Metode Berbasis Model (*Model-Based*) dalam Analisis Cluster

Oleh

Timbul Pardede

FMIPA
UNIVERSITAS TERBUKA
2005

Metode Berbasis Model (*Model-Based*) dalam Analisis Cluster

Mengetahui,
Ketua
Jurusan Statistika FMIPA UT

Ir. Isfarudi, M.Pd
131600400

Pondok Cabe Juli 2005

Disusun Oleh,

Timbul Pardede
131957295

ABSTRAK

TIMBUL PARDEDE. Metode Berbasis Model (*Model-Based*) dalam Analisis Gerombol.

Metode cluster berbasis model adalah metode cluster yang didasarkan pada aspek statistik, yaitu kriteria kemungkinan maksimum. Metode cluster berbasis model ini mempunyai beberapa model dengan berbagai macam sifat geometris yang diperoleh melalui komponen Gauss. Penyekatan data dilakukan dengan menggunakan kemungkinan maksimum melalui algoritma Ekspektasi-Maksimum (EM), kemudian dengan pendekatan model Bayes berdasarkan *Bayesian Information Criterion* (BIC) diperoleh model terbaik. Penelitian ini bertujuan untuk mengkaji hasil pengelompokan metode gerombol berbasis model. Hasil pengelompokan dari 2 contoh data penerapan (data Iris dan data Diabetes) menunjukkan bahwa metode gerombol berbasis model cukup efektif memisahkan kelompok-kelompok yang saling tumpang tindih.

Kata kunci Metode cluster berbasis model, algoritma EM, BIC.

PENDAHULUAN

Latar Belakang

Analisis cluster merupakan suatu metode cluster satuan objek pengamatan menjadi beberapa kelompok objek pengamatan berdasarkan peubah-peubah yang dimiliki, sedemikian sehingga objek-objek yang terletak dalam kelompok yang sama relatif lebih homogen dibandingkan dengan objek-objek pada kelompok yang berbeda.

Dewasa ini terdapat beberapa metode cluster yang dapat dikelompokkan berdasarkan algoritma proses yang dilakukan, yakni teknik yang berdasarkan ukuran jarak sebagai basis pengelompokannya. Metode berbasis ukuran jarak ini terdiri dari metode cluster berhierarki dengan penggabungan (*agglomerative*), antara lain metode pautan tunggal (*single linkage*), metode pautan lengkap (*complete linkage*), metode pautan rataan (*average linkage*), metode terpusat (*centroid*) dan metode Ward (*Ward's method*) dan juga metode cluster tak berhierarki, misalnya metode K-rataan (Anderberg 1973). Metode cluster ini memiliki teknik-teknik yang berbeda-beda dalam proses pembentukan kelompok, namun teknik-teknik tersebut hanya memperhatikan ukuran jarak antar objek-objek pengamatan. Metode-metode ini tidak mempertimbangkan aspek statistiknya, seperti sebaran datanya.

Metode cluster berbasis model (*model-based*) adalah suatu metode yang berbeda dengan metode cluster yang didasarkan pada ukuran jarak. Metode ini merupakan suatu algoritma pengelompokan yang tergolong baru, analisis dilakukan berdasarkan pada aspek statistik yaitu kriteria kemungkinan maksimum di dalam memutuskan hasil pengelompokan. Metode ini mempunyai beberapa model dengan berbagai macam sifat geometris yang diperoleh melalui komponen Gauss dengan parameter yang berbeda-beda. Penyekatan data dilakukan dengan menggunakan kemungkinan maksimum melalui algoritma Ekspektasi-Maksimum (EM), kemudian dengan pendekatan model Bayes berdasarkan *Bayesian Information Criterion* (BIC) diperoleh model terbaik (Fraley & Raftery, 1998). Oleh karena metode cluster

berbasis model ini masih tergolong baru, hasil pengelompokan dengan metode cluster berbasis model menjadi suatu hal yang perlu dan menarik untuk dikaji.

Tujuan

Berdasarkan permasalahan di atas, maka tujuan makalah ini adalah untuk mengkaji hasil pengelompokan metode berbasis model dalam analisis cluster.

TINJAUAN PUSTAKA

Metode cluster Berbasis Model

Model Campuran

Dalam pengelompokan berbasis model (Fraley & Raftery, 1998), diasumsikan bahwa data dibangkitkan oleh sebaran peluang campuran dengan setiap subpopulasi mewakili suatu cluster yang berbeda. Misalkan $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ adalah contoh acak peubah ganda-p dari suatu populasi, dimana p menyatakan dimensi data dan n menyatakan banyaknya objek pengamatan. Ke-n objek pengamatan ini dianggap berasal dari campuran G subpopulasi $G_1, G_2, ..., G_g$ yang masing-masing terdiri atas n_j data dengan $\sum_{j=1}^{g} n_j = n$.

