KAJIAN METODE PENGKLASTERAN HIRARKI DENGAN BERBAGAI PENGUKURAN JARAK

Isma Afrita Lubis[1], Sigit Nugroho[2] dan Baki Swita[2]
[1]Alumni Jurusan Matematika, Fakultas MIPA, Universitas Bengkulu
[2]Dosen Jurusan Matematika, Fakultas MIPA, Universitas Bengkulu

ABSTRACT

The aim of this research are to study distance measurement methods in hierarchical clustering and to know the difference of hierarchical clustering methods according distance measurement concept. Clustering is based on similarity or dissimilarity measure between objects. Eight distance measurement methods are used in this research such as *Euclid*, *Euclid* Square, *Mahalanobis*, *Minkowski*, *City-Block* atau *Manhattan*, *Chebychev, Canberra*, and *Czekanowski* distance.

The method of this research is literature riview. Based that, data of percentage of resident according to the city and the last sertificate in Bengkulu province. Is used to apply some clustering methods. The result of dendogram and cluster membership table, indicate that *Mahalanobis* and *Czekanowski* distance have different clustering. *Euclid* and *Minkowski* distance have similar clustering. *Euclid* Square and *City-Block* or *Manhattan* are the best distance.

Keyword: *Hierarchical Clustering*, *Distance*, *Dendogram*, and *Cluster Membership*

PENDAHULUAN

Analisis klaster bertujuan meminimalisasikan variasi di dalam satu klaster dan memaksimalisasikan variasi antar klaster (Agusta, 2007). Setiap pengklasteran objek hanya masuk ke dalam satu klaster dan tidak tumpang tindih (*overlapping* atau *interaction*).

Pada analisis klaster ukuran kemiripan atau ketidakmiripan yang digunakan adalah jarak (*distance*). Jarak merupakan selisih antara dua objek. Pengklasteran didasarkan pada ukuran kemiripan (*similarity*) atau ketidakmiripan (*dissimilaritiy*) antar objek. Ada beberapa metode untuk mengukur jarak antara dua objek dengan ukuran kedekatan yang berbeda yaitu jarak *Euclid*, kuadrat *Euclid*, *Minkowski*, *City-Block* atau *Manhattan*, *Mahalonobis*, *Chebychev*, *Canberra*, dan *Czekanowski*.

Metode pengklasteran hirarki terdiri dari: metode penggabungan (*Agglomerative Methods*) dan pemecahan (*Devisive Methods*). Metode agglomeratif atau penggabungan terbagi menjadi tiga bagian yaitu: metode pautan (*Linkage Method*), centroid, dan varian. Metode pengukuran jarak berperan pada pengklasteran hirarki dan untuk mengetahui perbedaan yang berarti pada berbagai pengukuran jarak berdasarkan metode pengklasteran hirarki. Agar lebih terarah, pengklasteran dibatasi hanya untuk metode hirarki. Untuk lebih memahami konsep maka dilakukan peneapan data sebagai validasi dengan menggunakan *software* SPSS dan program microsoft excel.

LANDASAN TEORI

Pengertian dan Asumsi Analisis Klaster

Analisis klaster merupakan suatu analisis untuk mengelompokkan objek atau observasi berdasarkan karakteristik yang dimilikinya.

Dalam analisis klaster, setiap objek yang paling dekat kesamaannya akan berada pada kelompok yang sama. Kelompok-kelompok yang terbentuk harus memiliki homogenitas (kesamaan) yang tinggi antar anggota dalam satu klaster (*whithin-cluster*), dan heterogenitas (perbedaan) yang tinggi antar klaster yang satu dengan yang lainnya (*between cluster*).

Homogenitas dan heterogenitas merupakan dua hal yang harus dimiliki sebuah klaster agar kedua itu dapat dikatakan baik. Pengklasteran yang baik adalah pengklasteran yang setiap objek hanya masuk ke dalam satu klaster atau menjadi anggota dari salah satu klaster sehingga tidak terjadi tumpang tindih.

Istilah analisis klaster pertama kali digunakan oleh Tryon (1939) (Anonim, 2008c). Analisis ini dilakukan bertujuan untuk: mereduksi data menjadi data baru dengan jumlah lebih kecil, melakukan generalisasi suatu populasi untuk memperoleh suatu hipotesis, dan menduga karakteristik data.

Asumsi-asumsi yang harus dipenuhi dalam analisis klaster adalah (Anonim, 2008b): ukuran sampel harus benar-benar mewakili seluruh populasi dan tidak terjadi multikolinearitas.

