PERBANDINGAN METODE *MODEL-BASED* DENGAN METODE K-*MEAN* DALAM ANALISIS *CLUSTER*

Timbul Pardede (timbul@mail.ut.ac.id)
Universitas Terbuka

ABSTRACT

K-mean method is a clustering method in which grouping techniques are based only on distance measure among observed objects, without considering statistical aspects. Model-based clustering is a method that use statistical aspects, as its theoretical basis i.e. probability maximum criterion. This model has several variations with a variety of geometrical characteristics obtained by mean Gauss component. Data partition is conducted by utilizing EM (expectation-maximization) algorithm. Then by using Bayesian Information Criterion (BIC) the best model is obtained. This research aimed to comparing result of grouping methods between model-based clustering and K-mean clustering. The results showed that model-based clustering was more effective in separating overlap groups than K-mean.

Keywords: BIC, EM algorithm, K-mean clustering method, model-based clustering method.

Analisis *cluster* merupakan suatu metode pengelompokan satuan objek pengamatan menjadi beberapa kelompok objek pengamatan berdasarkan peubah-peubah yang dimiliki. Analisis ini menjadikan 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 Ward dan juga metode *cluster* tak berhierarki, misalnya metode K-*mean* (Anderberg 1973). Metode *cluster* ini memiliki teknik yang berbeda dalam proses pembentukan kelompok, namun teknik tersebut hanya memperhatikan ukuran jarak antar objek pengamatan. Metode-metode ini tidak mempertimbangkan aspek statistiknya, seperti sebaran datanya.

Metode *model-based* adalah suatu metode yang berbeda dengan metode *cluster* yang didasarkan pada ukuran jarak. Metode ini merupakan suatu algoritma *cluster* yang tergolong baru. Analisis ini dilakukan berdasarkan pada aspek statistik yaitu menggunakan kriteria kemungkinan maksimum dalam memutuskan hasil kelompoknya. Metode ini mempunyai beberapa model dengan berbagai macam sifat geometris yang diperoleh melalui komponen Gauss.

Dari metode-metode *cluster* yang telah diungkapkan di atas, tidak semua metode dapat digunakan untuk menganalisis data, khususnya apabila data pengamatannya cukup besar atau data pengamatannya merupakan data *cluster* yang saling tumpang tindih. Sering muncul pertanyaan mengenai *"Berapa banyak cluster yang ada?", "Metode cluster apa yang digunakan?"*. Pengguna statistika seringkali melakukan dengan coba-coba (*trial and error*) untuk mendapatkan hasil yang bermakna atau yang dapat diinterpretasikan sesuai dengan masalah kajiannya (Siswadi & Suharjo,

1999). Hal ini sudah barang tentu akan menimbulkan subjektivitas pengguna statisitika dalam memutuskan hasil *cluster* tersebut. Oleh karena itu upaya membandingkan hasil pengelompokan metode *model-based* dengan metode *cluster* yang didasarkan pada ukuran jarak (metode K-*mean*) menjadi suatu hal yang perlu dan menarik untuk dikaji.

Berdasarkan permasalahan tersebut, tujuan penelitian ini adalah untuk membandingkan hasil pengelompokan metode *model-based* dengan metode K-*mean* dalam analisis *cluster*. Penelitian ini diharapkan dapat memberi rekomendasi yang tepat tentang metode *cluster* yang digunakan dalam pengelompokan data yang cenderung terdapat *cluster* yang saling tumpang tindih.

Metode *Cluster* K-*Mean*

Metode K-*mean* merupakan suatu metode *cluster* tak berhierarki, yaitu metode *cluster* yang menyekat objek pengamatan ke dalam k *cluster*. Metode ini pada umumnya diaplikasikan pada gugus data yang berukuran relatif besar.

Macqueen *dalam* Johnson dan Wichern (1998) menggambarkan algoritma *cluster* untuk menyeleksi n unit data ke dalam k *cluster* adalah berdasarkan kedekatan pusat (rataan) yang disusun dengan tahapan berikut ini.