Secara umum fungsi kepekatan peubah acak ganda ini dapat dinyatakan sebagai fungsi kepekatan campuran berhingga

$$f(\mathbf{x}|\phi) = \sum_{j=1}^{g} \pi_j f_j(\mathbf{x}|\theta) \quad ; \quad \phi \in \Omega \qquad (1)$$

dimana $f_j(\mathbf{x}|\theta)$ merupakan fungsi kepekatan G_j, yaitu subpopulasi ke-j dengan vektor parameter θ yang tidak diketahui dan π_j merupakan proporsi data yang berasal dari subpopulasi ke-j dengan $\sum_{j=1}^{g} \pi_j = 1$ dan $\pi_j > 0 \quad (j = 1,2,.....,g)$, sedangkan $\phi = (\pi', \theta')'$ adalah gugus semua parameter dari fungsi kepekatan campuran yang berasal dari ruang parameter Ω (Mclachlan & Basford, 1988).

Dengan asumsi $\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_n$ bebas stokastik dan identik dengan fungsi kepekatan $f_j(\mathbf{x}_i|\theta_j)$ merupakan fungsi kepekatan pengamatan $\mathbf{x}_i$ dari cluster ke-j. Fungsi kemungkinan sebaran campuran (*mixture likelihood*) pada persamaan (1) adalah :

$$L(\phi) = \prod_{i=1}^{n}\left[\sum_{j=1}^{g}\pi_j f_j\left(\mathbf{x}_i\middle|\theta_j\right)\right] \quad (2)$$

Dalam penelitian ini difokuskan pada suatu kasus dimana $f_j\left(\mathbf{x}_i\middle|\theta_j\right)$ adalah fungsi kepekatan peubah ganda campuran normal (*Gauss*) dengan parameter θ_j terdiri dari vektor rataan μ_j dan matriks koragam Σ_j, yang dinyatakan dalam bentuk :

$$f_j(x_i|\mu_j;\Sigma_j) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{x_i}-\mu_j\right)'\Sigma_j^{-1}\left(\mathbf{x_i}-\mu_j\right)\right\}}{(2\pi)^{\frac{p}{2}}\left|\Sigma_j\right|^{\frac{1}{2}}} \quad (3)$$

Dalam pengelompokan berbasis model, diasumsikan bahwa data dibangkitkan dengan fungsi kepekatan peubah ganda campuran. Data bangkitan tersebut dicirikan oleh cluster-cluster *ellipsoidal* yang terpusat pada rataan μ_j (Fraley & Raftery, 2000). Karakteristik geometrik (bentuk *(shape),* volume, dan orientasi *(orientation)*) cluster dihitung dengan matriks koragam Σ_j, yang diparameterisasikan untuk menentukan batasan antar cluster.

Branfield & Raftery (1993) mengembangkan pengelompokan berbasis model dengan memparameterisasikan setiap matriks koragam melalui suku-suku dekomposisi nilai ciri dalam bentuk :

$$\Sigma_j = \lambda_j D_j A_j D_j' \quad (4)$$

dimana :

D_j adalah matriks ortogonal dari vektor ciri, yang menjelaskan orientasi dari komponen ke-j,

A_j adalah matriks diagonal dengan masing-masing unsurnya proporsional terhadap nilai ciri dari Σ_j, yang menjelaskan bentuk,

λ_j adalah skalar yang menjelaskan volume.

Pencirian sebaran geometrik (orientasi, volume, bentuk) mungkin akan diperoleh dari berbagai macam bentuk cluster, atau terbatas pada cluster yang sama. Sebagai

illustrasi, model $\sum_j = \lambda I$ mempunyai volume sama dan semua cluster berbentuk bola (*spherical*). Model $\sum_j = \lambda DAD'$ mempunyai ciri geometrik sama dan semua cluster berbentuk *ellipsoidal*. Model $\sum_j = \lambda_j D_j A_j D_j'$ mempunyai model tanpa batasan dimana setiap cluster mempunyai ciri geometrik yang berbeda. Tabel-1 menunjukkan matriks koragam $\sum_j$ untuk model campuran normal ganda dan interpretasi geometrik (Banfield dan Raftery, 1998).