Pengukuran Kemiripan atau Ketidak miripan antar Objek

Langkah pertama dalam analisis klaster adalah menentukan matriks data yaitu **X** (*n* x *p*), dimana n adalah ukuran kedekatan objek dan p merupakan variabel (Hardle & Simar, 2007). Pengelompokan didasarkan pada ukuran kemiripan (*similarity*) atau ketidakmiripan (*dissimilarity*) antar objek. Hal ini dilakukan untuk memperoleh matriks *proximity*, yaitu matriks persegi dan simetri dengan jumlah objek yang sama pada baris dan kolom. Matriks *proximity* memuat semua pasangan kemiripan atau ketidakmiripan diantara objek yang akan diobservasi. Jika x_i dan x_j adalah objek ke-i dan ke-j maka nilai pada baris ke-i dan kolom ke-j dari matriks *proximity* adalah *similarity* atau *dissimilarity* antara x_i dan x_j (Anonim, 2000).

Bentuk matriks *proximity* objek adalah **D** (*n* x *n*)

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \qquad (1)$$

matriks **D** merupakan ukuran *similarity* atau *dissimilarity* antara *n* objek. Matriks ini mempunyai elemen d_{ij}, dimana $i = 1, 2, ..., n$ dan $j = 1, 2, ..., n$, elemen $d_{ij} > 0$ jika $i \neq j$ dan $d_{ij} = 0$ jika $i = j$. Sehingga diperoleh matriks jarak sebagai berikut:

$$d = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix} \quad (2)$$

Semakin kecil nilai d, maka semakin besar kemiripan antar kedua pengamatan tersebut. Sebaliknya bila d besar, semakin besar ketidakmiripan dari pengamatan tersebut.

Ada beberapa metode yang digunakan untuk mengukur kemiripan antar objek (Santoso, 2004):

a. Mengukur jarak antar dua objek
   Mengukur jarak antar dua objek ini berbentuk matriks simetri $n \times n$ yang berisi kemiripan atau ketidakmiripan antar objek sehingga jarak dua objek bisa langsung diukur.
b. Mengukur korelasi antar sepasang objek pada beberapa variabel.
   Pada metode ini data berbentuk matriks. Kemiripan antar objek didapat dengan transformasi satu-satu sehingga indeks ketidakmiripan bisa dikonversi menjadi indeks.
c. Mengukur asosiasi antar objek.
   Untuk mengukur asosiasi antar objek data tidak berbentuk matriks (nominal atau ordinal). Nominal adalah skala pengukuran berbentuk angka yang digunakan untuk menggolongkan suatu objek atau karakteristik. Sedangkan ordinal adalah skala pengukuran berbentuk angka selain berguna sebagai nominal juga menunjukkan urutan.

Analisis klaster didasarkan pada ukuran kemiripan atau ketidakmiripan antar data. Ukuran kemiripan atau ketidakmiripan yang digunakan adalah jarak (*distance*). Jika $d(p,q)$ menyatakan jarak (*dissimilarity*) antar objek p dan q maka terdapat beberapa sifat-sifat jarak (Johnson & Winchern, 2002) yaitu:

i. $d(p,q) = d(q,p)$ untuk semua p dan q

ii $d(p,q) > 0$ jika $p \neq q$

iii $d(p,q) = 0$ jika dan hanya jika $p = q$

iv $d(p,q) \leq d(p,r) + d(r,q)$ untuk semua objek p, q dan r

Metode Pengklasteran Hirarki dengan Berbagai Pengukuran Jarak

Pengukuran Jarak Pada Pengklasteran Hirarki

Jarak *Euclid*

Jarak *Euclid* adalah ukuran ketidakmiripan yang sering digunakan, merupakan jarak geometris diruang multidimensional (Anonim, 2008c). Jarak ini digunakan jika variabel-variabel yang digunakan tidak berkorelasi satu sama lain atau saling ortogonal, yang memiliki satuan dan skala pengukuran yang sama (Anonim, 2008e). Jarak ini cukup fleksibel untuk dilakukan modifikasi dalam mengatasi kelemahan data. Jarak *Euclid* merupakan jarak terpendek yang didapat antara dua titik dalam perhitungan.

Ukuran jarak *Euclid* antar dua objek $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ dan $\mathbf{y}' = [y_1, y_2, \ldots, y_p]$ yang berdimensi p adalah (Johnson & Winchern, 2002):

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots (x_p - y_p)^2}$$

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p}\left(x_i - y_i\right)^2} \tag{3}$$

Jarak Kuadrat *Euclid*

Jarak kuadrat *Euclid* adalah jumlah kuadrat yang berbeda dari nilai antara dua objek pada seluruh variabel (Fiedling, 2007). Jarak kuadrat *Euclid* antara dua objek x dan y pada ruang berdimensi *p* adalah

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}\left(x_i - y_i\right)^2 \tag{4}$$

Jarak *Mahalanobis*

Jika diantara variabel-variabel yang digunakan saling berkorelasi maka perlu dilakukan transformasi data. Transformasi ini berfungsi untuk menghilangkan pengaruh keragaman data atau dengan kata lain, semua variabel memberikan kontribusi yang sama untuk jarak (Anonim, 2008a). Jika tidak dilakukan transformasi data dapat digunakan jarak *Mahalanobis*.