1. Mengambil k unit data pertama yang digunakan sebagai k pusat *cluster* awal.
2. Menggabungkan setiap (n-k) data yang merupakan sisa anggota ke pusat *cluster* terdekat. Kemudian dihitung masing-masing pusat (rataan) *cluster* baru yang terbentuk dari hasil penggabungan.
3. Setelah semua data digabungkan pada tahap 2, pusat *cluster* yang terbentuk dijadikan sebuah titik pusat (rataan) *cluster*. Kemudian dilakukan penggabungan kembali dari setiap unit data ke dalam titik pusat terdekat.

Suatu *cluster* yang konvergen diperoleh dengan memperbaiki secara berulang titik pusat *cluster* yang terbentuk pada tahap ke-3 melalui penggabungan semua n data ke titik pusat terdekat. *Cluster* yang konvergen ditandai dengan adanya titik pusat yang tetap dan tidak ada lagi perubahan anggota di antara *cluster*.

Ukuran ketakmiripan antar objek pengamatan yang digunakan dalam analisis *cluster* adalah jarak antar objek. Jarak antar dua objek harus didefinisikan sedemikian rupa sehingga semakin pendek jarak antar dua objek, semakin kecil ketakmiripannya. Nilai ukuran ketakmiripan yang sering digunakan adalah jarak Euclid dan jarak Mahalanobis. Jarak Mahalanobis digunakan bila semua peubah saling berkorelasi atau tidak saling ortogonal, sebaliknya jarak Euclid digunakan bila antar peubah saling bebas atau saling ortogonal (Johnson & Wichern, 1998). Jarak Euclid antara objek ke-i dan objek ke-j dengan p peubah didefinisikan dengan $d_{ij} = \left\{ \sum_{k=1}^{p} \left(X_{ik} - X_{jk} \right)^2 \right\}^{1/2}$ dan jarak Mahalanobis didefinisikan dengan $d_{ij} = \left\{ \left(\mathbf{x_i} - \mathbf{x_j} \right)' S^{-1} \left(\mathbf{x_i} - \mathbf{x_j} \right) \right\}^{1/2}$ dengan S adalah matriks kovariansi.

Metode *Cluster Model-Based*

Dalam metode *model-based* (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 variabel acak multivariat (p-variat) dari suatu populasi, dengan 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 $\mathrm{n_j}$ data dengan $\sum_{j=1}^{g} \mathrm{n_j} = \mathrm{n}$.

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

$$\mathrm{f}\left(\mathbf{x}|\boldsymbol{\phi}\right) = \sum_{\mathbf{j}=1}^{\mathbf{g}} \pi_{\mathbf{j}}\, \mathrm{f_j}\left(\mathbf{x}|\boldsymbol{\theta}\right) \quad ; \quad \boldsymbol{\phi} \in \Omega \tag{1}$$

dengan $\mathrm{f_j}\left(\mathbf{x}|\boldsymbol{\theta}\right)$ merupakan fungsi kepekatan G_j, yaitu subpopulasi ke-j dengan vektor parameter $\boldsymbol{\theta}$ yang tidak diketahui dan $\pi_{\mathbf{j}}$ merupakan proporsi data yang berasal dari subpopulasi ke-j dengan $\sum_{j=1}^{g} \pi_{\mathrm{j}} = 1$ dan $\pi_{\mathrm{j}} > 0 \quad \left(\mathrm{j} = 1,2,.....,\mathrm{g}\right)$, sedangkan $\boldsymbol{\phi} = \left(\boldsymbol{\pi}', \boldsymbol{\theta}'\right)'$ adalah gugus semua parameter dari fungsi kepekatan campuran yang berasal dari ruang parameter Ω (Mclachlan & Basford, 1988). Dengan asumsi $\mathbf{X_1, X_2, ...X_n}$ bebas stokastik dan identik dengan fungsi kepekatan $\mathrm{f_j}\left(\mathbf{x_i}|\theta_{\mathrm{j}}\right)$ merupakan fungsi kepekatan pengamatan $\mathbf{x_i}$ dari *cluster* ke-j maka Fungsi kemungkinan sebaran campuran (*mixture likelihood*) pada persamaan (1) adalah:

$$\mathrm{L}(\boldsymbol{\phi}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \left[\sum_{\mathrm{j}=1}^{\mathrm{g}} \pi_{\mathrm{j}} \mathrm{f_j}\left(\mathbf{x_i}|\theta_{\mathrm{j}}\right) \right] \tag{2}$$