Tabel 1. Interpretasi geometrik dan parameterisasi matriks koragam $\sum_j$ dalam model campuran normal ganda.

$\sum_j$	**Volume**	**Bentuk Geometri**	**Orientasi**	**Tebaran**	**Simbol Mclust**
λI	Sama	Sama	-	*Spherical*	**EI**
$\lambda_j I$	Berbeda	Sama	-	*Spherical*	**VI**
$\lambda DAD'$	Sama	Sama	Sama	*Ellipsoidal*	**EEE**
$\lambda_j D_j A_j D_j'$	Berbeda	Berbeda	Berbeda	*Ellipsoidal*	**VVV**
$\lambda D_j A D_j'$	Sama	Sama	Berbeda	*Ellipsoidal*	**EEV**
$\lambda_j D_j A D_j'$	Berbeda	Sama	Berbeda	*Ellipsoidal*	**VEV**

Penduga Kemungkinan Maksimum Model Campuran melalui Algorithma EM

Algoritma EM (*Expectation-Maximum*) merupakan metode perhitungan iterasi terhadap masalah pendugaan kemungkinan maksimum parameter pada data tidak lengkap. Model data lengkap $\mathbf{y}_i' = (\mathbf{x}_i', \mathbf{z}_i')$, dimana $\mathbf{z}_i' = (z_{i1}, z_{i2}, ..., z_{ig})$ merupakan vektor indikator yang didefinisikan dengan :

$$z_{ij} = \begin{cases} 1, & \mathbf{x}_i \in G_j \\ 0, & \text{lainnya.} \end{cases} \quad ; i = 1,...,n \quad ; j = 1,...,g \qquad (5)$$

Algoritma EM ini terdiri dari dua tahap yaitu tahap E untuk pendugaan dan tahap M untuk pemaksimuman. Dengan asumsi $\mathbf{Z}$ bebas dan identik menurut sebaran multinomial dengan peluang $\pi_1, \pi_2, ..., \pi_g$ dan fungsi kepekatan $\mathbf{x}_i$ dengan $\mathbf{z}_i$ adalah $\prod_{j=1}^{g} f_j(\mathbf{x}_i|\theta_j)^{z_{ij}}$, maka fungsi kemungkinan data lengkap (*complete-data likelihood*) adalah :

$$L(\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{z}\ |\mathbf{x}) = \prod_{i=1}^{n}\prod_{j=1}^{g}\left\{\pi_j f_j(\mathbf{x}_i|\theta_j)\right\}^{z_{ij}}$$

atau fungsi log kemungkinan data lengkap (*complete-data loglikelihood*) adalah :

$$\mathbf{L}(\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{z}\ |\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\left\{\ln\pi_j + \ln f_j(\mathbf{x}_i|\theta_j)\right\}$$

Jika $f_j(\mathbf{x}_i|\theta_j)$ merupakan model campuran sebaran normal ganda yaitu $f_j(\mathbf{x}_i|\theta_j) = f_j(\mathbf{x}_i|\boldsymbol{\mu}_j;\Sigma_j)$, maka fungsi log kemungkinan data lengkap pada model campuran normal ganda adalah:

$$\mathbf{L}(\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{z}\ |\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\left\{\ln\pi_j + \ln f_j(\mathbf{x}_i|\boldsymbol{\mu}_j;\Sigma_j)\right\} \qquad (6)$$

dengan tahap E pada iterasi EM pada model campuran normal ganda diperoleh

$$\hat{z}_{ij} = \frac{\hat{\pi}_j f_j\left(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_j,\hat{\Sigma}_j\right)}{\sum_{k=1}^{g}\hat{\pi}_k f_k\left(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_k,\hat{\Sigma}_k\right)} \qquad ;i=1,...,n\ ;j=1,...,g \qquad (7)$$

$\hat{z}_{ij}$ merupakan dugaan peluang akhir $\mathbf{x}_i$ dalam cluster ke-j. Penduga kemungkinan maksimum untuk parameter θ_j dan π_j diperoleh dengan memasukkan nilai $\hat{z}_{ij}$ ke dalam persamaan (6), yaitu :

$$\mathbf{L}^*(\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{z}\ |\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{g}\hat{z}_{ij}\left\{\ln\pi_j + \ln f_j(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_j;\hat{\Sigma}_j)\right\}$$

Kemudian $\mathbf{L}^*(\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{z}\ |\mathbf{x})$ dimaksimumkan dengan tahap M pada iterasi EM. Demikian proses iterasi ini berlangsung hingga diperoleh hasil iterasi yang konvergen. Tahapan

pendugaan dan pemaksimuman untuk kasus model campuran normal ganda diparameterisasikan melalui dekomposisi nilai ciri seperti pada persamaan (4).

