Djtechno : Jurnal Teknologi Informasi
Vol. 5, No. 2, Agustus 2024
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

REVIEW TERBARU TENTANG KLASTERISASI DATA MINING MENGGUNAKAN METODE K-MEANS: TANTANGAN DAN APLIKASI

Sabrina Aulia Rahmah
Teknologi Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Dharmawangsa, Indonesia

Article Info

Article history:

Received: 24 Juli 2024
Revised: 26 Juli 2024
Accepted: 06 Agustus 2024

ABSTRACT

Abstrak

Klasterisasi data mining adalah teknik penting yang digunakan untuk mengelompokkan data menjadi subset yang bermakna, dan metode K-Means adalah salah satu algoritma yang paling populer dalam bidang ini. Artikel ini memberikan tinjauan komprehensif mengenai literatur terbaru tentang penggunaan metode K-Means, dengan fokus pada tantangan yang dihadapi dan solusi yang diusulkan, serta aplikasi praktis di berbagai domain. Berbagai inovasi dalam metode K-Means, seperti K-Means++ dan algoritma hibrida, telah dikembangkan untuk mengatasi masalah penentuan jumlah klaster dan sensitivitas terhadap outlier. Selain itu, artikel ini mengeksplorasi aplikasi K-Means dalam bidang pemasaran, kesehatan, dan teknologi informasi. Meskipun metode K-Means memiliki beberapa keterbatasan, berbagai studi menunjukkan peningkatan kinerja yang signifikan melalui modifikasi dan adaptasi yang tepat. Dengan demikian, review ini tidak hanya menyoroti kemajuan terbaru dalam klasterisasi menggunakan K-Means tetapi juga mengidentifikasi area yang memerlukan penelitian lebih lanjut.

Kata Kunci: Klasterisasi, K-Means, Tantangan dan Aplikasi

Abstract

Data mining clustering is an important technique used to group data into meaningful subsets, and the K-Means method is one of the most popular algorithms in this field. This article provides a comprehensive review of the recent literature on the use of K-Means methods, focusing on the challenges faced and proposed solutions, as well as practical applications in various domains. Various innovations in the K-Means method, such as K-Means++ and hybrid algorithms, have been developed to address the problems of determining the number of clusters and sensitivity to outliers. In addition, this article explores the applications of K-Means in marketing, healthcare, and information technology. Although the K-Means method has some limitations, various studies show significant performance improvements through appropriate modifications and adaptations. Thus, this review not only highlights recent advances in clustering using K-Means but also identifies areas that require further research.

Keywords: Clustering, K-Means, Challenges and Applications

Corresponding Author:
Pilih penulis yang akan menjadi korespondensi author
E-mail : sabrinaaulia@dharmawangsa.ac.id

1. PENDAHULUAN

Metode K-Means telah menjadi salah satu algoritma klasterisasi yang paling populer dan sering digunakan dalam data mining karena kesederhanaan dan efisiensinya. Algoritma ini pertama kali diperkenalkan oleh MacQueen pada tahun 1967 dan terus berkembang hingga saat ini. K-Means digunakan untuk mengelompokkan data ke dalam k klaster berdasarkan kedekatan centroid, dimana setiap klaster memiliki centroid yang menjadi pusat dari klaster tersebut. Popularitas metode ini didorong oleh berbagai aplikasinya di berbagai bidang seperti pemasaran, kesehatan, teknologi informasi, dan lain-lain (Zahroh, 2024).

Meskipun memiliki banyak keunggulan, metode K-Means juga menghadapi beberapa tantangan. Salah satu tantangan utama adalah penentuan jumlah klaster (k) yang optimal, yang dapat sangat mempengaruhi hasil akhir klasterisasi. Penelitian oleh Maryam (2024) menunjukkan bahwa penentuan jumlah klaster yang optimal dapat meningkatkan akurasi dalam pengelompokan data sekolah dasar, namun hal ini masih menjadi area yang memerlukan lebih banyak eksplorasi dan penelitian lebih lanjut. Selain itu, K-Means sensitif terhadap inisialisasi centroid yang dapat mempengaruhi konvergensi dan hasil akhir.