Dalam penelitian ini difokuskan pada suatu kasus dengan $\mathrm{f_j}\left(\mathbf{x_i}|\boldsymbol{\theta}_{\mathrm{j}}\right)$ adalah fungsi kepekatan peubah ganda campuran normal (*Gauss*) dengan parameter $\boldsymbol{\theta}_{\mathrm{j}}$ terdiri dari vektor rataan $\boldsymbol{\mu}_{\mathrm{j}}$ dan matriks koragam $\sum_j$, yang dinyatakan dalam bentuk:

$$\mathrm{f_j}(\mathrm{x_i}|\boldsymbol{\mu_j}; \sum_{\mathbf{j}}) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{x_i} - \boldsymbol{\mu_j}\right)' \sum_{\mathbf{j}}^{-1} \left(\mathbf{x_i} - \boldsymbol{\mu_j}\right)\right\}}{\left(2\pi\right)^{\frac{\mathbf{p}}{2}} \left|\sum_{\mathbf{j}}\right|^{\frac{1}{2}}} \tag{3}$$

Dalam Metode *model-based*, diasumsikan bahwa data dibangkitkan dengan fungsi kepekatan peubah ganda campuran. Data bangkitan tersebut dicirikan oleh *cluster ellipsoidal* yang terpusat pada rataan $\boldsymbol{\mu}_{\mathrm{j}}$ (Fraley & Raftery, 1999). Karakteristik geometrik (bentuk *(shape),* volume, dan orientasi *(orientation)*) *cluster* dihitung dengan matriks koragam $\sum_j$ yang diparameterisasikan untuk menentukan batasan antar *cluster*.

Banfield & Raftery (1993) mengembangkan metode *model-based* dengan memparameterisasikan setiap matriks koragam melalui suku dekomposisi nilai ciri dalam bentuk :

$$\sum_j = \lambda_j D_j A_j D_j' \tag{4}$$

dengan :

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 $\sum_j$, yang menjelaskan bentuk,

λ_j adalah skalar yang menjelaskan volume.

Sebagai ilustrasi, 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 dengan setiap *cluster* mempunyai ciri geometrik yang berbeda. Tabel 1 menunjukkan matriks koragam $\sum_j$ untuk model campuran normal ganda dan interpretasi geometrik (Fraley & Raftery, 1999).

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- maximization*) merupakan metode perhitungan iteratif 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 :

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

Algoritma EM ini terdiri dari dua tahap yaitu tahap E untuk pendugaan dan tahap M untuk pemaksimuman. Dengan asumsi **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

$L(\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{z}\ |\mathbf{x}) = \prod_{i=1}^{n}\prod_{j=1}^{g}\left\{\pi_j f_j\left(\mathbf{x_i}|\theta_j\right)\right\}^{z_{ij}}$ atau fungsi log kemungkinan data lengkap (*complete-data loglikelihood*) adalah $\mathbf{L}(\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{z}\ |\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}\left\{\ln\pi_j + \ln f_j\left(\mathbf{x_i}|\theta_j\right)\right\}$.

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

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

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

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

$\hat{z}_{ij}$ merupakan dugaan peluang akhir $\mathbf{x}_i$ dalam kelompok 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).

Pemilihan Model *Cluster* 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 (*Bayes Factor*). 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, M_1 dan 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:

$$P\left(\boldsymbol{X}\left|M_k\right.\right) = \int P\left(\boldsymbol{X}\left|\theta_k,M_k\right.\right) P\left(\theta_k\left|M_k\right.\right) d\theta_k \qquad k= 1,2$$

dengan $P\left(\theta_k\left|M_k\right.\right)$ adalah sebaran awal θ_k, dengan θ_k adalah parameter model M_k.