Fraley dan Raftery (1998) membuat algoritma EM pada model campuran Gauss sebagai berikut :

Mulai

Tahapan E

Hitung $\hat{z}_{ij} = \dfrac{\hat{\pi}_j f_j\left(x_i \middle| \hat{\mu}_j, \hat{\Sigma}_j\right)}{\sum_{k=1}^{g} \hat{\pi}_k f_k\left(x_i \middle| \hat{\mu}_k, \hat{\Sigma}_k\right)}$ dimana f_j dari persamaan (3)

atau $\hat{z}_{ij} = \dfrac{\hat{\pi}_j \left|\hat{\Sigma}_j\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_i - \hat{\mu}_j\right)' \hat{\Sigma}_j^{-1}\left(\mathbf{x}_i - \hat{\mu}_j\right)\right\}}{\sum_{k=1}^{g} \hat{\pi}_k \left|\hat{\Sigma}_k\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_i - \hat{\mu}_k\right)' \hat{\Sigma}_k^{-1}\left(\mathbf{x}_i - \hat{\mu}_k\right)\right\}}$

Tahapan M

Maksimumkan $\hat{z}_{ij}$ dari persamaan (6)

$$n_j \longleftarrow \sum_{i=1}^{n} z_{ij}$$

$$\hat{\pi}_j \longleftarrow \frac{n_j}{n}$$

$$\hat{\mu}_j \longleftarrow \frac{\sum_{i=1}^{n} \hat{z}_{ij} \mathbf{x}_i}{n_j}$$

$\hat{\Sigma}_j$: sesuai dengan model pada Tabel 1.

Ulang Sampai kriteria konvergen dipenuhi.

Pemilihan Model Pengelompokan dengan Faktor Bayes

Dalam aplikasi analisis cluster ada dua masalah yang dihadapi, yaitu pemilihan metode cluster dan memutuskan jumlah cluster. Untuk menangani kedua masalah ini dilakukan pendekatan model campuran melalui faktor Bayes. Salah satu keuntungan pendekatan model campuran dengan menggunakan pendekatan faktor Bayes adalah dapat membandingkan antar model. Sistematika pemilihan tidak hanya untuk parameterisasi model (metode cluster yang digunakan), tetapi juga banyaknya cluster.

Misalkan $\mathbf{X}$ adalah data pengamatan, $\mathcal{M}_1$ dan $\mathcal{M}_2$ adalah dua model yang berbeda dengan parameter masing-masing adalah θ_1 dan θ_2. Integral atau marginal kemungkinan (*Integral or marginal likelihood*) didefinisikan sebagai

$$P(\mathbf{X}|\mathcal{M}_k) = \int P(\mathbf{X}|\theta_k, \mathcal{M}_k) P(\theta_k|\mathcal{M}_k) d\theta_k \qquad k = 1,2$$

dimana $P(\theta_k|\mathcal{M}_k)$ adalah sebaran awal θ_k, dengan θ_k adalah parameter model $\mathcal{M}_k$. Faktor Bayes (*Bayes Factor*) didefinisikan sebagai rasio dari integral kemungkinan dari kedua model, yakni $B_{12} = \frac{P(\mathbf{X}|\mathcal{M}_1)}{P(\mathbf{X}|\mathcal{M}_2)}$. Metode ini dikembangan secara umum untuk lebih dari dua model (Kass & Raftery, 1995).

Kesulitan utama dalam penggunaan faktor Bayes adalah perhitungan integral kemungkinannya. Kass & Raftery (1995) mengemukakan bahwa integral kemungkinannya dapat didekati dengan pendekatan faktor Bayes melalui algoritma EM. Pendekatan ini disebut dengan BIC (*Bayesian Information Criterion*) dengan formulasi sebagai berikut :

$$2 \ln P(\mathbf{X}|\mathcal{M}_k) \approx 2 \ln P\left(\mathbf{X}\middle|\hat{\theta}_k, \mathcal{M}_k\right) - V_k \ln(n) \equiv BIC_k$$

dimana

$P(\mathbf{X}|\mathcal{M}_k)$: adalah integral kemungkinan untuk model $\mathcal{M}_k$,

$P\left(\mathbf{X}\middle|\hat{\theta}_k, \mathcal{M}_k\right)$: adalah kemungkinan maksimum model campuran untuk model $\mathcal{M}_k$,

V_k : adalah banyaknya parameter bebas yang diduga pada model $\mathcal{M}_k$,

$\hat{\theta}_k$: adalah dugaan kemungkinan maksimum untuk parameter θ pada model $\mathcal{M}_k$.

