ISSN: 2614-6754 (print)
ISSN: 2614-3097(online)

Halaman 22961-22967
Volume 9 Nomor 2 Tahun 2025

Analisis Performa Clustering: K-Means dan Similarity Matrix dalam Evaluasi Silhouette, DBI, CHI, dan Dunn Index

Bhima Fajar Ramadhan[1], Bimo Musthafa Abdillah[2], Miftahul Hidayatullah[3], Muhamad Nur Faizal[4], Rio Aditya Winata[5], Zurnan Alfian[6]
[1,2,3,4,5,6] Teknik Informatika, Universitas Pamulang
e-mail: fajarramadan238@gmail.com

Abstrak

Clustering merupakan teknik penting dalam data mining yang bertujuan untuk mengelompokkan data berdasarkan kemiripan antar objek. Penelitian ini membahas analisis performa dua pendekatan clustering, yaitu K-Means dan Similarity Matrix, dalam konteks evaluasi kualitas cluster. Pendekatan Similarity Matrix diterapkan menggunakan hierarchical clustering dengan metode complete-linkage, sedangkan K-Means menggunakan data fitur numerik secara langsung. Keduanya diuji pada beberapa dataset dan dievaluasi menggunakan metrik kuantitatif seperti Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), dan Dunn Index. Hasil eksperimen menunjukkan bahwa pendekatan K-Means cenderung unggul dalam pemisahan cluster (Silhouette dan CHI lebih tinggi), sedangkan pendekatan Similarity Matrix lebih baik dalam kepadatan dan keseragaman cluster (DBI dan Dunn Index lebih rendah). Temuan ini menegaskan pentingnya pemilihan metode clustering yang sesuai dengan karakteristik data dan tujuan analisis.

Kata kunci: *Clustering, K-Means, Similarity Matrix, Silhouette, DBI, CHI, Dunn Index*

Abstract

Clustering is a fundamental technique in data mining used to group data based on similarity between objects. This study analyzes the performance of two clustering approaches—K-Means and Similarity Matrix—in evaluating cluster quality. The Similarity Matrix approach is implemented using hierarchical clustering with the complete-linkage method, while K-Means utilizes raw numerical feature data directly. Both approaches are applied to various datasets and evaluated using quantitative matrix including Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), and Dunn Index. The experimental results indicate that K-Means tends to produce better-separated clusters (with higher Silhouette and CHI scores), whereas the Similarity Matrix approach performs better in terms of cluster compactness and uniformity (as reflected by lower DBI and higher Dunn Index values). These findings highlight the importance of selecting a clustering method that aligns with the characteristics of the data and the objectives of the analysis

Keywords : *Clustering, K-Means, Similarity Matrix, Silhouette, DBI, CHI*

PENDAHULUAN

Clustering merupakan salah satu teknik penting dalam data mining yang digunakan untuk mengelompokkan data berdasarkan tingkat kemiripan antar objek. Salah satu metode yang umum digunakan adalah Hierarchical Clustering, karena mampu menghasilkan struktur hirarki dan divisualisasikan dalam bentuk dendrogram yang informatif. Dalam penerapannya, metode ini dapat dijalankan dengan dua pendekatan data yang berbeda, yaitu menggunakan data fitur numerik asli dan menggunakan proximity matrix yang berisi informasi jarak antar objek.

Kedua pendekatan ini sering digunakan dalam kondisi yang berbeda. Data fitur numerik memberikan fleksibilitas dalam evaluasi performa clustering melalui metrik statistik, sedangkan proximity matrix berguna ketika hanya tersedia data jarak antar entitas tanpa informasi fitur eksplisit. Meski menggunakan metode linkage yang sama, hasil pengelompokan dari kedua pendekatan tersebut dapat menghasilkan struktur cluster yang berbeda.

Oleh karena itu, penelitian ini memfokuskan pada perbandingan hasil clustering antara data fitur dan proximity matrix, menggunakan metode complete-linkage pada hierarchical clustering. Hasil clustering dievaluasi menggunakan matrix kuantitatif seperti Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, dan Dunn Index, untuk menilai kualitas dan efektivitas masing-masing pendekatan dalam membentuk cluster yang padat dan terpisah dengan baik.

Berdasarkan latar belakang di atas, maka penelitian ini bertujuan untuk: (1) Untuk menerapkan metode Hierarchical Clustering dengan complete-linkage menggunakan dua pendekatan data, yaitu proximity matrix dan data fitur asli. (2) Untuk menghasilkan visualisasi struktur pengelompokan data dalam bentuk dendrogram dari masing-masing pendekatan. (3) Untuk mengevaluasi performa hasil clustering dari kedua pendekatan dengan menggunakan matrix kuantitatif seperti Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, dan Dunn Index. (4) Untuk membandingkan efektivitas kedua pendekatan dalam membentuk cluster yang representatif dan terpisah dengan baik.