Faktor Bayes didefinisikan sebagai rasio dari integral kemungkinan dari kedua model, yakni $B_{12} = \frac{P(\mathbf{X}|M_1)}{P(\mathbf{X}|M_2)}$. Metode ini dikembangan secara umum untuk lebih dari dua model (Kass & Raftery, 1995).

Kass & Raftery (1995) mengemukakan bahwa integral kemungkinannya dapat didekati dengan pendekatan faktor Bayes melalui algoritma EM (Ekspektasi-Maksimum). Pendekatan ini disebut BIC (*Bayesian Information Criterion*) dengan formulasi sebagai berikut.

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

dengan

$P\left(\boldsymbol{X} \middle| M_k\right)$: adalah integral kemungkinan untuk model M_k

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

V_k : adalah banyaknya parameter bebas yang diduga pada model M_k

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

Nilai BIC terbesar mengindikasikan bahwa model tersebut adalah model yang paling layak.

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

1. Tentukan banyak *cluster* maksimum (m) dan himpunan model campuran ganda normal.
2. Lakukan *cluster* berhierarki penggabungan untuk setiap model campuran normal ganda. Hasil *cluster* ini ditransformasi ke dalam peubah indikator pada persamaan (5), yang kemudian digunakan sebagai nilai awal untuk algoritma EM.
3. Lakukan algoritma EM untuk setiap model dan masing-masing banyak *cluster* 2, 3, …., m, yang diawali dengan klassifikasi *cluster* berhierarki.
4. 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*.
5. Plotkan nilai BIC untuk setiap model. Nilai BIC terbesar mengindikasikan bahwa model tersebut adalah model yang paling layak.

METODOLOGI

Data yang digunakan dalam penelitian ini adalah data hasil simulasi, yang dibangkitkan dengan menggunakan perangkat lunak program *Minitab 11.12*. Data yang dibangkitkan terdiri dari 3 *cluster* dengan 3 peubah dan jumlah pengamatan tiap *cluster* yang dicobakan sebesar 50, 100 dan 150. Ketiga *cluster* yang akan dibangkitkan dibuat dalam 3 macam kondisi, (1) ketiga *cluster* saling terpisah, (2) satu *cluster* terpisah dan dua *cluster* tumpang tindih, dan (3) ketiga *cluster* saling tumpang tindih. Untuk membangkitkan banyaknya pengamatan yang tumpang tindih, maka dicobakan 3 jenis ukuran jarak antara dua nilai tengah (pusat) *cluster*, yang disesuaikan dengan jauh dekatnya jarak antara vektor rataan *cluster*. Untuk melihat pengaruh tingkat korelasi antara peubah terhadap hasil akhir *cluster*, dicobakan 3 tingkat korelasi, yaitu tingkat rendah (0.20), tingkat korelasi

sedang (0.50) dan tingkat korelasi tinggi (0.80). Selain itu sebagai contoh penerapan untuk mendukung hasil penelitian ini, diambil data sekunder, yaitu data Iris yang banyak digunakan dalam buku teks statistika multivariat dan dalam paket program statistika seperti S-plus dan Minitab.

Prosedur analisis data dilakukan sebagai berikut.

