Panduan Praktikum
Clustering (Advance)

Durasi: 3 × 110 menit

Capaian Pembelajaran Mata Kuliah

Mahasiswa mampu menjelaskan metode clustering dan penggunaannya pada suatu permasalahan.

Tools

- Google Colab
- Jupyter Notebook
- PyCharm
- Spyder
- Python IDE yang lain

Materi Praktikum

- Pengertian Clustering

Clustering adalah teknik dalam data mining yang bertujuan untuk mengelompokkan data ke dalam beberapa grup berdasarkan kesamaan karakteristik. Grup-grup ini disebut kluster, di mana data dalam satu kluster memiliki kesamaan yang lebih tinggi dibandingkan dengan data di kluster lain.

- Konsep Lanjutan dalam Clustering

1. Hierarchical Clustering

- Teknik clustering yang menghasilkan hierarki data berbentuk pohon atau dendrogram.

- Metode:

  - Agglomerative: Memulai dengan setiap data sebagai kluster individu, kemudian digabungkan.

  - Divisive: Memulai dengan seluruh data dalam satu kluster besar, lalu dibagi.

2. Density-Based Clustering

- Mengelompokkan data berdasarkan area dengan kepadatan tinggi.

- Contoh Algoritma: DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

- Keunggulan:

  - Dapat menangani bentuk kluster yang tidak teratur.

  - Tahan terhadap outlier.

3. Model-Based Clustering

- Menggunakan pendekatan statistik untuk memodelkan data dan menentukan kluster.

- Contoh Algoritma: Gaussian Mixture Models (GMM).

- Keunggulan:

  - Dapat menangkap kluster dengan bentuk distribusi berbeda.

4. Spectral Clustering

- Menggunakan teori spektral graf untuk mempartisi data ke dalam kluster.

- Berguna untuk dataset dengan struktur kompleks.

- Evaluasi Clustering

1. Internal Evaluation (Tanpa Label)

- Silhouette Score: Mengukur seberapa baik data dalam satu kluster dibandingkan dengan kluster lainnya.

  Formula: Silhouette = (b - a) / max(a, b)

  - a: Rata-rata jarak antar data dalam kluster.

  - b: Rata-rata jarak data ke kluster terdekat.

2. External Evaluation (Dengan Label)

- Rand Index: Mengukur kesesuaian antara hasil clustering dengan label sebenarnya.

- Normalized Mutual Information (NMI): Mengukur informasi bersama antara kluster dan label aktual.

3. Stability-Based Evaluation

- Menguji konsistensi hasil clustering dengan perubahan kecil pada dataset.

- Algoritma Clustering Lanjutan

1. DBSCAN

- Ide Utama: Menemukan area dengan kepadatan tinggi berdasarkan dua parameter utama:

  - Epsilon (eps): Jarak maksimum antar titik.

  - MinPts: Jumlah minimum titik dalam radius eps.

- Keunggulan:

  - Dapat menemukan kluster berbentuk tidak teratur.

  - Mengabaikan noise atau outlier.

2. Gaussian Mixture Models (GMM)

- Ide Utama: Memodelkan distribusi data sebagai kombinasi dari distribusi Gaussian.

- Aplikasi:

  - Pengenalan pola.

  - Pemodelan data yang kompleks.

3. Spectral Clustering

- Langkah:

  1. Representasikan data sebagai graf.

  2. Gunakan nilai eigen dari matriks adjacency graf untuk mempartisi data.

- Keunggulan:

  - Cocok untuk data non-linear atau tidak terpisah dengan baik dalam ruang Euclidean.

4. Agglomerative Clustering

- Pendekatan: Dimulai dengan data sebagai kluster individu dan secara iteratif digabungkan berdasarkan metrik jarak (misalnya, jarak rata-rata atau single linkage).

- Contoh Aplikasi Clustering Lanjutan

1. Segmentasi Pelanggan:

- Menggunakan GMM untuk membagi pelanggan berdasarkan pola pembelian kompleks.

2. Pengenalan Pola dalam Citra:

- DBSCAN untuk mendeteksi objek dalam citra berbasis kepadatan.

3. Analisis Jaringan Sosial:

- Spectral Clustering untuk menemukan komunitas dalam jaringan sosial.

4. Deteksi Anomali:

- DBSCAN untuk mengidentifikasi outlier pada dataset transaksi keuangan.

Panduan Praktikum

Download lalu pelajari dan running code pada link berikut:
https://github.com/aldinata/Modul-Praktikum-Data-Mining/blob/main/Materi/6%20-%20Clustering%20(Advance).ipynb