Machine Learning

Cluster Analysis

Ali Ridho Barakbah
Knowledge Engineering Research Group
Soft Computing Laboratory
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

Politeknik Elektronika Negeri Surabaya
Departemen Teknik Informatika dan Komputer

Konten

- Cluster Analysis
- Variance
- Good Cluster
- Variance Within Cluster
- Variance Between Cluster
- Centroid Proximity Cluster
- Error Ratio

Tujuan Instruksi Umum

Mahasiswa mampu menyelesaikan masalah – masalah menggunakan metode mesin pembelajaran yang tepat berdasarkan supervised, unsupervised dan reinforcement learning, baik secara individu maupun berkelompok/kerjasama tim.

Tujuan Instruksi Khusus

- Memahami bagaimana menilai sebuah hasil klasterisasi yang baik
- Mampu menerapkan Klasterisasi dan Analisa Klaster

Cluster Analysis

- Variance
- Sum of Squared Error
- Centroid Proximity Index
- 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)

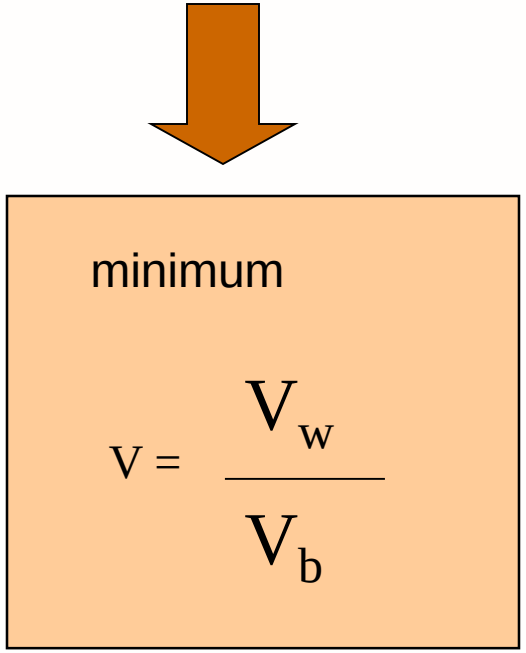
Variance & homogeneity

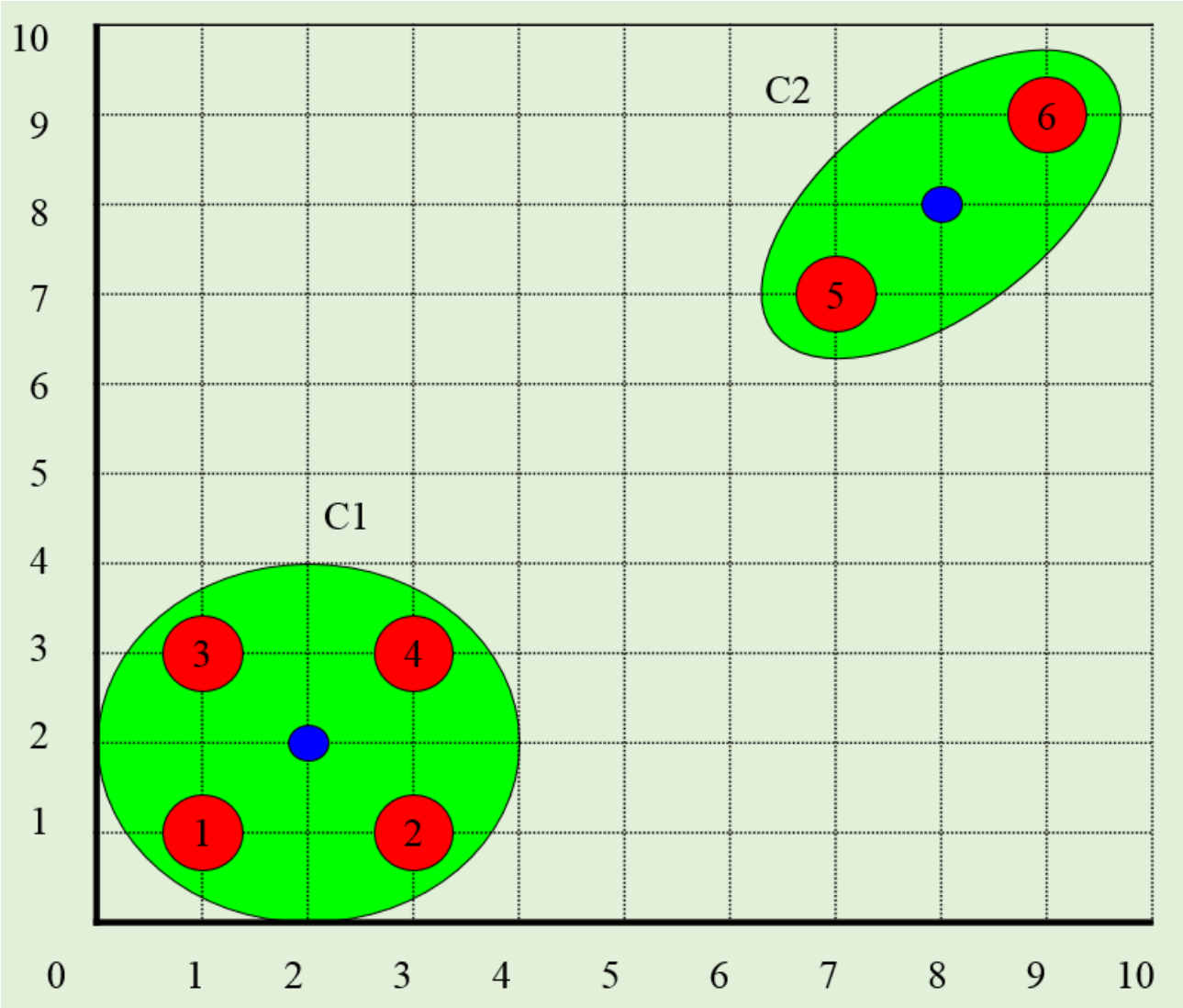
internal homogeneity → Variance within cluster (V_w)

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.




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

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

Sum of Squared Error

- The most widely used criterion to quantify cluster homogeneity is the Sum of Squared Error (SSE) criterion

$$SSE = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n(s_i)} \left\| m_{ij} - \bar{s}_i \right\|^2}{N}$$

Centriod Proximity Index

- Proposed by Barakbah and Kiyoki[1]
- To analyze the closeness of the final centroids of the clustering result to the centroids of the real data sets.

$$CPI = \min \sum_{i=1}^{k} (\|c_i - r_i\|)$$

where c_i is *i*-th final centroid of clustering result and r_i is *i*-th real centroid of datasets.

[1]Ali Ridho Barakbah, Yasushi Kiyoki, "A Fast Algorithm for K-Means Optimization using Pillar Algorithm", The 2nd International Workshop with Mentors on Databases, Web and Information Management for Young Researchers, August 2-4, 2010, Tokyo, Japan.