Jarak ini menggunakan variabel dengan sampel matriks varian-kovarian, karena matriks kovarian juga menggunakan rata-rata korelasi diantara variabel. Jarak *Mahalanobis* antara objek x dan y dapat dinyatakan dalam bentuk

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\left(x-y\right)' S^{-1}\left(x-y\right)} \tag{5}$$

dimana S^{-1} adalah invers matriks varian-kovarian.

Jarak *Minkowski*

Jarak *Minkowski* ditemukan oleh Herman Minkowski (1864-1909) (Anonim, 2008d). Pengukuran jarak *Euclid* merupakan turunan umum dari rumus jarak *Minkowski*, yaitu jika $m = 2$. Jarak *Minkowski* dapat dinyatakan sebagai berikut:

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p}\left|x_i - y_i\right|^m\right]^{\frac{1}{m}} \tag{6}$$

dimana *m* adalah parameter. Jarak *Minkowski* menghasilkan penekanan yang lebih besar pada perbedaan antar koordinat jika dipilih $m > 1$.

Jarak *City-Block* atau *Manhattan*

Jarak *Block* atau *Manhattan* adalah jumlah nilai perbedaan mutlak untuk setiap variabel (Supranto, 2004). Ukuran jarak ini menghasilkan jarak yang serupa dengan jarak *Euclid* untuk beberapa kasus tertentu. Jarak ini mempunyai kelebihan yaitu dapat mendeteksi keberadaan outlier dengan baik (Agusta, 2007). Jarak *City-Block* atau *Manhattan* diturunkan dari persamaan (6) untuk $m = 1$, sehingga jarak *City-Block* atau *Manhattan* dinyatakan sebagai berikut:

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \sum_{i=1}^{p}\left|x_i - y_i\right| \\ &= \left|x_1 - y_1\right| + \left|x_2 - y_2\right| + \cdots + \left|x_p - y_p\right| \end{aligned} \tag{7}$$

Jarak *Chebychev*

Jarak *Chebychev* antara dua objek adalah nilai perbedaan mutlak yang maksimum pada tiap variabel. Pengukuran jarak ini sangat sensitif terhadap objek yang mempunyai outlier. Jarak *Chebychev* dinyatakan dalam bentuk

$$d(\mathbf{x},\mathbf{y}) = \max\left|x_i - y_i\right| \tag{8}$$

Jarak *Canberra*

Jarak *Canberra* adalah jumlah nilai perbedaan mutlak dibagi dengan jumlah antara dua variabel. (Johnson & Wichern, 2002). Jarak ini dapat dinyatakan dalam bentuk sebagai berikut:

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{\left|x_i - y_i\right|}{\left(x_i + y_i\right)} \tag{9}$$

Ukuran jarak *Canberra* digunakan hanya untuk variabel yang bernilai positif.

Jarak *Czekanowski*

Jarak *Czekanowski* merupakan selisih antara satu dengan dua kali jumlah minimum dari nilai variabel satu dengan variabel lainnya dibagi dengan jumlah antara dua variabel. Jarak ini dapat dirumuskan sebagai berikut (Johnson & Wichern, 2002):

$$d(\mathbf{x},\mathbf{y}) = 1 - \frac{2\sum_{i=1}^{p} \min\left(x_i, y_i\right)}{\sum_{i=1}^{p}\left(x_i + y_i\right)} \tag{10}$$

Jarak *Czekanowski* juga digunakan untuk variabel yang bernilai positif.

Macam-macam Metode Pengklasteran Hirarki

Pengklasteran hirarki atau tingkatan merupakan salah satu metode dari beberapa jenis analisis klaster. Metode pengklasteran hirarki atau *hierarchical clustering methods* adalah metode yang digunakan untuk mencari struktur pengklasteran dari objek-objek dan banyaknya klaster yang dibentuk belum diketahui.

Proses pengklasteran diawali dengan melihat setiap objek n sebagai satu klaster, sehingga jumlah klaster sebanyak jumlah objeknya. Dua objek atau klaster yang sangat mirip adalah objek yang pertama kali digabungkan menjadi satu klaster, sehingga jumlah klaster menjadi $n-1$. Jarak klaster baru dengan klaster sebelumya dihitung kembali. Demikian seterusnya sehingga klaster akan membentuk semacam dendogram atau grafik pohon, dimana ada hirarki atau tingkatan yang jelas antar objek dari yang sangat mirip sampai dengan yang tidak mirip. Sehingga semua objek pada akhirnya membentuk satu klaster.