Nilai BIC terbesar mengindikasikan bahwa model tersebut adalah model yang paling layak. Nilai BIC ini dapat digunakan untuk membandingkan model dengan parameterisasi matriks koragam yang berbeda dan banyaknya cluster yang berbeda.

Fraley & Raftery (1998) membuat strategi metode cluster berbasis model dengan cara mengkombinasikan pengelompokan berhierarki penggabungan, algoritma EM dan faktor Bayes dengan langkah-langkah sebagai berikut :

*. Tentukan banyak cluster maksimum (m), dan himpunan model campuran ganda normal.

*. Lakukan pengelompokan berhierarki penggabungan untuk setiap model campuran normal ganda. Hasil pengelompokan ini ditransformasi ke dalam peubah indikator pada persamaan (5), yang kemudian digunakan sebagai nilai awal untuk algoritma EM

*. Lakukan algoritma EM untuk setiap model dan masing-masing banyak cluster 2, 3, ….,m, yang diawali dengan klassifikasi pengelompokan berhierarki.

*. Hitung nilai BIC untuk kasus satu cluster pada setiap model dan untuk model kemungkinan campuran dengan parameter optimal dari algoritma EM untuk 2, 3, …., m cluster.

*. Plotkan nilai BIC untuk setiap model. Nilai BIC terbesar mengindikasikan bahwa model tersebut adalah model yang paling layak.

BAHAN DAN METODE PENELITIAN

Bahan Penelitian

Bahan atau data yang digunakan dalam makalah ini adalah data sekunder, yaitu data Iris dan data Diabetes yang banyak digunakan dalam buku-buku teks statistika peubah ganda dan dalam paket program statistika seperti *S-plus* dan *Minitab*. Data Iris dan Diabetes dapat dilihat pada Lampiran 1 dan Lampiran 2.

Metode Analisis

Perangkat lunak yang digunakan untuk menganalisis data dengan model berbasis model adalah paket program Mclust dengan *interface Splus 2000* dengan prosedur sebagai berikut :

1. Install perangkat lunak MCLUST ke S-plus2000 dengan cara :

- source("c:/mclust/mclust.s")
- dyn.load("c:/mclust//mclust.obj")

2. Masukkan data Iris dan data Diabetes ke dalam tabel *Splus 2000*.
3. Tentukan banyak kelompok maksium
4. Hitung nilai EM untuk setiap model dan untuk setiap kelompok
5. Hitung dan plotkan nilai BIC untuk masing-masing model
6. Tentukan model terbaik berdasarkan nilai BIC yang paling besar.

Perintah pengelompokan dengan metode berbasis mode untuk data diabetes :

- diabetes.m← matriks(aperm(diabetes,c(1,2)),145,3,dimnames(diabetes)[1:2])
- bicdiabetes<-emclust(diabetes.m,nclus=2:6,modelid= c("EI","VI","EEE","VVV","EEV","VEV"))
- sumdiabetes<-summary(bicdiabetes,diabetes.m)
- plot(bicdiabetes)

Pada Gambar 1 disajikan diagram alir metode analisis yang akan dilakukan pada data Iris dan data Diabetes.

Gambar 1. Diagram alir metode cluster dengan berbasis model

HASIL DAN PEMBAHASAN

Data Iris

Data Iris merupakan contoh klasik yang sering digunakan dalam buku-buku teks statistik untuk mengilustrasikan masalah pengklasifikasian. Data Iris ini adalah sejenis bunga yang terdiri dari 4 peubah yaitu, Panjang Petal (PP), Lebar Petal (LP), Panjang Sepal (PS) dan Lebar Sepal (LS). Masing-masing peubah terdiri dari 150 pengamatan, setiap ukuran peubah terbagi atas tiga spesies yaitu *Iris Setosa* (IS), *Iris Versicolor* (IC), dan *Iris Virginica* (IV) yang masing-masing terdiri dari 50 pengamatan (lihat Lampiran 1).