Clustering merupakan salah satu metode yang sering digunakan dalam data mining untuk mengelompokkan data berdasarkan kemiripan karakteristik antar objek. Salah satu teknik paling populer adalah K-Means, yang pertama kali diperkenalkan oleh MacQueen [1]. Kelebihan utama K-Means terletak pada kecepatan dan kesederhanaannya dalam mengelompokkan data numerik. Namun, algoritma ini memiliki beberapa keterbatasan, seperti ketergantungan terhadap inisialisasi awal dan asumsi bentuk cluster yang cenderung konveks [2].

Seiring berkembangnya kebutuhan analisis data, pendekatan alternatif seperti penggunaan similarity matrix menjadi semakin relevan, terutama untuk data yang tidak memiliki representasi fitur numerik secara eksplisit. Dalam pendekatan ini, informasi kedekatan antar objek disusun dalam bentuk matrix jarak atau kemiripan, yang kemudian dapat dianalisis menggunakan metode seperti hierarchical clustering [3]. Salah satu varian hierarchical clustering yang banyak digunakan adalah complete-linkage, karena mampu membentuk struktur hierarki yang dapat divisualisasikan dalam bentuk dendrogram [4].

Untuk menilai kualitas hasil clustering, peneliti biasanya menggunakan matrix evaluasi internal. Matrix yang sering digunakan antara lain Silhouette Score, yang mengukur sejauh mana sebuah objek mirip dengan cluster-nya sendiri dibandingkan dengan cluster lain, serta Davies-Bouldin Index (DBI) yang menilai rasio jarak antar cluster terhadap sebaran dalam cluster [5]. Selain itu, Calinski-Harabasz Index (CHI) dan Dunn Index juga umum digunakan sebagai indikator pemisahan dan kekompakan cluster.

Studi komparatif oleh Xu dan Wunsch menunjukkan bahwa tidak ada matrix evaluasi tunggal yang bisa secara mutlak digunakan untuk semua jenis data. Oleh karena itu, pemilihan metrik sebaiknya disesuaikan dengan karakteristik data dan tujuan analisis [6]. Di sisi lain, Ester et al. menemukan bahwa performa algoritma clustering sangat dipengaruhi oleh struktur distribusi data, serta adanya outlier atau noise [7].

Dalam konteks data yang memiliki bentuk cluster tidak teratur atau tidak konveks, pendekatan berbasis similarity matrix sering kali lebih mampu menangkap pola alami dalam data [8]. Meskipun demikian, K-Means tetap menjadi pilihan utama ketika berhadapan dengan data yang besar, terstruktur, dan berdimensi rendah karena efisiensinya yang tinggi [9]

Berdasarkan kajian tersebut, penelitian ini berusaha membandingkan secara sistematis performa K-Means dan hierarchical clustering berbasis similarity matrix, dengan menggunakan matrix evaluasi yang beragam, agar diperoleh gambaran yang lebih komprehensif tentang kelebihan dan kekurangan masing-masing pendekatan

Penelitian ini menggunakan metode eksperimen kuantitatif komparatif berbasis simulasi komputasi, yang bertujuan untuk membandingkan performa dua pendekatan clustering: Hierarchical Clustering berbasis proximity matrix dan Hierarchical Clustering berbasis data fitur asli (fitur numerik).

Kedua pendekatan diimplementasikan menggunakan metode complete-linkage dan dianalisis berdasarkan struktur cluster yang dihasilkan. Proses clustering dilakukan melalui simulasi dengan bantuan perangkat lunak Python, dan hasilnya dievaluasi menggunakan matrix kuantitatif seperti Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI),

dan Dunn Index. Evaluasi ini dilakukan untuk mengukur kepadatan, pemisahan, serta kualitas keseluruhan dari hasil pengelompokan. Dengan demikian, metode ini memungkinkan analisis objektif terhadap efektivitas masing-masing pendekatan dalam menghasilkan cluster yang representatif.

METODE

Desain Penelitian

Penelitian ini bersifat eksperimen kuantitatif yang membandingkan hasil clustering pada berbagai dataset menggunakan matrix evaluasi yang disebutkan.

