ITS
Institut
Teknologi
Sepuluh Nopember

Cluster Analysis

Ali Ridho Barakbah

Cluster Analysis

- Variance
- Error ratio

Variance

- Digunakan untuk mengukur nilai penyebaran dari data-data hasil clustering
- Dipakai untuk data yang bertipe unsupervised
- Variance pada clustering ada 2 macam:
  - Variance within cluster
  - Variance between clusters

Good cluster

is when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external homogeneity)

Variance & homogeneity

internal homogeneity → Variance within cluster (V_w)

external homogeneity → Variance between clusters (V_b)

Ideal cluster

- The ideal cluster has minimum V_w to express internal homogeneity and maximum V_b to express external homogeneity.

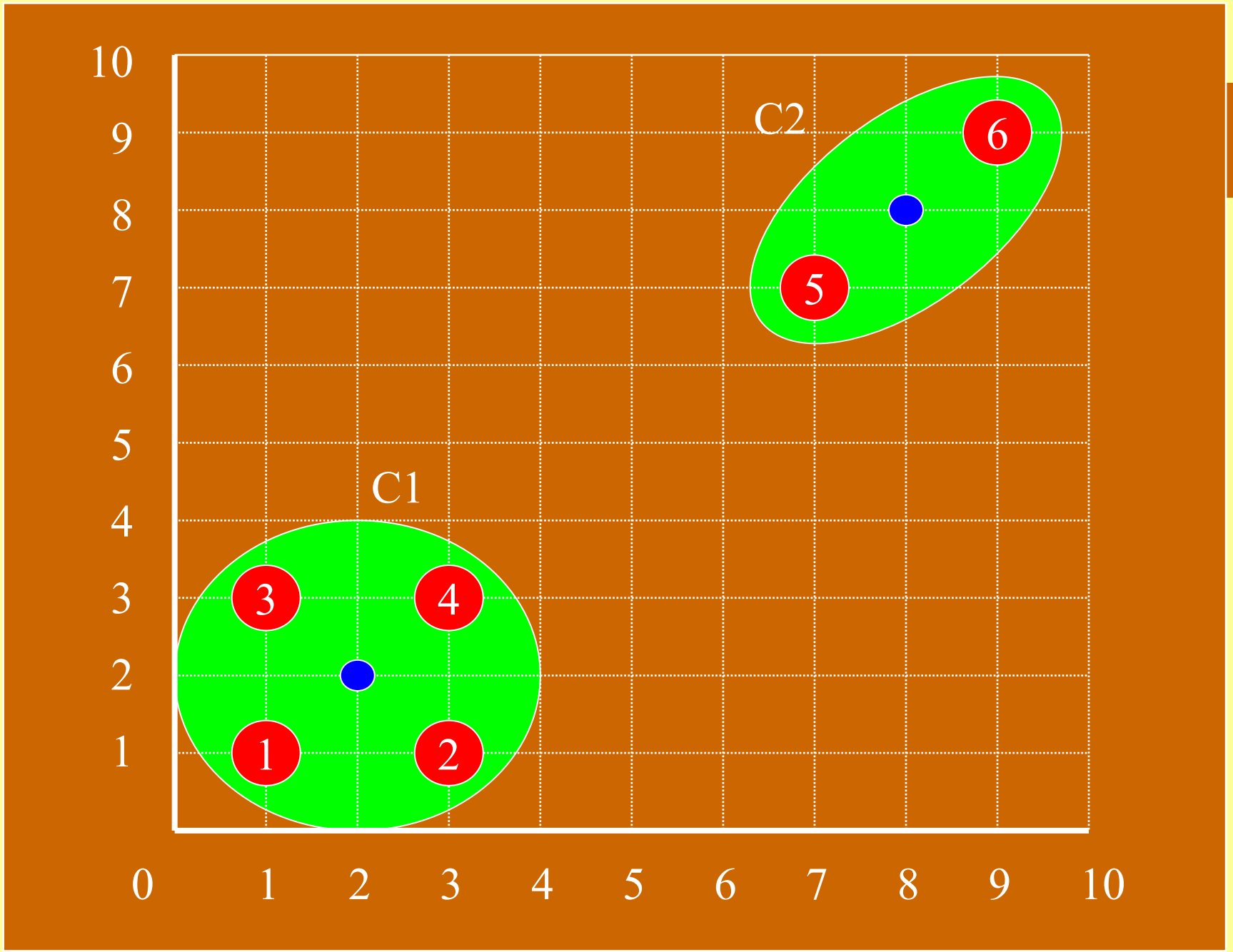
minimum

$$V = \frac{V_w}{V_b}$$

atau

maximum

$$V = \frac{V_b}{V_w}$$


10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
C1
C2
1
2
3
4
5
6

Cluster Variance

$$v_c^{\,2} = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} \left(d_i - \overline{d}_i \right)^2$$

$v_c^{\,2}$ = variance pada cluster c
$c = 1..k$, dimana k = jumlah cluster
n_c = jumlah data pada cluster c
d_i = data ke-i pada suatu cluster
$\overline{d}_i$ = rata-rata dari data pada suatu cluster

Variance within cluster

$$v_w = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) \cdot v_i^2$$

v_w = variance within cluster
N = jumlah semua data

Variance between clusters

$$v_b = \frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\overline{d_i} - \overline{d}\right)^2$$