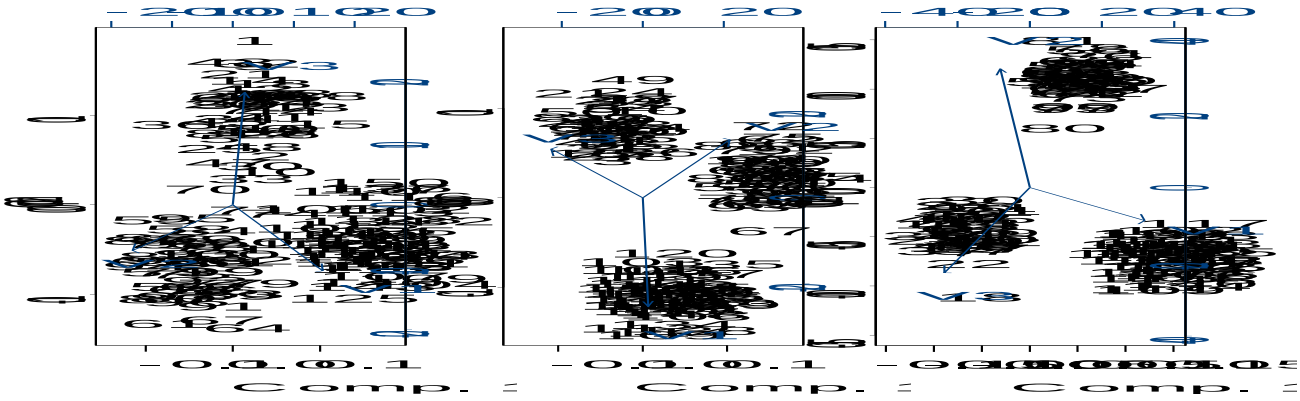
1. Data hasil bangkitan, selanjutnya dilakukan analisis *cluster* dengan menggunakan paket program *Splus 2000* untuk metode K-*mean*, sedangkan untuk metode *model-based* digunakan paket program *Mclust* dengan *interface Splus 2000*.
2. Bandingkan hasil *cluster* masing-masing metode dengan *cluster* yang sebenarnya (ditentukan saat simulasi).
3. Hitung persentase salah pengelompokan dari masing-masing metode, kemudian hasilnya dibandingkan.
4. Persentase salah pengelompokan yang terkecil menunjukkan bahwa metode yang digunakan lebih baik.

HASIL DAN PEMBAHASAN

Data yang dibangkitkan terdiri dari 81 kasus data simulasi dengan masing-masing kasus terdiri dari 3 *cluster*. Semua kasus data dibedakan atas kondisi pengelompokan, jarak antar pusat *cluster*, tingkat korelasi dan banyak data. Setiap kasus data digunakan sebagai data awal bagi masing-masing metode K-*mean* dan metode *model-based.*

Ketiga *Cluster* Saling Terpisah

Persentase salah pengelompokan untuk kondisi ketiga *cluster* saling terpisah diperoleh hasil yang sama besar pada kedua metode yang dibandingkan, yaitu 0%. Ini menunjukkan bahwa pengelompokan setiap metode sesuai dan tepat dengan pengelompokan yang sebenarnya (ditentukan saat simulasi). Hal ini disebabkan oleh ukuran jarak antar vektor rataan *cluster* yang relatif jauh dan variansi setiap peubah cenderung kecil sehingga objek pengamatan mengelompok di sekitar vektor rataannya (lihat Gambar 1). Persentase salah pengelompokan ini tidak terpengaruh terhadap banyak objek pengamatan tiap *cluster*, ketiga jarak antar pusat *cluster* dan juga tingkat korelasi antar peubah.

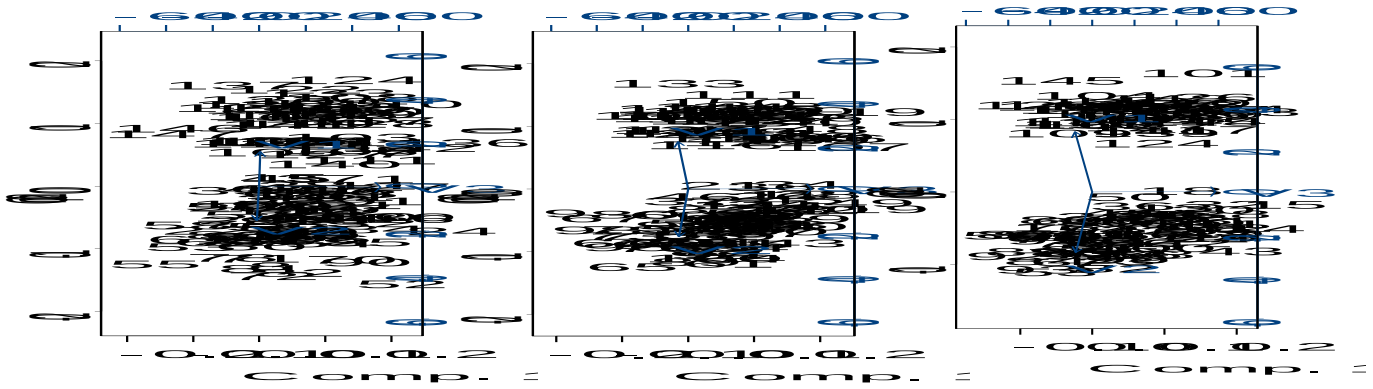


d dekat, tingkat korelasi rendah dan n=50 | b. d sedang, tingkat korelasi rendah dan n=50 | c. d jauh, tingkat korelasi rendah dan n=50