Metode Penggabungan (*Agglomerative Methods*)

Metode penggabungan sering juga disebut sebagai metode aglomeratif. Metode ini adalah metode yang masing-masing objek dianggap satu klaster tersendiri, kemudian objek-objek yang sangat mirip berdasarkan ukuran jarak yang terdekat bergabung menjadi satu klaster.

Metode aglomeratif terdiri dari tiga metode yaitu: metode pautan (*Linkage Methods*), *Centroid*, dan Varian (*Ward Methods*).

Metode Pautan (*Linkage Methods*)

Menurut Johnson & Winchern (2002), metode pautan (*Linkage Methods*) terdiri dari tiga jenis, yaitu: pautan tunggal (*Single Linkage*), pautan lengkap (*Complete Linkage*), dan pautan rata-rata (*Average Linkage*).

1. Metode Pautan Tunggal (*Single Linkage Methods*)

Metode pautan tunggal dilakukan dengan meminimumkan jarak antar klaster yang digabungkan, atau dengan kata lain metode yang mengklasterkan dua objek yang mempunyai jarak terdekat terlebih dahulu. Jarak antar klaster di bentuk dari individu-individu dalam dua klaster yang mempunyai jarak terkecil atau kemiripan terbesar. Metode ini dimulai dengan menentukan jarak terkecil dalam matriks *proximity* $\mathbf{D} = \{d_{ik}\}$ dan gabungkan objek-objek yang memiliki jarak terkecil tersebut, misal U dan V, sehingga diperoleh klaster(UV). Untuk mencari klaster antara (UV) dan klaster W lainnya dapat diperoleh dengan cara sebagai berikut:

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\} \tag{11}$$

keterangan:

d_{UW} adalah jarak terdekat dari klaster U dan W

d_{VW} adalah jarak terdekat dari klaster V dan W

2. Metode Pautan Lengkap (*Complete Linkage Methods*)

Metode ini hampir sama dengan metode pautan tunggal hanya saja metode pautan lengkap menggunakan jarak terjauh. Metode pautan lengkap adalah metode yang mengklasterkan dua objek yang mempunyai jarak terjauh terlebih dahulu. Pengklasteran dimulai dengan mencari jarak pada matriks *proxmity* $\mathbf{D} = \{d_{ik}\}$ dan penggabungan antara U dan V untuk mendapatkan klaster pertama atau (UV). Selanjutnya jarak antara (UV) dan setiap klaster dihitung dengan:

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\} \tag{12}$$

keterangan:

d_{UW} adalah jarak terjauh dari klaster U dan W

d_{VW} adalah jarak terjauh dari klaster V dan W

3. Metode Pautan Rata-rata (*Average Linkage Methods*)

Metode ini mengklasterkan objek berdasarkan jarak rata-rata yang didapat dengan melakukan rata-rata semua jarak antar objek terlebih dahulu. Jarak antara dua klaster didefinisikan sebagai rata-rata jarak antara semua pasangan objek, di mana salah satu anggota dari pasangan berasal dari setiap klaster. Jarak metode pautan rata-rata dinyatakan dengan:

$$d_{(UV)W} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n} d_{ik}}{N_{(UV)}N_W} \tag{13}$$

keterangan:

d_{ik} adalah jarak antara objek i pada klaster (UV) dengan objek k pada klaster W

$N_{(UV)}$ adalah jumlah objek pada klaster (UV)

N_W adalah jumlah objek pada klaster W

Metode *Centroid*

Metode *centroid* adalah metode yang menggunakan rata-rata jarak pada sebuah klaster, yang diperoleh dengan cara menghitung rata-rata pada setiap variabel untuk semua objek. Pada metode ini, jarak antar dua klaster adalah jarak diantara dua *centroid* klaster-klaster tersebut. Dengan metode ini, setiap terjadi klaster baru segera terjadi perhitungan ulang *centroid* sampai terbentuk klaster yang tetap (Sokal & Michener, 1958 *dalam* Seber, 1984).