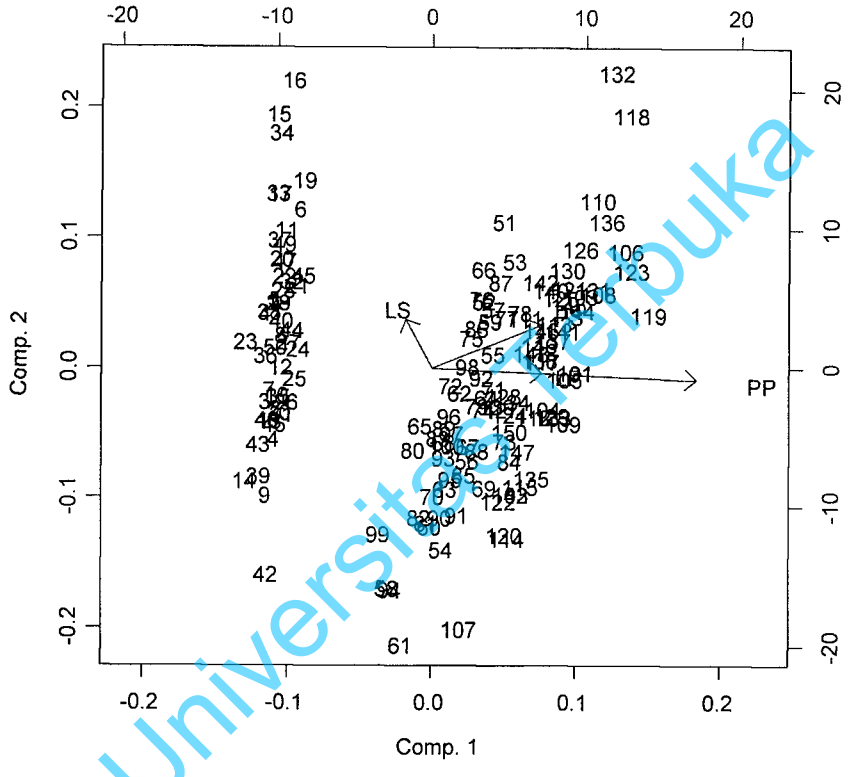
Sebelum menerapkan analisis cluster terhadap data Iris, terlebih dahulu diberikan gambaran umum mengenai data tersebut berupa statistik deskriptif ke empat peubah yang diamati (lihat Tabel 2).

Tabel 2. Statistik deskriptif peubah-peubah data Iris

Data	Peubah	Rataan	Simpangan baku
IS	PS	5.006	0.353
	LS	3.428	0.379
	PP	1.462	0.174
	LP	0.246	0.105
IC	PS	5.936	0.516
	LS	2.770	0.314
	PP	4.260	0.470
	LP	1.326	0.198
IV	PS	6.577	0.636
	LS	2.974	0.323
	PP	5.552	0.552
	LP	2.026	0.275

Dari Tabel 2 tampak bahwa rataan dan simpangan baku peubah PP untuk spesies IS jauh lebih kecil dibandingkan dengan spesies IC dan IV, demikian juga

untuk peubah LP dan PS, walaupun perbedaannya tidak sebesar PP. Akan tetapi peubah LS untuk spesies IS, rataan dan simpangan bakunya sedikit lebih besar daripada spesies IC dan IV. Berdasarkan data deskriptif dan dari hasil plot dua komponen utama pertama (lihat Gambar 2) dapat digunakan sebagai petunjuk awal bahwa spesies IS terpisah dari kedua spesies lainnya. Ilustrasi data Iris ini dapat mewakili kondisi satu cluster terpisah dan dua cluster tumpang tindih. Hasil analisis cluster untuk masing-masing metode cluster untuk data Iris tertera pada Tabel 6.



Gambar 2. Plot dua komponen utama pertama pada data Iris

Tabel 3. Hasil pengelompokan data Iris menjadi 3 cluster dan persentase salah pengelompokannya.

Metode cluster	**IS (50,0,0)**	**IC (0,50,0)**	**IV (0,0,50)**	**Salah pengelompokan**
Berbasis model	(50,0,0)	(0,45,5)	(0,0,50)	5 (3.33%)

Ket. (50,0,0) : 50 masuk kelompok IS, 0 masuk kelompok IC dan 0 masuk kelompok IV

Dari Tabel 3 terlihat bahwa semua metode dapat memisahkan kelompok IS dari dua cluster lainnya. Untuk kelompok IC, metode cluster berbasis model memperoleh 5 amatan masuk dalam kelompok IV, yang seharusnya masuk dalam kelompok IC. Untuk kelompok IV, metode cluster berbasis model dapat memisahkan dengan tepat kelompok IV dari dua kelompok lainnya.