Gambar 1. Pola simulasi data untuk kondisi ketiga *cluster* saling terpisah dengan banyak objek pengamatan untuk tiap *cluster* sebesar n=50

Satu *Cluster* Terpisah dan Dua *Cluster* Tumpang Tindih

Pada Gambar 2 disajikan hasil plot dua komponen utama pertama dari kondisi pengelompokan satu *cluster* terpisah dan dua *cluster* tumpang tindih, tingkat korelasi rendah dengan banyak objek pengamatan tiap *cluster* sebesar 50 pengamatan.



a. d dekat, tingkat korelasi rendah dan n=50

b. d sedang, tingkat korelasi rendah dan n=50

c. d jauh, tingkat korelasi rendah dan n=50

Gambar 2. Pola simulasi data untuk kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih dengan banyak objek pengamatan tiap *cluster* sebesar n=50

Persentase salah pengelompokan untuk kondisi ini diperoleh hasil seperti tercantum pada Tabel 2. Metode *model-based* memperoleh persentase salah pengelompokan terkecil dan bahkan jauh lebih kecil persentase pengelompokannya dibandingkan metode K-*mean*. Ditinjau dari jarak antar pusat *cluster*, terjadi penurunan salah pengelompokan dengan semakin jauh jarak antar pusat *cluster* untuk kedua metode *cluster*. Ditinjau dari tingkat korelasi antar peubah menunjukkan bahwa metode *model-based* terjadi penurunan persentase salah pengelompokan dari tingkat korelasi rendah ke tingkat korelasi tinggi, walaupun penurunan ini hampir tidak ada perbedaan yang berarti. Untuk banyak objek pengamatan tiap *cluster* sebesar 50 mempunyai pola persentase salah pengelompokan yang tidak jauh berbeda dengan objek pengamatan tiap *cluster* sebesar 100 dan 150.

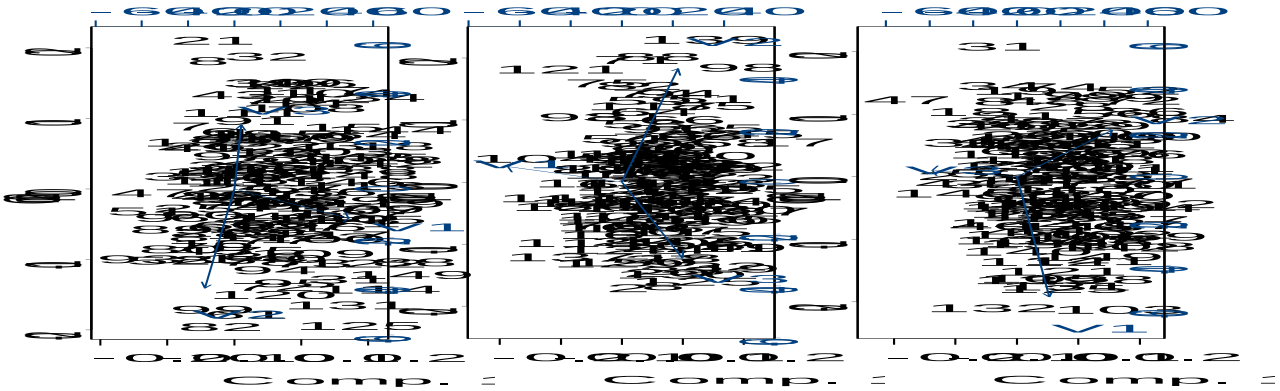
Tabel 2. Persentase Salah Pengelompokan pada Kondisi Satu *Cluster* Terpisah dan Dua *Cluster* Saling Tumpang Tindih.