Selain penentuan jumlah klaster, inisialisasi centroid juga menjadi tantangan signifikan dalam penggunaan K-Means. Studi yang dilakukan oleh Paridah (2024) menunjukkan bahwa penggunaan metode K-Means dalam klasterisasi penerima dana bantuan dapat menghasilkan variasi yang signifikan dalam hasil akhir tergantung pada inisialisasi centroid yang digunakan. Oleh karena itu, berbagai teknik seperti K-Means++ telah diusulkan untuk mengatasi masalah ini dengan memilih centroid awal secara lebih efektif dan meningkatkan kinerja algoritma secara keseluruhan.

Penelitian terbaru juga menunjukkan aplikasi praktis metode K-Means di berbagai bidang. Sebagai contoh, Subur (2024) menggunakan algoritma K-Means untuk mengklasifikasikan penduduk miskin sebagai penerima bantuan sosial, yang

membantu pemerintah dalam mendistribusikan bantuan secara lebih tepat sasaran. Demikian pula, Rivaldi (2024) mengembangkan sistem informasi geografis untuk memetakan sebaran keluarga miskin menggunakan metode K-Means, yang dapat digunakan untuk analisis spasial dan perencanaan kebijakan.

Meskipun demikian, masih banyak area yang memerlukan penelitian lebih lanjut. Studi oleh Helia (2024) mengindikasikan bahwa metode K-Means dapat digunakan untuk menganalisis daya tarik obyek wisata berdasarkan jenisnya, namun hasilnya menunjukkan perlunya penyesuaian lebih lanjut agar metode ini dapat diterapkan secara efektif dalam berbagai konteks data yang berbeda. Dengan demikian, meskipun metode K-Means telah mengalami banyak pengembangan, tantangan yang ada menunjukkan perlunya penelitian yang lebih mendalam untuk meningkatkan akurasi dan aplikasi dari algoritma ini dalam berbagai domain.

2. MASALAH PENELITIAN UTAMA

Salah satu masalah utama dalam penggunaan metode K-Means adalah penentuan jumlah klaster yang optimal. Penentuan jumlah klaster yang tidak tepat dapat mengakibatkan hasil klasterisasi yang kurang akurat dan tidak dapat diandalkan. Penelitian oleh Maryam (2024) menunjukkan bahwa optimalisasi jumlah klaster dalam data sekolah dasar dapat meningkatkan keakuratan pengelompokan data, namun proses ini sering kali kompleks dan memerlukan pemahaman yang mendalam tentang data yang dianalisis. Tantangan ini menjadi semakin signifikan ketika berhadapan dengan dataset besar dan heterogen, di mana variasi dalam data dapat mempengaruhi hasil akhir secara signifikan.

Masalah lain yang signifikan dalam metode K-Means adalah inisialisasi centroid. Inisialisasi centroid yang buruk dapat menyebabkan hasil klasterisasi yang suboptimal dan mempengaruhi konvergensi algoritma. Studi oleh Paridah (2024) dalam konteks penerima dana bantuan menunjukkan bahwa variasi inisialisasi centroid dapat menghasilkan hasil yang sangat berbeda. Pendekatan K-Means++ telah diusulkan untuk mengatasi masalah ini dengan memilih centroid awal secara lebih efektif, namun tetap diperlukan penelitian lebih lanjut untuk memastikan bahwa metode ini dapat diterapkan dengan konsisten dalam berbagai jenis data.

Sensitivitas metode K-Means terhadap outlier juga merupakan tantangan yang perlu diatasi. Outlier dapat sangat mempengaruhi hasil klasterisasi dengan menggeser centroid dan menciptakan klaster yang tidak representatif. Penelitian oleh Subur (2024) tentang klasterisasi penduduk miskin untuk penerima bantuan sosial menunjukkan bahwa keberadaan outlier dapat mengganggu distribusi data dalam klaster, sehingga mengurangi akurasi hasil. Oleh karena itu, teknik pra-pemrosesan data, seperti penghapusan outlier atau penggunaan algoritma klasterisasi yang lebih tahan terhadap outlier, menjadi penting dalam aplikasi praktis K-Means.

Selain itu, metode K-Means sering kali menghadapi tantangan dalam menangani data berdimensi tinggi. Ketika jumlah fitur dalam dataset meningkat, efektivitas metode K-Means dapat menurun karena kesulitan dalam menemukan struktur klaster yang bermakna. Penelitian oleh Riyandoro (2023) mengenai pengelompokan merek laptop menunjukkan bahwa dengan peningkatan dimensi data, metode K-Means memerlukan pendekatan tambahan seperti reduksi dimensi untuk meningkatkan kinerja klasterisasi. Ini menunjukkan perlunya pengembangan teknik yang dapat mengatasi tantangan data berdimensi tinggi dalam aplikasi K-Means.