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

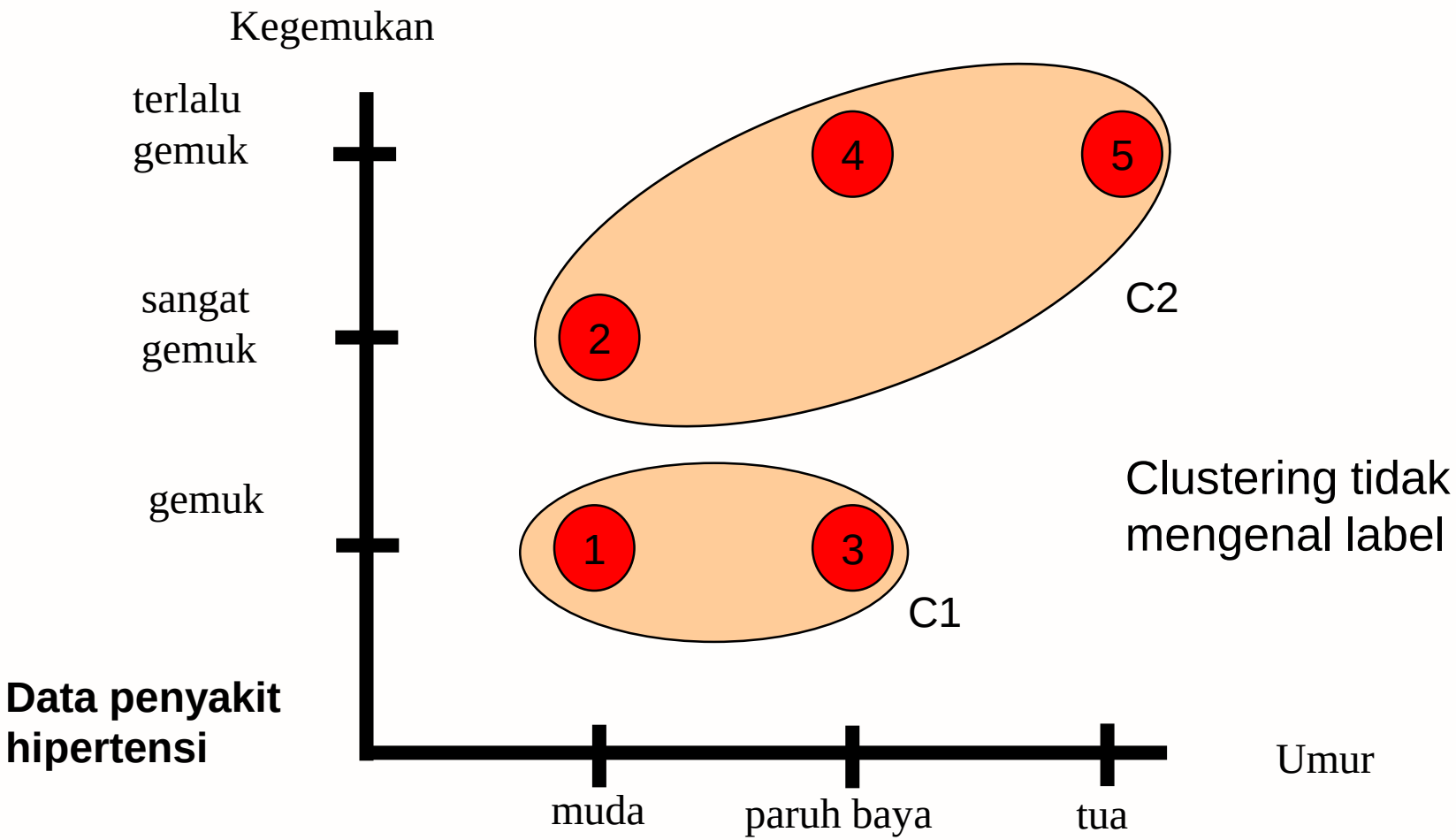
Data penyakit hipertensi

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

label

Supervised data

Contoh Hasil Clustering



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

Latihan Soal

1. Tambahkan analisa klaster pada implementasi K-means untuk data Ruspini dalam program!
2. Tambahkan analisa klaster pada implementasi Hirarki untuk data Ruspini dalam program!

Referensi

- Tom Michael, Machine Learning, McGraw-Hill publisher, 1997.
- Ali Ridho Barakbah, *Machine Learning*, Lecture Handout, Electronic Engineering Polytechnic Institute of Surabaya.
- UCI Repository, Ruspini Dataset.

bridge to the future

http://www.eepis-its.edu