Dataset

- Data biomekanik tulang belakang (vertebral column)
- Data pengangguran provinsi Indonesia
- Data laba-rugi perusahaan
- Data luas daerah jakarta
- Similarity matrix antar objek

Prosedur Clustering

- persiapan data
- penentuan jumlah cluster
- clustering dengan similarity
- evaluasi performa
- perbandingan dan interpretasi

Proses Evaluasi

1. Siapkan hasil clustering
   Dari proses clustering (baik K-Means maupun Similarity Matrix), dapatkan label cluster untuk setiap data poin. Misal, hasil label cluster untuk dataset Pengangguran Indonesia, Vertebral Column, dll.
2. Hitung matrix evaluasi untuk setiap clustering
   Gunakan rumus dan definisi berikut untuk menghitung metrik:

Silhouette Score

Mengukur seberapa mirip sebuah objek dengan cluster-nya dibandingkan dengan cluster lain.
Rumus singkat:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

di mana:
a(i)a(i)a(i) = rata-rata jarak objek i ke anggota cluster-nya sendiri
b(i)b(i)b(i) = jarak rata-rata objek i ke cluster terdekat berikutnya
Nilai s(i)s(i)s(i) berkisar dari -1 sampai 1, semakin tinggi semakin baik.

Davies-Bouldin Index (DBI)

Mengukur rata-rata kesamaan antara cluster dan cluster lain, dihitung dengan rasio jarak intra-cluster dan inter-cluster.

$$BI = \frac{1}{k} \sum_{i=0}^{n} \max_{i \neq j} \left(\frac{Si + Si}{Mij} \right)$$

di mana:
SiS_iSi = jarak rata-rata dari semua titik dalam cluster i ke centroid cluster i
$Mij\_\{ij\}$ = jarak antara centroid cluster i dan j
Nilai lebih kecil menunjukkan cluster yang baik.

Calinski-Harabasz Index (CHI)

Mengukur rasio variansi antar cluster terhadap variansi intra cluster.
Formula:

$$CHI = \frac{trace(B_k)}{trace(W_k)} \times \frac{n - k}{k - 1}$$

di mana:
BkB_kBk = scatter matrix antar cluster
WkW_kWk = scatter matrix intra cluster
nnn = jumlah total data
kkk = jumlah cluster
Nilai lebih besar menunjukkan clustering yang baik.

Dunn Index

Mengukur rasio jarak minimum antar cluster terhadap jarak maksimum dalam cluster.di mana:

$$Dunn = \frac{min_{1\le i\le j\le k} d(C_i, C_j)}{max_{1\le i\le j\le k} diam(C_l)}$$

$d(Ci, Cj) d(C\_i, C\_j) d(Ci, Cj)$ = jarak antara cluster i dan j
$diam(C_l) diam(C\_l) diam(C_l)$ = diameter cluster l (jarak maksimum antara dua titik dalam cluster)

Nilai lebih tinggi berarti cluster lebih baik.

Implementasi Perhitungan

Gunakan tools atau library statistik/ML seperti:
Python (sklearn.metrics untuk Silhouette, DBI, CHI).

Implementasi manual atau library tambahan untuk Dunn Index (karena tidak tersedia langsung di sklearn). Hitung semua metrik di atas untuk hasil clustering dari setiap dataset dan metode.

Bandingkan Hasil Evaluasi

Buat tabel perbandingan nilai matrix dari K-Means dan Similarity Matrix (seperti yang sudah kamu punya).

Interpretasikan nilai-nilai tersebut:

Metode dengan Silhouette dan Calinski-Harabasz Index (CHI) lebih tinggi, Davies-Bouldin Index (DBI) lebih rendah, Dunn lebih tinggi $\rightarrow$ lebih baik.

Fokuskan pada dataset dengan jumlah cluster yang sama agar perbandingan valid.

Simpulkan Performa Clustering

Dari tabel perbandingan, tentukan metode clustering terbaik berdasarkan keseimbangan matrix. Jelaskan kelebihan dan kelemahan masing-masing metode dari hasil evaluasi.

HASIL DAN PEMBAHASAN

- Silhouette: Nilai lebih tinggi (maks 1) berarti cluster lebih baik (pemisahan dan kepadatan cluster bagus).
- Davies-Bouldin Index (DBI): Nilai lebih rendah lebih baik (cluster compact dan terpisah dengan baik).
- Calinski-Harabasz Index (CHI): Nilai lebih tinggi lebih baik (cluster lebih terpisah dan padat).
- Dunn Index: Nilai lebih tinggi lebih baik (jarak antar cluster besar dan jarak intra cluster kecil).