$\overline{d}$ = rata-rata dari $\overline{d_i}$

Variance dari semua cluster

$$v = \frac{v_w}{v_b}$$

Error ratio

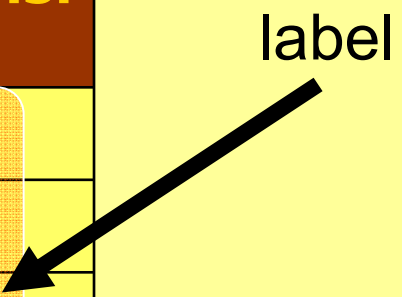
- Dipakai jika dataset yang digunakan adalah supervised
- Biasanya digunakan untuk mengukur tingkat presisi dari metode clustering
- Rumus:

$$Error = \frac{missclassified}{jumlahdata} \times 100\%$$

Contoh sederhana

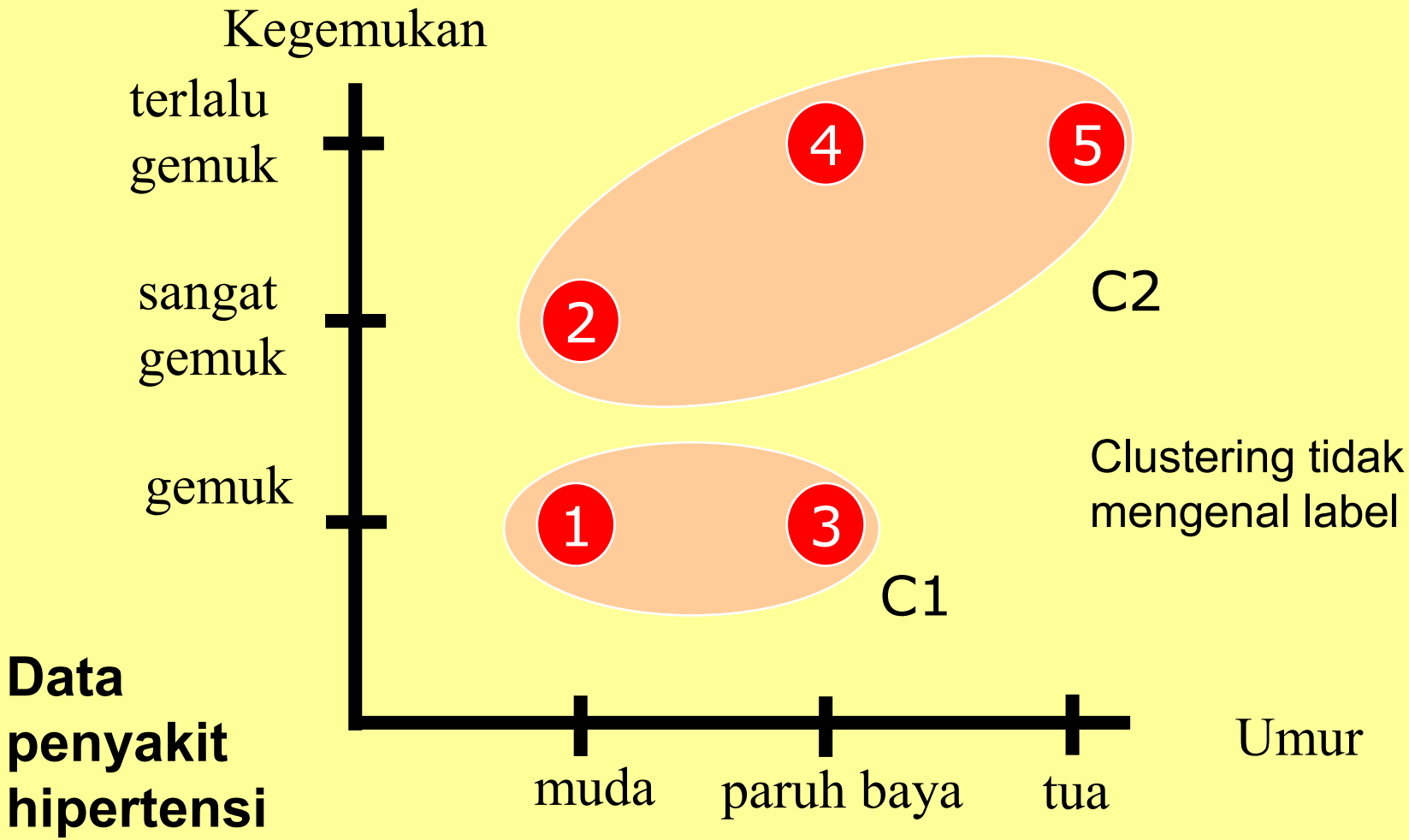
Data penyakit hipertensi

Data ke-	Umur	Kegemukan	Hipertensi
1	muda	gemuk	Tidak
2	muda	sangat gemuk	Tidak
3	paruh baya	gemuk	Tidak
4	paruh baya	terlalu gemuk	Ya
5	tua	terlalu gemuk	Ya



Supervised data

Contoh Hasil Clustering



Data penyakit hipertensi

Menghitung error ratio

	Label	Kombinasi 1 C1 → Tidak C2 → Ya	Kombinasi 2 C1 → Ya C2 → Tidak
Data 1	Tidak	Tidak	Ya
Data 2	Tidak	Ya	Tidak
Data 3	Tidak	Tidak	Ya
Data 4	Ya	Ya	Tidak
Data 5	Ya	Ya	Tidak
Misclassified		1	4
Error ratio		20%	80%

Semua kemungkinan label dicoba sehingga ada n! kombinasi

Ambil error ratio yang terkecil