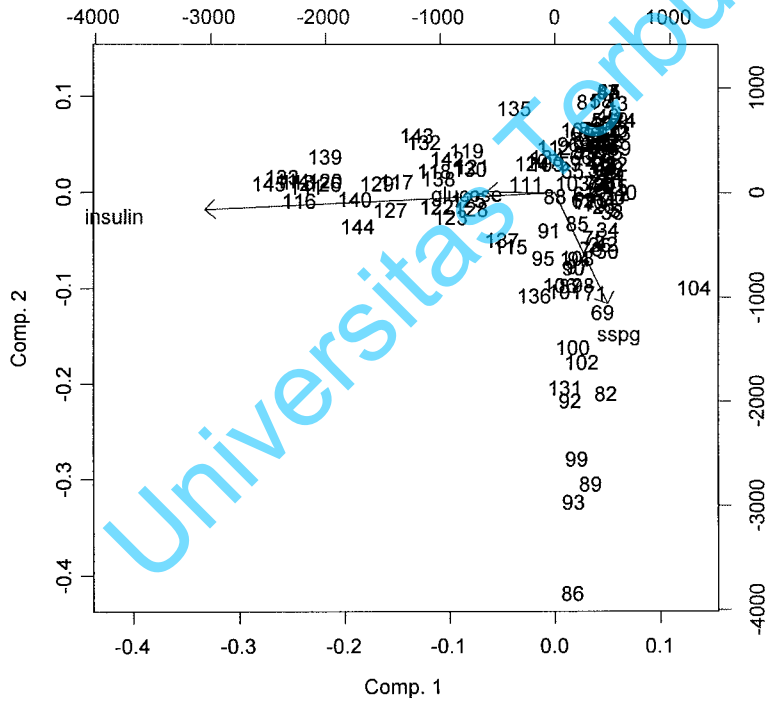
Persentasi salah pengelompokan yang terjadi dengan menggunakan metode cluster berbasis model ini adalah sebesar 3.33% (5 pengamatan). Salah pengelompokan terjadi hanya melibatkan spesies IC dan IV, sementara untuk spesies IS tidak terpengaruh untuk semua metode yang dicobakan. Hal ini disebabkan oleh cukup dekatnya jarak antar pusat cluster spesies IC dengan spesies IV (d= 1,62), sementara jarak antar pusat cluster spesies IS dengan spesies IC (d=3.21) dan jarak antar pusat cluster spesies IS dengan spesies IV (d=4.75) cukup jauh, sehingga menyebabkan spesies IS memang benar-benar terpisah dari dua spesies lainnya.

Data Diabetes

Data Diabetes adalah data diagnosa Diabetes yang dikembangkan oleh Reaven dan Miller, 1979 (data diperoleh dari perangkat lunak *MCLUST*). Data ini terdiri dari 3 jenis ukuran peubah yaitu: glucoce, Insulin dan SSPG (Steady-State Plasma Glucose) dengan masing-masing peubah terdiri dari 145 pengamatan. Secara klinikal data Diabetes ini telah diklasifikasikan atas tiga jenis, yaitu : Normal (NO), Chemical Diabetes (CD) dan Overt Diabetes (OD), yang masing-masing terdiri dari 76, 36, dan 33 pengamatan (lihat Lampiran 2). Tabel 3 memberikan gambaran umum mengenai data Diabetes, yang merupakan data deskriptif dari ketiga peubah yang diamati.

Tabel 4. Statistik deskriptif peubah-peubah data Diabetes

Data	Peubah	Rataan	Simpangan baku
NO	Glucose	91.184	8.228
	Insulin	349.974	36.871
	Sspg	172.645	68.854
CD	Glucose	99.306	9.489
	Insulin	482.556	93.018
	Sspg	288.000	157.832
OD	Glucose	217.667	76.563
	Insulin	1043.757	309.395
	Sspg	106.000	93.425



Gambar 3. Plot dua komponen utama pertama pada data Diabetes

Dari Tabel 4 menunjukkan bahwa rataan dan simpangan baku peubah glucose, insulin untuk jenis Diabetes OD jauh lebih besar dibandingkan dengan jenis Diabetes NO dan CD. Untuk peubah SSPG, rataan dan simpangan baku jenis Diabetes CD

lebih besar dari kedua jenis Diabetes lainnya. Dari hasil data deskriptif ini dan hasil dua komponen utama pertama (lihat Gambar 3) terlihat bahwa data Diabetes ini tidak jelas pengelompokannya. Illustrasi data Diabetes ini dapat mewakili kondisi ketiga cluster saling tumpang tindih (tidak jelas pengclusterannya).

