BAB II

ANALISA CLUSTER

Dalam analisa clustering, setiap elemen di dalam himpunan N disajikan dalam vektor, dengan elemen-elemen berupa karakteristik untuk atribute yang digunakan sebagai informasi awal untuk pengelompokan (Grouping). Diasumsikan bahwa $x_i = (x_{1i}, x_{2i},\ldots,x_{zi}) \in R^z$; $i \in N$. Didefinisikan jarak antara dua elemen dari N, seperti pada definisi berikut :

2.1. Mencari Jarak Antara Dua Elemen

Definisi :

Suatu fungsi riil d_{ij} , i dan j $\in$ N adalah metrik atau ukuran jarak, jika d_{ij} memenuhi kondisi berikut :

1. $d_{ij} \geq 0$, $d_{ij} = 0$ jika $i = j$
2. $d_{ij} = d_{ji}$
3. $d_{ik} + d_{kj} \geq d_{ij}$; $i, j, k \in N$

Dengan menganggap beberapa jarak yang digunakan , d_{ij} dinyatakan dalam bentuk :

1. $d_{ij} = \left[\sum_{k=1}^{z} \left| X_{ik} - X_{jk} \right|^r \right]^{1/r}$; r integer positip dalam fungsi i dan j disebut metrik minkowski. Jika r = 1 disebut matriks absolut, jika r = 2 disebut metrik euclid.

2. $d_{ij} = \left[\sum_{k=1}^{z} W_k \left|X_{ik} - X_{jk}\right|^r\right]^{1/r}$, r integer positip disebut jarak terboboti minkowski (*weighted Minkowski distance*). Jika r = 1,2 didapat hubungan antara berat absolut dan metrik euclid.

3. $d_{ij} = \left(X_i - X_j\right)' P \left(X_i - X_j\right)$ dimana P adalah matriks simetrik semi definite positip z x z disebut metrik euclid umum (*general Euclidean metric*) dan diberikan berat suatu pasangan karakteristik $d_{ij} = \left(X_i - X_j\right)' \sum_z^{-1} \left(X_i - X_j\right)$ dimana Σ_z adalah matriks varian-covarian yang menghubungkan karakteristik z.

Metrik ini lebih umum dari metrik euclid berat pada $\Sigma_z = I_z$. Jarak ini disebut Mahalanobis D^2 . Dengan d_{ij} adalah jarak antara i dan j untuk i,j $\in$N. Metrik dari d_{ij} disebut matriks disimilaritas.

Contoh aplikasi matriks Mahalanobis :

Jika diberikan n obyek dan p variabel, maka diberikan jajaran matriks $x = (x_{i1}, x_{i2}, \ldots, x_{in}) \in R^p$ dimana i $\in$ N matriks X diberikan seperti di bawah :

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{nxp}$$

Definisi mean, varian - kovarian dan product momen korelasi

- Mean dari variabel x dan y

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \quad \text{dan} \quad \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

dari mean di atas diperoleh

$$\hat{x} = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} \quad \text{dan} \quad \hat{x}^T = [(x_1 - \bar{x}) \cdots (x_n - \bar{x})]$$

$$\hat{y} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \quad \text{dan} \quad \hat{y}^T = [(y_1 - \bar{y}) \cdots (y_n - \bar{y})]$$

- varian dan covarian

$$\text{cov}(x,y) = \hat{x}^T \hat{y}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y})$$

$$\text{var}(x) = \hat{x}^T \hat{x}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- product momen korelasi

$$r = (x,y)$$

$$= \frac{\text{cov}(x,y)}{\{\text{var}(x)\,\text{var}(y)\}}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left(\left[\Sigma (x_i - \bar{x})^2 \left[(y_i - \bar{y})^2\right]\right]\right)^{\frac{1}{2}}}$$

product momen korelasi dihitung sebagai ukuran gabungan antara nilai vektor dari dua data unit. Cronbach and Gleser (1953) menunjukkan koefisien korelasi mempunyai karakter limit matriks khususnya jika x_{ij} adalah nilai untuk j data unit pada i variabel, maka didefinisikan mean dan scater untuk j data unit sebagai

$$\bar{x}_j = \sum_{i=1}^{n} \frac{x_{ij}}{n} \qquad s_j = \left[\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)^2 \right]^{\frac{1}{2}}$$

dari statistik ini, untuk mentransformasikan nilai dari setiap data unit pada rata-rata nol dan unit scatter diperoleh jarak euclid antara unit data sebagai ukuran ditransformasikan nilai $\hat{x}_{ij} = \dfrac{\left(x_{ij} - \bar{x}_j \right)}{s_j}$ adalah

$$d^2\left(x_j, x_k\right) = \sum_{i=1}^{n} \left[\frac{x_{ij} - \bar{x}_j}{s_j} - \frac{x_{ik} - \bar{x}_k}{s_k} \right]^2$$

$$= \frac{1}{s_j^2 s_k^2} \left[s_j^2 s_k^2 - 2 s_j s_k \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right) \left(x_{ik} - \bar{x}_k \right) + s_j^2 s_k^2 \right]$$

$$= 2 \left[1 - \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right) \left(x_{ik} - \bar{x}_k \right)}{s_j s_k} \right]$$

$$= 2 \left(1 - r_{jk} \right)$$

dari jarak euclid di atas diperoleh matriks jarak kuadrat

$$d_{ij}^2 = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix}$$ disebut matriks dissimilaritas