Jika $\bar{x}_1 = \sum_{i \in U} \frac{x_i}{N_1}$ adalah *centroid* dari objek-objek dalam klaster U dan $\bar{x}_2 = \sum_{i \in U} \frac{x_i}{N_2}$ adalah *centroid* dari objek-objek dalam klaster V, maka jarak antara dua klaster didefinisikan sebagai berikut:

$$d_{(UV)W} = d(\bar{x}_1, \bar{x}_2) \tag{14}$$

dimana d adalah ukuran matriks *proximity*

Centroid klaster baru yang terbentuk didapat dengan rumus

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} \tag{15}$$

keterangan:

$N_1 = N_2$ adalah banyaknya objek

Metode Varian

Metode varian sering juga disebut sebagai metode *Ward*. Metode ini menggunakan perhitungan yang lengkap dan memaksimumkan homogenitas di dalam satu klaster. Homogenitas atau kemiripan pada setiap klaster diukur dengan jumlah kuadrat objek-objek di dalam klaster, yang lebih dikenal sebagai *error sum of squares* (ESS). Nilai ESS digunakan sebagai fungsi objektif dan didefinisikan sebagai berikut (Dillon & Goldstein (1984)):

$$ESS = \sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} x_{ij}{}^2 - \frac{1}{n_j}\left(\sum_{i=1}^{n_j} x_{ij}\right)^2\right) \tag{16}$$

Keterangan

x_{ij} : Nilai objek ke- i pada klaster ke- j

k : Jumlah klaster setiap *stage*

n_j : Jumlah objek ke- i pada klaster ke- j

Ward (1963) *dalam* Anonim (2008a), mengusulkan penggunaan metode yang didasarkan pada hasil informasi yang minimum dari kenaikan pada jumlah kuadrat deviasi rata-rata klaster. Metode *Ward* juga dikenal dengan metode varian minimum dan harus menggunakan jarak kuadrat *Euclid* dengan menggunaakan bantuan komputer.

Metode Pemecahan (*Devisive Methods*)

Metode pemecahan sering juga disebut sebagai metode devisif. Metode ini merupakan kebalikan dari metode aglomeratif. Metode devisif yaitu metode yang pada mulanya seluruh

objek dianggap berada dalam satu klaster. Kemudian klaster tersebut dipecah menjadi beberapa klaster kecil. Objek yang berada dalam satu klaster kecil sangat berbeda dengan objek yang berada pada klaster kecil lainnya. Selanjutnya setiap klaster kecil dipecah lagi menjadi beberapa klaster kecil berdasarkan ukuran ketidakmiripannya.

William & Lance *dalam* Seber (1984), menyatakan metode devisif lebih menguntungkan dibandingkan dengan metode aglomeratif. Metode devisif tidak dilanjutkan ketika didapat n klaster yang mempunyai satu objek. Jika terdapat jumlah variabel yang lebih sedikit dibandingkan objek, maka perhitungan yang dibutuhkan juga sedikit yaitu d^2 jika dibandingkan dengan perhitungan devisif yang didekati oleh $(n-1)^2$. Tetapi metode ini jarang digunakan karena tidak semua *software* menyediakan fasilitas metode devisif (Anonim, 2008a).

Aplikasi Metode Pengklasteran Hirarki

Deskripsi Data

Data yang digunakan pada penelitian ini adalah data sekunder,pengujian nilai awal, Pada data tersebut tidak terlihat adanya variasi dalam satuan atau tidak terdapat perbedaan yang mencolok. Perbedaan data yang besar akan membuat perhitungan menjadi tidak valid (Santoso, 2004). Oleh karena itu, data tidak perlu ditransformasi.

Sebelum dilakukan pengklasteran hirarki terlebih dahulu dilakukan pengujian terhadap asumsi-asumsi yang harus dipenuhi. Asumsi-asumsi tersebut yaitu sampel yang diambil harus mewakili populasi yang ada dan tidak terjadi multikolinearitas.

Penelitian ini menggunakan total sampling yaitu populasi. Dengan demikian asumsi bahwa sampel yang diambil harus benar-benar dapat mewakili populasi yang ada telah terpenuhi.

Multikolinearitas terjadi jika ada hubungan linier yang sempurna di antara beberapa atau semua variabel bebas (Rahardiantoro, 2008). Untuk mengetahui ada tidaknya multikolinearitas antar variabel dilakukan dengan menggunakan matriks korelasi (Lampiran 2). Untuk nilai korelasi negatif yang digunakan adalah nilai mutlaknya (Winchern, 2002).

Proses dan Hasil Pengklasteran

Proses pengklasteran hirarki dilakukan dengan menggunakan *software* SPSS versi 11.5 *for Windows* dan Microsoft EXCEL. Pemilihan *software* SPSS dikarenakan hampir setiap literatur menyarankan menggunakan *software* ini. Sedangkan Microsoft EXCEL digunakan karena ada tiga metode jarak yang tidak terdapat di *software* SPSS. Metode jarak tersebut yaitu: jarak *Mahalanobis*, *Canberra*, dan *Czekanowski*.

Jarak *Mahalanobis* menggunakan sampel matriks varian-kovarian. Untuk mendapatkan nilai matriks varian-kovarian dengan jarak *Mahalanobis* dapat dicari dengan menggunakan *software* MINITAB 14.