Tabel 5. Hasil pengelompokan data Diabetes menjadi 3 cluster dan persentasi salah pengelompokannya.

Metode cluster	**NO (76,0,0)**	**CD (0,36,0)**	**OD (0,0,33)**	**Salah pengelompokan**
Berbasis model	(74,2,0)	(8,26,2)	(0,5,28)	17 (11.7%)

Ket. (76,0,0) : 76 masuk kelompok NO, 0 masuk kelompok CD dan 0 masuk kelompok OD

Dari Tabel 5, metode cluster berbasis model dapat mengelompokkan kelompok NO dengan tepat sebesar 74 amatan dan hanya ada 2 amatan salah mengelompokkan ke dalam kelompok CD, dan mengelompok CD dengan tepat sebesar 26 amatan, 8 amatan salah mengelompokkan ke dalam kelompok NO dan 2 amatan salah mengelompokkan ke dalam kelompok OD. Untuk Kelompok OD, metode cluster berbasis model terdapat 28 amatan dengan tepat pengelompokannya, semetara 5 amatan lainnya masuk dalam kelompok CD. Sedangkan untuk metode Ward hanya dapat mengelompokkan OD sebesar 14 amatan, 12 amatan masuk dalam kelompok CD dan 7 amatan masuk dalam kelompok NO. Metode K-rataan dapat mengelompokkan OD sebesar 17 amatan, 12 amatan masuk dalam kelompok CD dan 4 amatan masuk dalam kelompok NO.

Persentasi salah pengelompokan yang terjadi dengan menggunakan metode cluster berbasis model ini adalah sebesar 11.7% (17 pengamatan). Salah pengelompokan terjadi dengan melibatkan ketiga kelompok

KESIMPULAN DAN SARAN

Kesimpulan

Data Iris merupakan suatu kondisi satu kelompok terpisah dan 2 kelompok saling tumpang tindih. Persentasi salah pengelompokan yang terjadi dengan menggunakan metode cluster berbasis model adalah sebesar 3.33 % (amatan). Kesalahan pengelompokan ini terjadi pada 2 kelompok yang saling tumpang tindih, yaitu kelompok 5 amatan masuk kelompok IV yang seharusnya masuk kelompok IC.

Data Diabetes merupakan kondisi ketiga kelompok saling tumpang tindih. Persentasi salah pengelompokan sebesar 11.7% (17 pengamatan). Kesalahan pengelompokan ini terjadi pada ketiga kelompok, 2 amatan dari kelompok NO masuk kedalam kelompok CD, 8 amatan dari kelompok CD masuk NO dan 2 amatan masuk kelompok OD. Sedangkan kelompok OD terdapat 5 amatan masuk ke dalam kelompok CD.

Hasil pengelompokan dari 2 contoh data penerapan di atas (data Iris dan data Diabetes) menunjukkan bahwa metode gerombol berbasis model cukup efektif memisahkan kelompok-kelompok yang saling tumpang tindih.

Saran

Perlu adanya penelitian lebih lanjut untuk membandingkan hasil pengelompokan menggunakan metode cluster berbasis model dengan metode cluster berbasis jarak.

DAFTAR PUSTAKA

1. Anderberg, M.R. (1973) *Cluster analysis for applications*, New York: Academic Press.
2. Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997) Inference in *model-based* cluster analysis, *Statistics and Computing*, 7, 1-10.
3. Fraley, C. (1999) Algorithms for *model-based* Gaussian hierarchical clustering. *SIAM Journal Science Computations*, 20, 270-281.
4. Fraley, C. and Raftery, A. E. (1998) How many clusters ? Which clustering method ? Answers via *model-based* cluster analysis. *The Computer Journal* 41, 578-588.
5. Fraley, C. and Raftery, A. E. (1999) MCLUST:Software for *model-based* clustering analysis. *Journal of Classifications*. 16, 297-306.
6. Fraley, C. and Raftery, A. E. (2000) *Model-based* clustering, discriminant analysis, and density estimation. Technical Report, Department of Statistics University of Washington.
7. Hartigan, J. A. (1975) *Clustering Algorithm*, New York : Wiley.
8. Johnson, R. A. and Wichern, D. W. (1998) *Applied Multivariate Statistical Analysis*, 4 th Edition, New Jersey: Prentice-Hall.
9. Kass, R. E. and Raftery, A. E. (1995) Bayes Factor. *Journal of the American Statistical Association*, 90, 773-795.
10. Mclachlan, G.J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering,* New York : Marcel Dekker.