Diberikan suatu matriks data dengan n baris (himpunan subyek dari 8 jenis mesin textile) dan p kolom (himpunan variabel dari 6 minggu)

	A	B	C	D	E	F
1	12	20	5	12	7	12
2	3	10	3	4	6	6
3	4	3	3	3	4	7
4	4	5	4	3	12	4
5	7	2	4	4	4	7
6	6	3	2	2	3	2
7	2	2	2	2	2	6
8	2	3	1	2	2	4
Σ	40	48	24	32	40	48
$\overline{X}$	5	6	3	4	5	6

Mencari jarak antara dua elemen d_{ij}^2 :

$r_{ij} = \dfrac{s_{ij}}{s_{ii}s_{jj}}$ Produk Momen Korelasi

$$r_{11} = \frac{S_{11}^2}{S_{11}S_{11}} = 1 \qquad r_{11} = r_{22} = r_{33} = r_{44} = r_{55} = r_{66} = 1$$

$$r_{12} = \frac{S_{12}}{S_{11}S_{22}} = \frac{13}{\sqrt{9,75}\sqrt{34}} = 0,71$$

$r_{13} = 0,75$ $r_{23} = 0,63$ $r_{34} = 0,75$ $r_{46} = 0,88$

$r_{14} = 0,87$ $r_{24} = 2,38$ $r_{35} = 0,29$ $r_{56} = 0,15$

$r_{15} = 0,27$ $r_{25} = 1,03$ $r_{36} = 0,1$

$r_{16} = 0,67$ $r_{26} = 0,75$ $r_{45} = 0,36$

dari varian, covarian dan produk momen korelasi,

$$d_{11} = d_{22} = d_{33} = d_{44} = d_{55} = d_{66} = 0$$

$$d_{12} = 2\,(1-r_{12}) = 0,58$$

$$d_{13} = 0,5$$

$$d_{ij}^2 = \begin{bmatrix} 0 & 0,58 & 0,5 & 0,26 & 1,46 & 0,66 \\ 0,58 & 0 & 0,74 & -2,76 & -0,06 & 0,5 \\ 0,5 & 0,74 & 0 & 0,5 & 1,42 & 1,8 \\ 0,26 & -2,76 & 0,5 & 0 & 2,28 & 0,26 \\ 1,46 & -0,06 & 1,42 & 1,28 & 0 & 1,7 \\ 0,66 & 0,5 & 1,8 & 0,26 & 1,7 & 0 \end{bmatrix}$$

d_{ij}^2 = menggambarkan hubungan fungsional antara variabel ke-*i* dan variabel ke-*j*

2.1 Metode Clustering

Di dalam tugas akhir ini pemecahan permasalahan analisa cluster digunakan metode clustering hierarchical.

Metode Clustering Hierarchical

Metode ini menggunakan struktur yang sama. Dengan memberikan inisial pada n cluster yang menghubungkan n elemen dalam N, didefinisikan jarak $d_{J_iJ_j}^2$ antara dua cluster J_i dan J_j dalam model tertentu. Dengan menggunakan jarak ini dua cluster dikelompokkan menjadi satu cluster.

J_i dan J_j adalah dua cluster yang dikelompokkan menjadi satu cluster

$$d_{J_iJ_j}^2 = \min_{r,s} d_{J_rJ_s}^2$$

Definisi 2.1

$p \in J_i$ dan $q \in J_j$ adalah dua cluster yang dikelompokkan menjadi satu cluster jika $d^2_{J_i J_j} = \min_{\substack{p \in J_i \\ q \in J_j}} d^2_{pq}$ merupakan metode persekitaran terdekat.

Definisi 2.2

$p \in J_i$ dan $q \in J_j$ adalah dua cluster yang dikelompokkan menjadi satu cluster jika $d^2_{J_i J_j} = \max_{\substack{p \in J_i \\ q \in J_j}} d^2_{pq}$ merupakan metode persekitaran terjauh.