Tabel 1. Hasil Evaluasi

Data set	Cluster	Silhouette	DBI	CHI	Dunn Index
Luas Daerah Jakarta	2	0.5267	0.2177	11.8951	1.6723
Laba-Rugi PT Matahari	2	0.1691	0.3725	2.8637	0.5354
Pengangguran Indonesia	3	0.584	0.490	51.748.0.251	0.251

Biomekanik Tulang Belakang	3	0.449	0.631	242.356	0.069
Similarity Matrix	3	0.40	-0.39	8.02	1.28

Evaluasi Matrix

1. Silhouette Score

   Mengukur seberapa baik tiap data berada di dalam cluster-nya sendiri dibandingkan dengan cluster lain. Nilai berkisar dari -1 sampai 1, nilai tinggi berarti cluster yang terbentuk jelas dan terpisah.

   Data : Pengangguran Indonesia (K-Means) = 0.584 (paling tinggi → cluster paling jelas), Similarity Matrix = 0.40 (cukup baik),Vertebral Column (K-Means) = 0.449 Interpretasi: K-Means pada Pengangguran Indonesia menghasilkan cluster yang lebih terpisah dan jelas dibanding Similarity
2. Davies-Bouldin Index (DBI)

   Mengukur compactness dan pemisahan cluster, nilai kecil lebih baik.

   Data :Similarity Matrix = -0.39 (nilai negatif sangat bagus, menunjukkan cluster sangat baik), Luas Daerah Jakarta (K-Means) = 0.2177, Pengangguran Indonesia (K-Means) = 0.490, Vertebral Column (K-Means) = 0.631

   Interpretasi: Similarity Matrix unggul dalam compactness dan pemisahan cluster, bahkan jauh lebih baik dibanding K-Means.
3. Calinski-Harabasz Index (CHI)

   Mengukur rasio variansi antar cluster terhadap variansi dalam cluster, nilai besar lebih baik.

   Data :Vertebral Column (K-Means) = 242.356 (sangat tinggi), Pengangguran Indonesia (K-Means) = 51.748, Similarity Matrix = 8.02

   Interpretasi: K-Means terutama pada Vertebral Column menghasilkan cluster dengan perbedaan yang sangat jelas antar cluster dibanding Similarity Matrix. Namun, nilai CHI yang terlalu tinggi kadang bisa menandakan overfitting.
4. Dunn Index

   Mengukur jarak antar cluster relatif terhadap ukuran cluster, nilai tinggi lebih baik.

   Data :Luas Daerah Jakarta (K-Means) = 1.6723 (tertinggi), Similarity Matrix = 1.28, Pengangguran Indonesia (K-Means) = 0.251

   Interpretasi: Luas Daerah Jakarta dan Similarity Matrix menunjukkan cluster yang cukup baik secara pemisahan dan compactness.

Interpretasi Hasil

1. Silhouette Score

   Dataset Pengangguran Indonesia dengan K-Means memiliki nilai Silhouette tertinggi (0.584), yang menunjukkan bahwa cluster yang terbentuk cukup jelas dan setiap data lebih dekat ke cluster-nya sendiri dibanding cluster lain.

   Nilai Silhouette Similarity Matrix (0.40) masih cukup baik, tetapi lebih rendah dari K-Means pada dataset ini, menunjukkan bahwa K-Means mampu membentuk cluster yang lebih terpisah.
2. Davies-Bouldin Index (DBI)

   Similarity Matrix menunjukkan nilai DBI negatif (-0.39), yang mengindikasikan cluster yang sangat baik dalam hal compactness dan pemisahan antar cluster.

   Nilai DBI pada K-Means cenderung positif dan lebih tinggi, menunjukkan bahwa cluster K-Means masih kurang compact dan memiliki pemisahan yang lebih rendah dibanding Similarity Matrix.
3. Calinski-Harabasz Index (CHI)

   Nilai CHI tertinggi ditemukan pada K-Means untuk dataset Vertebral Column (242.356), mengindikasikan pemisahan antar cluster yang sangat jelas dan compactness yang baik.

Nilai CHI Similarity Matrix jauh lebih rendah (8.02), menunjukkan bahwa meskipun cluster terbentuk dengan compact, namun pemisahan antar cluster kurang optimal dibanding K-Means pada dataset tersebut.

4. Dunn Index

Nilai Dunn Index tertinggi diperoleh pada dataset Luas Daerah Jakarta dengan K-Means (1.6723), menunjukkan jarak antar cluster yang besar relatif terhadap ukuran cluster, sehingga cluster lebih terpisah dan compact.

Similarity Matrix juga menunjukkan nilai Dunn yang cukup baik (1.28), masih cukup kompetitif dengan K-Means pada beberapa dataset.

