.8)	Bernoulli(0.8)	Bernoulli(0.5)
	3	Bernoulli(0.8)	Bernoulli(0.5)	Bernoulli(0.2)	Bernoulli(0.2)
4	1	Bernoulli(0.2)	Bernoulli(0.4)	Bernoulli(0.6)	Bernoulli(0.8)
	2	Bernoulli(0.4)	Bernoulli(0.6)	Bernoulli(0.8)	Bernoulli(0.2)
	3	Bernoulli(0.6)	Bernoulli(0.8)	Bernoulli(0.2)	Bernoulli(0.4)
	4	Bernoulli(0.8)	Bernoulli(0.2)	Bernoulli(0.4)	Bernoulli(0.6)

Terdapat beberapa tahapan dalam penelitian ini.

1. Data yang telah dibangkitkan dari sebaran-sebaran di atas dianalisis dengan menggunakan analisis bergerombol berhierarkhi dengan kombinasi ukuran kemiripan dan metode perbaikan jarak sebagai berikut :
   a. Metode perbaikan jarak: jarak gerombol T dan R adalah rata-rata jarak semua pasangan objek yang mungkin yang ada pada gerombol T dan R atau:

$$d_m = \frac{\sum_i \sum_j d_{ij}}{N_T N_R}$$

d_{ij} : jarak objek i pada gerombol T dan j pada gerombol R
N_T : Jumlah anggota gerombol T
N_R : Jumlah anggota gerombol R

- metode pautan rataan dalam kelompok; jarak gerombol T dan R didasarkan jarak antara setiap pasangan objek yang ada pada gerombol T dan R melalui formula:

$$d_{TR} = \frac{\sum_i d_i + \sum_j d_j + \sum_i \sum_j d_{ij}}{1/2(N_T + N_R)(N_T + N_R - 1)}$$

- metode pautan lengkap; jarak gerombol UV yang merupakan gabungan objek U dan V dengan gerombol W didefinisikan sebagai :

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

b. Ukuran kemiripan :

- Kesesuaian sederhana; $d_{ij} = \frac{a+d}{a+b+c+d}$
- Dice $d_{ij} = \frac{2a}{2a+b+c}$
- Hamann $dij = \frac{(a+c)-(b+d)}{a+d+2(b+c)}$
- Jaccard $d_{ij} = \frac{a}{a+b+c}$
- Rogers & Tanimoto $d_{ij} = \frac{a+d}{a+d+2(b+c)}$
- Russel & Rao $d_{ij} = \frac{a}{a+b+c+d}$
- Sokal & Sneath 1 $d_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$
- Sokal & Sneath 2 $d_{ij} = \frac{a}{a+2(b+c)}$

2. Menghitung persentase salah klasifikasi untuk masing-masing kombinasi tersebut. Prosedur pembangkitan dan analisis data di atas diulang sebanyak 25 kali.
3. Menghitung rata-rata persentase salah klasifikasi untuk setiap kombinasi konsep jarak dan metode perbaikan jarak untuk 25 kali ulangan simulasi yang dilakukan
4. Memandingkan kedelapan konsep jarak tersebut berdasarkan rata-rata persentase kesalahan pengelompokan yang terjadi.

3. HASIL DAN DISKUSI

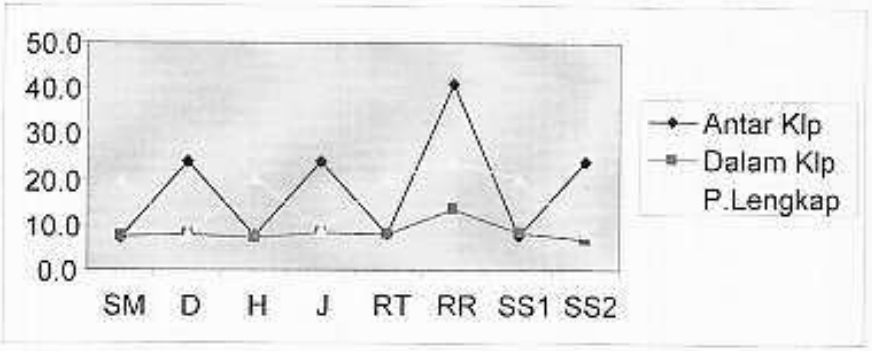
Dari simulasi yang dilakukan, untuk banyak kelompok k = 2, diperoleh rata-rata kesalahan klasifikasi sebagaimana yang disajikan pada Tabel 1 dan Gambar 1 berikut.