Tingkat Korelasi	Metode	n=50			n=100			n=150		
		Jarak			Jarak			Jarak		
		Dekat	Sedang	Jauh	Dekat	Sedang	Jauh	Dekat	Sedang	Jauh
Rendah	K-*mean*	54.67	49.33	12.67	56.00	50.33	15.33	54.44	49.78	13.56
	Model-based	2.67	2.00	0.00	4.67	0.00	0.00	2.89	0.89	0.00
Sedang	K-*mean*	58.67	55.33	14.00	51.67	40.00	15.67	54.89	55.33	12.00
	Model-based	2.00	0.00	0.00	1.67	0.00	0.00	0.89	0.00	0.00
Tinggi	K-*mean*	55.33	52.67	19.33	59.33	53.33	15.00	58.00	46.44	12.44
	Model-based	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Hasil pengelompokan kedua metode *cluster* yang dibandingkan, menunjukkan bahwa metode *model-based* lebih efektif dalam memisahkan kelompok-kelompok pada kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih dibandingkan metode K-*mean*.

Ketiga *Cluster* Saling Tumpang Tindih

Persentase salah pengelompokan untuk kondisi ketiga *cluster* saling tumpang tindih disajikan pada Tabel 3. Metode *model-based* pada tingkat korelasi rendah dan tingkat korelasi sedang dengan ketiga jenis jarak yang dicobakan tidak mampu memisahkan kelompok-kelompok yang saling tumpang tindih. Hal ini mungkin disebabkan oleh objek-objek pengamatannya mengelompok pada satu *cluster* (lihat Gambar 3). Sehingga secara geometris dari 6 model dari metode *model-based* tidak mampu memisahkan *cluster* yang saling tumpang tindih. Bahkan metode *model-based* ini menganjurkan bahwa akan lebih efektif jika pengelompokannya dibagi dalam 2 atau 4 *cluster*.



a. d dekat, tingkat korelasi rendah dan n=50 b. d sedang, tingkat korelasi rendah dan n=50 c. d jauh, tingkat korelasi rendah dan n=50

Gambar 3. Pola simulasi data untuk kondisi ketiga *cluster* saling tumpang tindih dengan banyak objek pengamatan untuk tiap *cluster* sebesar n=50

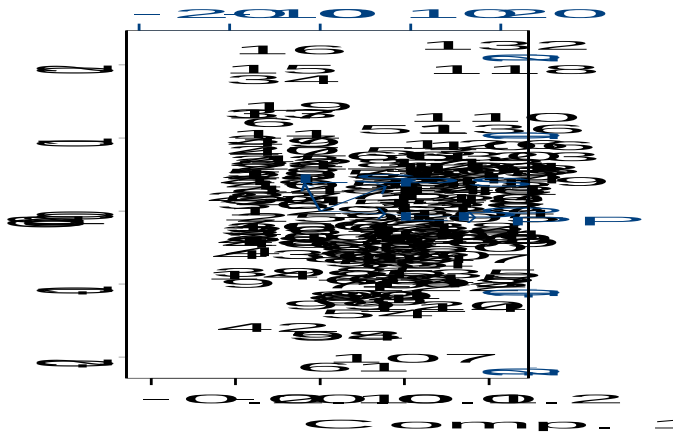
Hasil pengelompokan kedua metode yang dibandingkan menunjukkan bahwa metode *model-based* lebih efektif memisahkan *cluster* yang saling tumpang tindih apabila tingkat korelasi tinggi dan jarak antar pusat *cluster* relatif sedang dan jauh. Sebaliknya, apabila tingkat korelasi tinggi dengan jarak antar pusat *cluster* relatif dekat dan juga pada tingkat korelasi rendah dan sedang dengan jarak antar pusat *cluster* dekat, sedang dan jauh, kedua metode yang dibandingkan tidak efektif dalam memisahkan *cluster* yang tumpang tindih.