Definisi 2.3

$\overline{X}_{J_i} = \frac{1}{n_i} \sum_{p \in J_i} X_p$ dan $\overline{X}_{J_j} = \frac{1}{n_i} \sum_{q \in J_j} X_q$ adalah dua cluster yang dikelompokkan menjadi satu cluster sehingga diperoleh metode terpusat (metode sentroid)

$$d^2_{J_i J_j} = \frac{n_1 n_2}{n_1 + n_2} \left(\overline{X}_{J_i} - \overline{X}_{J_j} \right)' \left(\overline{X}_{J_i} - \overline{X}_{J_j} \right)$$

dimana : $d^2_{J_i J_j}$ = Kuadrat Jarak Euclid antara $\overline{X}_{J_i}$ dan $\overline{X}_{J_j}$.

Dari definisi 2.1 akan dicari jarak terpendek dari data matriks d^2_{ij} dengan metode persekitaran tedekat. Diberikan ilustrasi dari sebuah pabrik textile mempunyai 6 jenis mesin tenun. Dengan menggunakan matriks d^2_{ij} elemen-elemen dalam matriks tersebut menggambarkan hubungan antar variabel sehingga akan digunakan untuk menetapkan tingkat produksi yang optimal dengan efisiensi waktu. Disajikan dalam contoh aplikasi sebagai berikut :

$$((d_{ij}^2)) = \begin{array}{c} \{1\}\ \{2\}\ \{3\}\ \{4\}\ \{5\}\ \{6\} \\ \begin{bmatrix} 0 & 8 & 9 & 4 & 7 & 7 \\ 8 & 0 & 3 & 7 & 4 & 5 \\ 9 & 3 & 0 & 8 & 3 & 5 \\ 4 & 7 & 8 & 0 & 7 & 6 \\ 7 & 4 & 3 & 7 & 0 & 6 \\ 7 & 5 & 5 & 6 & 6 & 0 \end{bmatrix} \end{array}$$

Langkah 1 : mencari jarak terpendek $d_{j_1 j_J}^2 = \min_{\substack{p \in J_i \\ q \in J_j}} d_{pq}^2$

$d_{12}^2 = 8 \quad d_{13}^2 = 9 \quad d_{14}^2 = 4 \quad d_{15}^2 = 7$

$d_{16}^2 = 5 \quad d_{23}^2 = 3 \quad d_{24}^2 = 7 \quad d_{25}^2 = 4$

$d_{26}^2 = 5 \quad d_{34}^2 = 8 \quad d_{35}^2 = 3 \quad d_{36}^2 = 5$

$d_{45}^2 = 7 \quad d_{46}^2 = 7 \quad d_{56}^2 = 6$

didapatkan $d_{23}^2 = 3$ jarak terpendek selanjutnya elemen {2} dan elemen {3} dipadatkan dalam satu cluster sehingga didapat $J_1=\{1\}$, $J_2=\{2,3\}$, $J_3=\{4\}$, $J_5=\{5\}$, $J_6=\{6\}$. kemudian digunakan metoda persekitaran terdekat

$d_{j_1 j_J}^2 = \min_{\substack{p \in J_i \\ q \in J_j}} \{ d_{pq}^2 \}$ $d_{j_1 j_J}^2 = \min_{\substack{p \in (2,3) \\ q \in 1}} \{ d_{12}^2, d_{13}^2 \}$ $= \min \{8,9\} = 8$

$d_{j_1 j_J}^2 = \min_{\substack{p \in (2,3) \\ q \in 4}} \{ d_{24}^2, d_{34}^2 \}$

$= \min \{7,8\}$

$= 7$

$d_{j_1 j_J}^2 = \min_{\substack{p \in (2,3) \\ q \in 5}} \{ d_{25}^2, d_{35}^2 \}$

$= \min\{3,3\}$

$= 3$

$$d^2_{j_1 j_J} = \min_{\substack{p\in(2,3)\\ q\in 6}} \{\, d^2_{26}, d^2_{36} \,\}$$

$= \min\{5,5\}$

$= 5$

kemudian dibentuk matriks baru dengan menggunakan jarak kuadrat di atas dan matriks dinotasikan dengan $((d^2_{ij(1)}))$ dan i,j meneruskan ke-i dan ke-j. Matriks diberikan seperti di bawah:

$$((d^2_{ij(1)})) = \begin{array}{c} \{1\}\ \ \{2,3\}\ \ \{4\}\ \ \{5\}\ \ \{6\} \\ \begin{bmatrix} 0 & 8 & 4 & 7 & 7 \\ 8 & 0 & 7 & 3 & 5 \\ 4 & 7 & 0 & 7 & 6 \\ 7 & 3 & 7 & 0 & 6 \\ 7 & 5 & 6 & 6 & 0 \end{bmatrix} \end{array}$$

Langkah 2 :