Tabel 2. Perbandingan Persentase Salah Klasifikasi Konsep-konsep Jarak untuk k = 2

Metode Perbaikan Jarak	Konsep Jarak							
	PS	D	H	J	RT	RR	SS1	SS2
P.Rataan antar kelompok	7.1	23.8	7.5	23.8	7.9	40.8	7.5	23.8
P.Rataan dalam Kelompok	7.5	7.9	7.1	8.3	7.9	13.3	8.3	6.7
P.Lengkap	20.0	7.9	20.4	8.8	20.4	23.8	20.4	8.3

Keterangan : PS = Penyesuaian sederhana; D = Dice; H = Hamann; J = Jaccard; RT = Rogers&Tanimoto; RR = Russel&Rao; SS1 = Sokal&Sneath 1; SS2 = Sokal&Sneath 2

Dari tabel tersebut dapat diketahui bahwa dalam analisis gerombol digunakan sebagai metode perbaikan jarak, maka konsep jarak penyesuaian sederhana, kedelapan metode tersebut memberikan rata-rata nilai salah klasifikasi yang berbeda. Kesimpulan tersebut lebih dipertegas oleh pengujian kesamaan nilai tengah yang dilakukan terhadap persentase salah klasifikasi, yang memberikan nilai-p = 0.013>0.05. Hal tersebut berarti bahwa pada taraf nyata 5%, dapat disimpulkan bahwa paling tidak satu konsep jarak memberikan rata-rata persentase salah klasifikasi yang berbeda.



Gambar 1. Perbandingan Persentase Salah Klasifikasi Konsep-konsep Jarak untuk k = 2

Dari tabel dan gambar dapat diketahui bahwa konsep penyesuaian sederhana, Hamann, Rogers&Tanimoto dan Sokal&Sneath 1 memberikan persentase salah klasifikasi yang hampir sama. Demikian juga dengan konsep jarak Dice, Jaccard dan Sokal&Sneath 2. Dibandingkan dengan persentase salah klasifikasi dengan konsep jarak pada kelompok pertama, terlihat bahwa persentase salah klasifikasi pada kelompok kedua

ini lebih besar yaitu sekitar 20%. Persentase salah klasifikasi terbesar terjadi bila digunakan konsep jarak Russel & Rao, yaitu sebesar 40%.

Berbeda dengan sebelumnya, jika dalam analisis gerombol digunakan metode pautan rataan dalam kelompok sebagai metode perbaikan jarak, hasil simulasi memberikan pola yang sangat berbeda, dimana semua konsep jarak memberikan persentase salah klasifikasi yang relatif sama dan persentase salah klasifikasi hanya berkisar 6.5% – 8.5%. Namun demikian, konsep jarak Russel & Rao tetap memberikan salah klasifikasi yang relatif paling besar, yaitu sebesar 13.3%. Persentase salah klasifikasi terkecil diberikan oleh konsep jarak Sokal & Sneath 2, yaitu sebesar 6.7%. Uji kesamaan nilai tengah yang dilakukan terhadap persentase salah klasifikasi memberikan nilai-p sebesar 0.372, yang berarti bahwa secara statistik kedelapan konsep jarak tersebut memberikan rata-rata kesalahan yang sama.

Pola yang juga berbeda diperlihatkan jika digunakan metode Pautan Lengkap sebagai metode perbaikan jarak. Uji hipotesis mengenai kesamaan nilai tengah yang dilakukan terhadap persentase salah klasifikasi memberikan nilai-p = 0.000, yang berarti secara statistik, terdapat sedikitnya satu konsep jarak yang memberikan rata-rata salah klasifikasi yang berbeda dengan yang lain.