Komparasi Metode

- Similarity Matrix unggul dalam compactness dan pemilahan cluster secara global (ditunjukkan oleh DBI dan Dunn Index).
- K-Means unggul dalam menghasilkan cluster dengan pemisahan jelas dan kepadatan yang baik secara lokal (terlihat dari Silhouette Score dan CHI).

Pilihan metode tergantung pada kebutuhan analisis:

Jika mengutamakan cluster yang kompak dan terpisah secara keseluruhan → Similarity Matrix lebih cocok. Jika fokus pada pemisahan cluster yang jelas dan padat secara lokal → K-Means lebih disarankan.

SIMPULAN

Berdasarkan hasil evaluasi clustering menggunakan metrik Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), dan Dunn Index, dapat disimpulkan bahwa setiap metode clustering memiliki keunggulan pada aspek yang berbeda.

Dengan demikian, dapat disimpulkan bahwa pemilihan metode clustering sangat bergantung pada tujuan analisis: Jika fokus pada pemisahan cluster secara tegas dan visualisasi yang jelas → K-Means lebih tepat digunakan. Jika mengutamakan kestabilan dan kepadatan cluster berdasarkan hubungan antar objek → Similarity Matrix lebih sesuai.

DAFTAR PUSTAKA

Asegaf, M. M., Arfianti, U. I., & Hamdani, A. R. (2022). Implementasi metode K-Means clustering dalam pengelompokan penyebaran COVID-19 di Surabaya. *Open Science Framework*.

docplayer.info. (2025). [Sumber dokumen digital]. [Online]. Tersedia: https://docplayer.info. [Diakses: 16-Jun-2025].

garuda.kemdikbud.go.id. (2025). *Garuda | Garba Rujukan Digital*. [Online]. Tersedia: https://garuda.kemdikbud.go.id. [Diakses: 16-Jun-2025].

Hasan, Y. (2024). Pengukuran silhouette score dan Davies-Bouldin index pada hasil cluster K-Means dan DBSCAN. *Jurnal Informatika dan Teknik Elektro Terapan*.

Hermawati, F. A. (2013). *Data mining*. Surabaya: Andi.

inass.org. (2025). *International Association of Applied Science & Engineering*. [Online]. Tersedia: https://www.inass.org. [Diakses: 16-Jun-2025].

jurnal.uia.ac.id. (2025). *Jurnal Universitas Islam As-Syafi'iyah*. [Online]. Tersedia: https://jurnal.uia.ac.id. [Diakses: 16-Jun-2025].

Johannes, E. B. (2021). Indexing pada sistem penalaran berbasis kasus menggunakan metode complete-linkage clustering. *ALE Proceeding*.

kc.umn.ac.id. (2025). *Beranda | Knowledge Center UMN*. [Online]. Tersedia: https://kc.umn.ac.id. [Diakses: 16-Jun-2025].

Prasetyo, E. (2012). *Data mining: Konsep dan aplikasi menggunakan MATLAB*. Gresik: Andi Yogyakarta.

Raharja, P. (2019). *Pengantar ilmu ekonomi*. Jakarta: Salemba Empat.

Sigit, R. (2024). Penerapan algoritma K-Means clustering dalam menganalisis pola peminjaman buku di perpustakaan. *The Indonesian Journal of Computer Science*.

Simbolon, I. N., & Friskila, P. D. (2024). Analisis dan evaluasi algoritma DBSCAN pada tuberkulosis. *Jurnal Informatika dan Teknik Elektro Terapan*.

Sjafrizal. (2016). *Perencanaan pembangunan daerah*. Jakarta: Rajawali Pers.
Sujarweni, V. W. (2017). *Manajemen keuangan: Teori, aplikasi dan hasil penelitian*. Yogyakarta: Pustaka Baru Press.
www.chungnam.net. (2025). *Chungnam Official Website*. [Online]. Tersedia: https://www.chungnam.net. [Diakses: 16-Jun-2025].
www.coursehero.com. (2025). *Course Hero Academic Resources*. [Online]. Tersedia: https://www.coursehero.com. [Diakses: 16-Jun-2025].
www.scribd.com. (2025). *Scribd Document Sharing Platform*. [Online]. Tersedia: https://www.scribd.com. [Diakses: 16-Jun-2025].
Yeung, C., Wang, J., Du, Y., Cao, J., Zhou, Q., Du, Z., Fan, Y., Ding, Y., & Cai, L. (2024). Clearness index cluster analysis for photovoltaic weather classification based on solar irradiation measurement data. *Energy, 360*.
zombiedoc.com. (2025). [Sumber dokumen digital]. [Online]. Tersedia: https://zombiedoc.com. [Diakses: 16-Jun-2025].