Tabel 3. Persentase salah pengelompokan pada kondisi ketiga *cluster* saling tumpang tindih

Tingkat Korelasi	Metode	n=50			n=100			n=150		
		Jarak			Jarak			Jarak		
		Dekat	Sedang	Jauh	Dekat	Sedang	Jauh	Dekat	Sedang	Jauh
Rendah	K-*mean*	49.33	46.67	23.33	56.67	47.00	25.33	53.78	40.44	27.56
	Model-based	52.67	53.33	24.67	64.00	55.00	31.00	58.67	52.89	22.67
Sedang	K-*mean*	58.00	50.00	44.00	62.33	52.33	41.33	60.67	51.78	48.44
	Model-based	64.67	60.00	43.33	62.67	60.33	16.00	64.89	25.11	14.00
Tinggi	K-*mean*	64.00	62.00	52.00	64.67	63.00	60.67	65.11	60.22	55.56
	Model-based	45.33	11.33	2.67	26.67	10.33	4.00	39.56	8.89	2.44

Data Iris

Data Iris merupakan contoh klasik yang sering digunakan dalam buku-buku teks statistik untuk mengilustrasikan masalah pengelompokan data. 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.



Gambar 4. Plot dua komponen utama pertama pada data Iris

Berdasarkan hasil plot dua komponen utama pertama (lihat Gambar 4) 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.

Dari Tabel 4 terlihat bahwa untuk kelompok IC, metode *model-based* memperoleh 5 amatan masuk dalam kelompok IV, yang seharusnya masuk dalam kelompok IC, sementara untuk metode K-*mean* memperoleh salah pengelompokan sebesar 2 amatan masuk dalam kelompok IV, yang seharusnya masuk dalam kelompok IC. Untuk kelompok IV, metode *model-based* dapat memisahkan dengan tepat kelompok IV dari dua kelompok lainnya, sedangkan untuk metode K-*mean* memperoleh salah pengelompokan yang cukup besar, yaitu sebesar 14 amatan masuk dalam kelompok IC.

Tabel 4. 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
K-*mean*	(50,0,0)	(0,48,2)	(0,14,36)	16 (10.67%)
Model-based	(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

Salah pengelompokan terkecil terjadi pada metode pengelompokan *model-based* sebesar 3.33% (5 pengamatan), sementara persentase salah pengelompokan metode K-*mean* sebesar 10.67% (16 pengamatan).

Salah pengelompokan yang terjadi disini hanya melibatkan spesies IC dan IV, sementara untuk spesies IS tidak terpengaruh untuk kedua metode yang dicobakan. Hal ini disebabkan oleh cukup dekatnya jarak antar pusat *cluster* spesies IC dengan spesies IV, sehingga menyebabkan spesies IS memang benar-benar terpisah dari dua spesies lainnya.

KESIMPULAN

Hasil penelitian ini menunjukkan bahwa secara umum metode *model-based* lebih efektif memisahkan *cluster* yang saling tumpang tindih dibandingkan dengan metode K-*mean*. Dalam hal untuk data lapangan yang cenderung terdapat *cluster* yang saling tumpang tindih disarankan menggunakan metode *model-based.*

REFERENSI

Anderberg, M.R. (1973). *Cluster analysis for applications*, New York: Academic Press.

Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49, 803-821.*

Fraley, C. & Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal,* 41, 578-588.

Fraley, C. & Raftery, A.E. (1999). MCLUST: Software for model-based clustering analysis. *Journal of Classifications,* 16, 297-306.

Johnson, R.A. & Wichern, D.W. (1998). *Applied multivariate statistical analysis*, 4 th Edition. New Jersey: Prentice-Hall.

Kass, R.E. & Raftery, A.E. (1995). Bayes Factor. *Journal of the American Statistical Association*, 90, 773-795.

Mclachlan, G.J. & Basford, K.E. (1988). *Mixture models: Inference and applications to clustering,* New York: Marcel Dekker.

Siswadi & Suharjo, B. (1999). *Analisis eksplorasi data peubah ganda.* Bogor: Jurusan Matematika FMIPA IPB. Bogor.