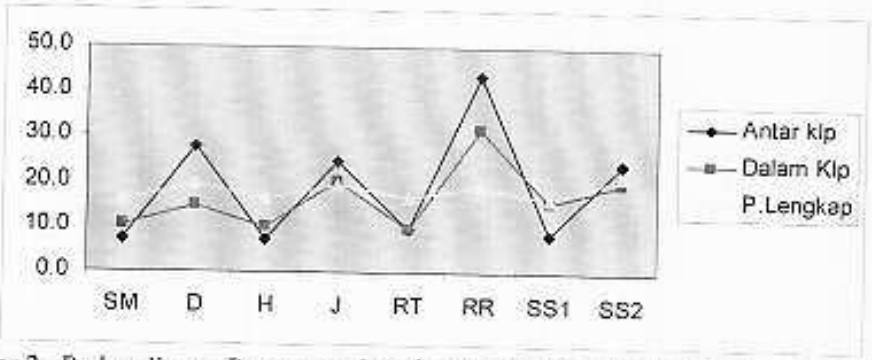
Seperti pada metode pautan rata-rata antar kelompok, dengan metode pautan tunggal ini, terdapat dua kelompok konsep jarak, yaitu kelompok konsep jarak penyesuaian sederhana-Hamann-Rogers&Tanimoto-Sokal&Sneath 1 dan kelompok konsep jarak Dice-Jaccard-Sokal&Sneath 2. Berbeda dengan metode pautan rataan antar kelompok, konsep jarak pada kelompok pertama memberikan nilai salah klasifikasi yang lebih besar. Konsep Russel & Rao memberikan salah klasifikasi yang paling besar, namun nilainya hampir sama dengan salah klasifikasi konsep jarak pada kelompok pertama.

Pada tabel dan gambar berikut disajikan hasil simulasi untuk k=3.

Tabel 3. Perbandingan Persentase Salah Klasifikasi Konsep-konsep Jarak untuk k = 3

Metode Perbaikan Jarak	Konsep Jarak							
	PS	D	H	J	RT	RR	SS1	SS2
P.Rataan antar kelompok	7.1	27.9	7.1	24.6	9.6	43.8	8.3	24.2
P.Rataan dalam Kelompok	10.4	14.6	10.0	20.4	9.6	31.7	15.4	20.0
P.Lengkap	15.8	19.6	16.3	19.6	16.7	19.2	16.3	21.3

Keterangan : PS = Penyesuaian sederhana; D = Dice; H = Hamann; J = Jaccard; RT = Rogers&Tanimoto; RR = Russel&Rao; SS1 = Sokal&Sneath 1; SS2 = Sokal&Sneath 2



Gambar 2. Perbandingan Persentase Salah Klasifikasi Konsep-konsep Jarak untuk k = 3

Dari gambar dan tabel di atas, dapat dijelaskan bahwa bila digunakan metode pautan rata-rata antar kelompok atau pautan rata-rata dalam kelompok sebagai metode perbaikan jarak, maka terdapat dua kelompok konsep jarak, yaitu kelompok I yang terdiri dari konsep jarak penyesuaian sederhana, Hamann, Rogers&Tanimoto dan Sokal&Sneath 1 dan kelompok II yang terdiri dari konsep jarak Dice-Jaccard-Sokal&Sneath 2. Konsep jarak Russel&Rao masih memberikan nilai salah klasifikasi yang paling besar, sebagaimana yang didapat pada kasus dengan k = 2.

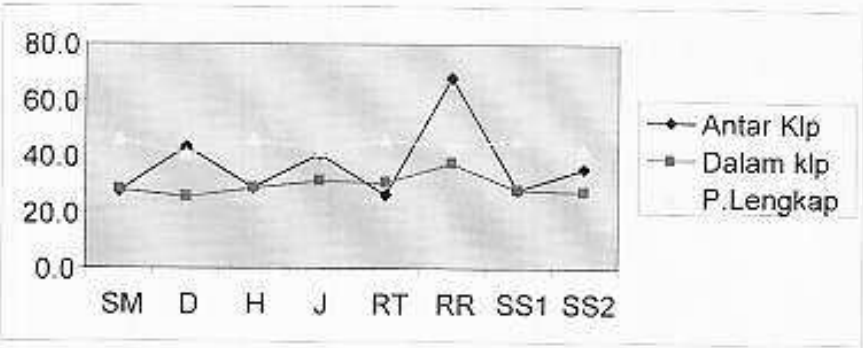
Bila digunakan metode pautan tunggal sebagai metode perbaikan jarak, terlihat bahwa kedelapan konsep jarak tersebut memberikan nilai salah klasifikasi yang hampir sama. Uji hipotesis yang dilakukan untuk menguji kesamaan nilai tengah persentase salah klasifikasi memberikan nilai-p = 0.983 yang berarti bahwa tidak cukup bukti untuk menyangkal bahwa kedelapan konsep jarak tersebut memiliki nilai tengah salah klasifikasi yang sama.