Metode pengklasteran hirarki dilakukan dengan memperhatikan ukuran kemiripan antar objek yang terdapat dalam matriks *proximity*.

Metode pautan (*Linkage Methods*) terdiri dari pautan tunggal (*Single Linkage*), pautan lengkap (*Complete Linkage*), dan pautan rata-rata (*Average Linkage*). Pada metode ini tidak ada sumber yang mengharuskan menggunakan jarak tertentu untuk mendapatkan matriks *proximity*nya. Untuk itulah digunakan semua metode jarak.

Hasil pengklasteran berdasarkan konsep pengukuran jarak menunjukkan bahwa jarak yang berbeda memberikan hasil pengklasteran yang hampir sama. Dari analisis yang dilakukan diperoleh hasil pengklasteran dengan metode pengukuran jarak *Euclid*, Kuadrat *Euclid*, dan *Minkowski* memberikan hasil pengklasteran yang sama persis pada setiap metode pengklasteran hirarki. Sedangkan pengukuran jarak *City-Block* atau *Manhattan*, *Chebychev*, *Mahalanobis*, *Canberra*, dan *Czekanowski* menghasilkan pengklasteran yang berbeda, namun perbedaannya tidak terlalu mencolok.

Output metode pengklasteran hirarki disajikan secara visual berbentuk dendogram yaitu suatu bagan yang menyajikan banyaknya klaster terbesar hingga terkecil. Dendogram masing-masing metode dengan berbagai konsep pengukuran jarak yang menghasilkan pengklasteran yang berbeda-beda dapat dilihat pada Lampiran 4.

Dendogram dibaca dari kiri ke kanan. Garis tegak lurus menunjukkan objek yang digabung membentuk satu klaster. Sedangkan posisi garis pada skala menunjukkan jarak klaster yang digabung. Adapun skala yang digunakan bukanlah koefisien yang ada pada tabel *Agglomeration*, namun telah dilakukan proses skala ulang dengan batasan 0 sampai 25. Proses agglomerasi dimulai pada skala 0, dengan ketentuan jika sebuah garis dekat dengan angka 0, maka variabel-variabel yang mewakili garis tersebut semakin membentuk sebuah klaster.

Tabel Agglomerasi di bawah ini menunjukkan cara penggabungan klaster setiap tahap pada metode pautan tunggal (*Single Linkage*) dengan menggunakan jarak *Euclid*.

Pada tahap pertama terbentuk satu klaster dengan objek 5 (Kaur) dan objek 9 (Kepahiang). Nilai pada kolom koefisien menyatakan jarak antar objek seperti yang terlihat pada matriks *proximity*. Nilai koefisien tersebut didapat dari jarak yang digunakan. Karena proses agglomerasi dimulai dengan menggabungkan dua variabel yang terdekat, maka jarak kedua variabel tersebut merupakan jarak terdekat dari sekian banyak kombinasi jarak dari 9 objek. Sehingga semakin kecil nilai koefisien pada tabel agglomerasi berarti semakin mirip atau dekat satu objek dengan objek yang lain. Sebaliknya, semakin besar nilai koefisien pada tabel agglomerasi berati semakin tidak mirip satu objek dengan objek yang lain.

Kolom *next stage* menyatakan tahapan lanjutan dari penggabungan objek dengan objek yang baru saja digabungkan. Pada tahap pertama terlihat angka 2. Hal ini berarti langkah pengklasteran selanjutnya adalah menggabungkan objek 2 dengan klaster yang baru saja terbentuk, yaitu objek 5 dan 9. Demikian seterusnya sampai langkah terakhir yakni langkah 8.

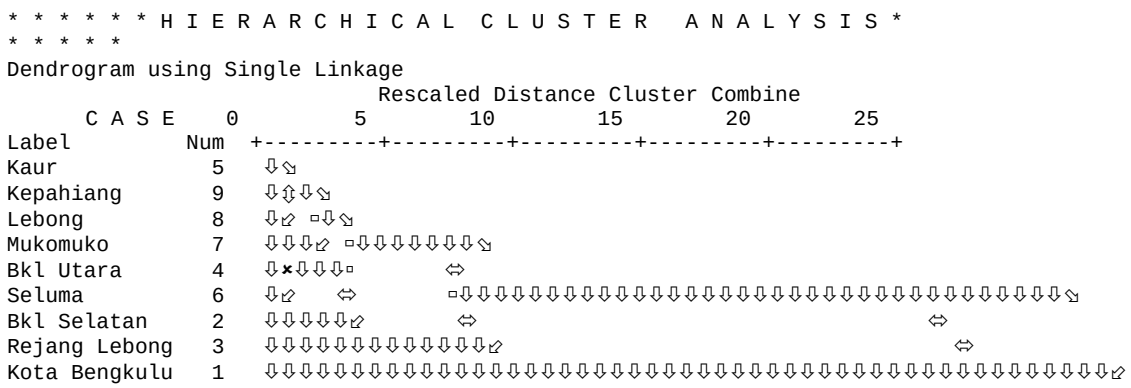
Pada dendogram metode pautan tunggal (*Single Linkage*) dengan jarak *Euclid*, terlihat sebagai tahap pertama objek (Kaur, Lebong, dan Kepahiang) membentuk satu klaster tersendiri, karena mempunyai panjang garis yang sama dan tergabung menjadi satu kesatuan. Demikian juga dengan objek (Bengkulu Utara dan Seluma) membentuk klaster tersendiri. Hasil dendogram metode ini dapat dilihat pada Gambar 4.1.