Dari jarak kuadrat di atas didapat $d^2_{j_1 j_J} = \min d^2_{ij} = d^2_{23(5)} = 3$.

sebagai jarak terpendek selanjutnya cluster {2,3} dan elemen {5} dipadatkan dalam satu cluster sehingga didapatkan $J_1=\{1\}$, $J_2=\{2,3,5\}$, $J_3=\{4\}$, $J_5=\{6\}$.Kemudian digunakan metoda persekitaran terdekat:

$$d^2_{j_1 j_J(2)} = \min_{\substack{p\in J_1(2)\\ q\in J_J(2)}} \{\, d^2_{pq(2)} \,\}$$

Kemudian dibentuk matriks baru dengan menggunakan jarak kuadrat di atas dan matriks dinotasikan dengan $((d^2_{ij(2)}))$ dan i,j meneruskan ke-i dan ke-j.

Matriks diberikan seperti di bawah ini :

$$((d^2_{ij(2)})) = \begin{array}{c} \{1\} \quad \{2,3,5\} \quad \{4\} \quad \{6\} \\ \begin{bmatrix} 0 & 7 & 4 & 7 \\ 7 & 0 & 7 & 5 \\ 4 & 7 & 0 & 6 \\ 7 & 5 & 6 & 0 \end{bmatrix} \end{array}$$

Langkah 3 :

Dari jarak kuadrat di atas didapat $d^2_{j_1 j_J} = \min d^2_{ij} = d^2_{14} = 4$.

Sebagai jarak terpendek selanjutnya elemen {1} dan elemen {4} dipadatkan dalam satu cluster sehingga didapatkan $J_1=\{1,4\}$, $J_2=\{2,3,5\}$, $J_3=\{6\}$. Kemudian digunakan metoda persekitaran terdekat:

$$d^2_{j_1 j_J(3)} = \min_{\substack{p \in J_{1(3)} \\ q \in J_{J(3)}}} \{ d^2_{pq(3)} \}$$

Kemudian dibentuk matriks baru dengan menggunakan jarak kuadrat di atas dan matriks dinotasikan dengan $((d^2_{ij(3)}))$ dan i,j meneruskan ke-i dan ke-j. Matriks diberikan seperti di bawah ini:

$$((d^2_{ij(3)})) = \begin{array}{c} \{1,4\} \quad \{2,3,5\} \quad \{6\} \\ \begin{bmatrix} 0 & 7 & 6 \\ 7 & 0 & 5 \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

Langkah 4 :

Dari jarak kuadrat di atas didapat $d^2_{j_i j_j} = \min d^2_{ij} = d^2_{(235)(6)} = 5$.

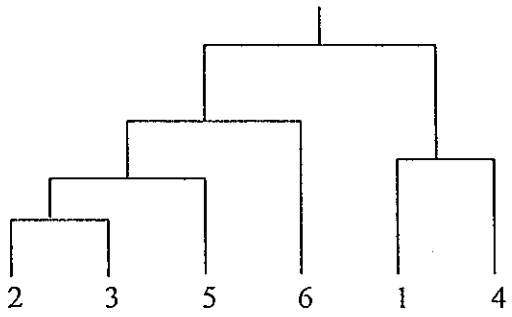
sebagai jarak terpendek selanjutnya elemen {2,3,5} dan elemen {6} dipadatkan dalam satu cluster sehingga didapatkan $J_1 = \{1,4\}$, $J_2=\{2,3,5,6\}$.kemudian digunakan metoda persekitaran terdekat:

$$d^2_{j_i j_j(4)} = \min_{\substack{p \in J_{i(4)} \\ q \in J_{j(4)}}} \{ d^2_{pq(4)} \}$$

kemudian dibentuk matriks baru dengan menggunakan jarak kuadrat di atas dan matriks dinotasikan dengan $((d^2_{ij(4)}))$ dan i,j meneruskan ke-i dan ke-j. Matriks diberikan seperti di bawah ini:

$$((d^2_{ij(4)})) = \begin{array}{c} \\ \{1,4\} \\ \{2,3,5,6\} \end{array} \begin{array}{c} \begin{array}{cc} \{1,4\} & \{2,3,5,6\} \end{array} \\ \begin{bmatrix} 0 & 7 \\ 7 & 0 \end{bmatrix} \end{array}$$

Langkah 5: didapatkan jarak terpendek $d^2_{14(2356)} = 7$ yang digunakan untuk memadatkan elemen {1,4} kedalam cluster {2,3,5,6} sehingga semua unit termuat dalam satu cluster. Dengan langkah-langkah di atas didapat skema clustering hirarki sebagai berikut :



Skema hirarki menunjukkan urutan pengelompokkan unit-unit variabel sehingga terletak dalam satu cluster.