Hasil simulasi untuk k = 4 dapat dilihat pada tabel dan gambar berikut.

Tabel 4. Perbandingan Persentase Salah Klasifikasi Konsep-konsep Jarak untuk k = 3

Metode Perbaikan Jarak	Konsep Jarak							
	PS	D	H	J	RT	RR	SS1	SS2
P.Rataan antar kelompok	27.5	43.3	28.3	40.4	26.3	67.9	27.9	35.8
P.Rataan dalam Kelompok	27.9	25.4	28.8	30.8	30.4	37.5	27.9	27.5
P.Lengkap	45.4	40.8	45.8	41.7	45.8	43.8	45.8	42.5

Keterangan : PS = Penyesuaian sederhana; D = Dice; H = Hamann; J = Jaccard; RT = Rogers&Tanimoto; RR = Russel&Rao; SS1 = Sokal&Sneath 1; SS2 = Sokal&Sneath 2



Gambar 3. Perbandingan Persentase Salah Klasifikasi Konsep-konsep Jarak untuk k = 4

Dari Tabel 4 dan Gambar 3 dapat diketahui bahwa bila digunakan metode pautan rata-rata antar kelompok sebagai metode perbaikan jarak, juga akan didapatkan dua kelompok jarak, seperti yang diperoleh untuk dua kasus sebelumnya. Konsep jarak Russel&Rao tetap memberikan nilai persentase salah klafisikasi yang paling besar, yaitu sebesar 67.8%

Bila digunakan metode pautan rata-rata dalam kelompok atau metode pautan lengkap sebagai metode perbaikan jarak, terlihat bahwa kedelapan konsep jarak tersebut memberikan nilai salah klasifikasi yang hampir sama. Hal tersebut juga dipertegas dengan pengujian hipotesis yang dilakukan untuk menguji kesamaan nilai tengah persentase salah klasifikasi. Pengujian tersebut menghasilkan p-value masing-masing sebesar 0.732 dan 0.795. Hal itu berarti bahwa tidak cukup bukti untuk menyangkal bahwa kedelapan konsep jarak tersebut memberikan nilai persentase salah klasifikasi yang sama.

Bila dibandingkan hasil yang diperoleh untuk k=2, 3 dan 4, diketahui bahwa secara umum persentase salah klasifikasi tersebut semakin besar bila banyak kelompok juga semakin banyak.

4. KESIMPULAN

Dari hasil simulasi yang dilakukan, dapat dismpulkan bahwa konsep jarak yang digunakan akan memberikan gerombol dengan kesalahan klasifikasi yang berbeda jika digunakan metode perbaikan jarak pautan rataan antar kelompok. Terdapat dua

kelompok konsep jarak yang memiliki persentase salah klasifikasi yang sama, yaitu kelompok penyesuaian sederhana-Hamann-Rogers&Tanimoto-Sokal&Sneath 1 dan kelompok konsep jarak Dice-Jaccard-Sokal&Sneath 2. Namun jika digunakan metode perbaikan jarak Konsep jarak yang memberikan salah klasifikasi terbesar adalah konsep jarak Russel & Rao. Semakin banyk kelompok, maka semakin besar persentase salah klasifikasi yang terjadi.

DAFTAR PUSTAKA

Andenberg, MR. 1973. *Cluster Analysis for Application*. Academic Press, Inc., New York

Chatfield, C & AJ Collins. 1980. *Introduction to Multivariate Analysis*. Chapman and Hall, New York.

Finch, H. 2005. Comparison of Distance Measures in Cluster Analysis with Dichotomous Data. *Journal of Data Science* 3 : 85 – 100.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. Prentice Hall, Inc., New York