Sebaliknya objek (Mukomuko, Bengkulu Selatan, Rejang Lebong, dan kota Bengkulu) tidak tergabung dengan objek-objek sebelumnya, karena mempunyai garis yang lebih panjang. Dengan demikian, pada proses pertama telah terbentuk enam klaster yaitu: dua klaster yang mempunyai anggota lebih dari satu objek (Kaur, Lebong, dan Kepahiang), (Bengkulu Utara dan Seluma), dan empat klaster berdiri sendiri.

Kemudian proses dilanjutkan, objek (Bengkulu Selatan) bergabung dengan objek (Bengkulu Utara dan Seluma). Demikian seterusnya, proses ini berjalan ke arah kanan, dengan menggunakan petunjuk panjang garis yang semakin ke kanan. Sehingga pada akhirnya, semua objek akan tergabung menjadi satu klaster.

Tabel 4.2 Agglomerasi Metode *Single Linkage* dengan Jarak *Euclid*

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	9	2.194	0	0	2
2	5	8	2.849	1	0	4
3	4	6	2.963	0	0	5
4	5	7	3.358	2	0	6
5	2	4	4.711	0	3	6
6	2	5	4.814	5	4	7
7	2	3	8.191	6	0	8
8	1	2	25.383	0	7	0



Gambar 4.1 Dendogram Metode *Single Linkage* dengan Jarak *Euclid*

Hasil Pengklasteran Metode Hirarki

Jika yang diambil 3 klaster sebagai pemecahannya, maka:

- Untuk semua konsep pengukuran jarak pada metode pautan tunggal (*Single Linkage*), maka didapat anggota klasternya sebagai berikut: (kota Bengkulu), (Rejang Lebong), dan (kabupaten-kabupaten selain Bengkulu dan Rejang Lebong).
- Metode pautan rata-rata (*Average Linkage*), berdasarkan konsep pengukuran jarak *Euclid*, Kuadrat *Euclid*, *City-Block* atau *Manhattan*, *Chebychev*, dan *Minkowski* anggota klasternya adalah (kota Bengkulu), (Rejang Lebong), dan (kabupaten-kabupaten selain kabupaten yang telah disebutkan pada klaster sebelumnya). Hasil metode ini juga sama dengan metode *Ward*, dan *Centroid*.
- Metode pautan rata-rata (*Average Linkage*) dengan konsep pengukuran jarak *Mahalanobis* anggota klasternya adalah (Bengkulu Selatan), (Rejang Lebong), dan (kabupaten-kabupaten selain kabupaten yang telah disebutkan pada klaster sebelumnya).
- Metode pautan rata-rata (*Average Linkage*) dengan konsep pengukuran jarak *Canberra* anggota klasternya adalah (kota Bengkulu), (Rejang Lebong, Bengkulu Utara, dan Lebong), dan (kabupaten-kabupaten selain kabupaten yang telah disebutkan pada klaster sebelumnya).
- Metode pautan rata-rata (*Average Linkage*) dengan konsep pengukuran jarak *Czekanowski* anggota klasternya adalah (kota Bengkulu, Bengkulu Selatan, Rejang Lebong), (Bengkulu Utara, Seluma), dan (kabupaten-kabupaten selain kabupaten yang telah disebutkan pada klaster sebelumnya).

Interpretasi Profil dan Akses Validitas Klaster

Hasil dari pemecahan metode pengklasteran hirarki diperoleh 3 pemecahan hasil pengklasteran dan pengambilan banyaknya klaster ini bersifat subjektif. Hasil pemecahan dari anggota metode pengklasteran hirarki dengan menggunakan metode pautan tunggal (*Single Linkage*), pautan lengkap (*Complete Linkage*), pautan rata-rata (*Average Linkage*), *Ward* method, dan metode *centroid* berdasarkan konsep pengukuran jarak Kuadrat *Euclid*, menghasilkan *Cluster membership* yang sama.

Hasil pemecahan pengklasteran dengan metode pautan tunggal (*Single Linkage*) berdasarkan konsep pengukuran jarak *Euclid*, *Minkowski*, *Chebychev*, *City-Block* atau *Manhattan*, dan *Canberra*, diperoleh hasil pemecahan pengklasteran yang sama persis dengan pengukuran jarak Kuadrat *Euclid*. Tetapi, pada pengukuran jarak *Mahalanobis* dan *Czekanowski* menghasilkan *cluster membership* yang berbeda dengan jarak-jarak yang lain.

Pada metode pautan rata-rata (*Average Linkage*) dengan menggunakan konsep pengukuran jarak *Euclid*, *Minkowski*, *Chebychev*, *City-Block* atau *Manhattan*, juga diperoleh hasil pemecahan *cluster membership* yang sama persis dengan jarak Kuadrat *Euclid*. Tetapi, pada pengukuran jarak *Mahalanobis*, *Canberra*, dan *Czekanowski* menghasilkan *cluster membership* yang berbeda dengan jarak-jarak yang lain.

Pada metode pautan lengkap (*Complete Linkage*) hanya pengukuran jarak *City-Block* atau *Manhattan* yang hasil pemecahan *cluster membership* sama dengan Kuadrat *Euclid*. Sedangkan pada pengukuran jarak-jarak yang lain hasil pemecahan *cluster membership* berbeda-beda.

Kesimpulan dan Saran

Dari delapan pengukuran jarak yang digunakan dengan data yang ada, jarak *Mahalanobis* dan *Czekanowski* menghasilkan pengklasteran yang sangat berbeda. Hal ini terlihat dari hasil dendogram dan tabel *cluster membership*. Jarak *Euclid* dan *Minkowski* menghasilkan pengklasteran yang sama, karena pengukuran jarak *Euclid* merupakan turunan umum dari jarak *Minkowski*. Selanjutnya, hasil analisis menunjukkan bahwa jarak yang paling baik adalah jarak Kuadrat *Euclid* dan *City-Block* atau *Manhattan*, karena setiap metode yang digunakan menghasilkan pengklasteran yang sama, baik dilihat dari dendogram maupun *cluster membership*nya. Tetapi, pada metode *Ward* dan *Centroid* tidak menggunakan jarak *City-Block* atau *Manhattan* melainkan harus menggunakan jarak Kuadrat *Euclid*.

Untuk mengetahui ada atau tidaknya perbedaan yang berarti dalam pengklasteran berdasarkan konsep pengukuran jarak yang berbeda-beda, disarankan untuk melanjutkan penelitian ini pada metode nonhirarki seperti *fuzzy clustering*

DAFTAR PUSTAKA

Agusta, Y. 2007. *K-Means-Penerapan, Permasalahan, dan Metode Terkait*. http://www.yudiagusta.file.wordpress.com/2008/03/K-Means.pdf.

Anonim. 2000. *An Introduction to Cluster Analysis for Data Mining*. http://www.clustan.com/whatisclusteranalysis.html-11k.

Anonim. 2008a. *Cluster Analysis*. http://paleo.cortland.edu/class/stats/lecture Notes/11-cluster.pdf.

Anonim. 2008b. *Cluster Analysis*. http:// www. ilmustatistik.org.

Anonim. 2008c. *Cluster Analysis*. http://www.stasoft.com/textbook/stcluan.html.

Anonim. 2008d. *Statistical Analysis of Microarray Data*. http://www-users.itlabs.umn.edu/clases/spring-2008/csci5980-FGB/lecture10.ppt.

Johnson, R.A and D.W. Wichern. 2002. *Applied Multivariate Statistical Analysis*. 5th edition. Prentice Hall. New Jersey. http://faculty.smu.edu/tfomby/eco5385/lecture/Scoring%20Measures%20for%20Prediction%20Problems.pdf

Rencher, A.C. 2002. *Methods of Multivariate Analysis*. 2nd edition. John Wiley & Sons, Inc. New York

Rahardiantoro, D. 2008. *Principal Component Analysis (PCA) Sebagai Metode Jitu untuk Mengatasi Masalah Multikolinearitas*. http://dickyrahardi.wordpress.com/2006/2/9/principal-component-analysis-pca-sebagai-metode-jitu-untuk-mengatasi multikolinearitas.

Santoso, S. 2004. *Buku Latihan SPSS Statistik Multivariat*. Elex Media Komputindo. Jakarta.

Seber, G.A.F. 1984. *Multivariate Observation*. John Wiley & Sons, Inc. NewYork.