Untuk mengatasi berbagai tantangan ini, berbagai solusi telah diusulkan dalam literatur. Salah satunya adalah penggunaan metode hibrida yang menggabungkan K-Means dengan algoritma lain seperti algoritma genetika atau optimasi kawanan partikel. Studi oleh Samsudin (2024) menunjukkan bahwa kombinasi metode ini dapat meningkatkan kinerja klasterisasi dalam manajemen persediaan bisnis modern. Pendekatan lain termasuk pengembangan algoritma yang lebih adaptif dan penggunaan teknik validasi klaster yang lebih robust untuk memastikan bahwa hasil klasterisasi akurat dan dapat diandalkan dalam berbagai kondisi data.

3. HASIL DAN PEMBAHASAN

Penelitian ini mengulas berbagai aplikasi metode K-Means dalam berbagai domain, menunjukkan bahwa algoritma ini tetap menjadi pilihan utama untuk klasterisasi data meskipun menghadapi beberapa tantangan. Misalnya, Zahroh (2024) berhasil menerapkan K-Means untuk mengklasifikasikan data kegemaran membaca

siswa di SMA Al-Islam Cirebon. Hasil klasterisasi menunjukkan bahwa terdapat tiga kelompok utama siswa berdasarkan minat bacanya, yaitu kelompok dengan minat baca tinggi, sedang, dan rendah. Temuan ini penting bagi pihak sekolah untuk merancang program literasi yang lebih efektif dan sesuai dengan kebutuhan setiap kelompok siswa.

Maryam (2024) menyoroti pentingnya optimalisasi jumlah klaster dalam klasterisasi data sekolah dasar (SD). Dengan menggunakan K-Means, ia mampu mengidentifikasi jumlah klaster optimal yang memaksimalkan keakuratan pengelompokan data siswa. Hasil penelitiannya menunjukkan bahwa penggunaan metode silhouette dan elbow membantu dalam menentukan jumlah klaster yang tepat, yang selanjutnya meningkatkan keakuratan hasil klasterisasi. Penerapan metode ini dapat membantu sekolah dalam mengelompokkan siswa berdasarkan kemampuan akademis mereka, sehingga program pendidikan dapat disesuaikan dengan kebutuhan tiap klaster.

Studi oleh Paridah (2024) mengenai klasterisasi penerima dana bantuan program keluarga harapan di Desa Gereba mengungkapkan bahwa metode K-Means efektif dalam mengelompokkan penerima bantuan berdasarkan kondisi ekonomi mereka. Dengan mengidentifikasi tiga klaster utama penerima bantuan, yaitu sangat membutuhkan, membutuhkan, dan kurang membutuhkan, penyaluran dana bantuan dapat dilakukan dengan lebih tepat sasaran. Hal ini menunjukkan bagaimana K-Means dapat digunakan untuk meningkatkan efisiensi program sosial pemerintah, memastikan bahwa bantuan diberikan kepada yang benar-benar membutuhkan.

Subur (2024) juga menunjukkan keberhasilan penggunaan K-Means dalam mengklasifikasikan penduduk miskin sebagai penerima bantuan sosial. Hasil klasterisasi menunjukkan bahwa penduduk dapat dikelompokkan menjadi beberapa klaster berdasarkan tingkat kemiskinan mereka. Dengan demikian, pihak berwenang dapat merancang kebijakan yang lebih terfokus dan efektif dalam menanggulangi kemiskinan. Ini menunjukkan pentingnya penggunaan teknik klasterisasi yang tepat dalam pengelolaan data sosial untuk meningkatkan kualitas hidup masyarakat.

Rivaldi (2024) mengembangkan sistem informasi geografis (SIG) untuk memetakan sebaran keluarga miskin menggunakan metode K-Means. Hasil penelitian menunjukkan bahwa K-Means dapat digunakan untuk mengidentifikasi wilayah-wilayah dengan konsentrasi kemiskinan yang tinggi, yang selanjutnya dapat membantu dalam perencanaan dan implementasi kebijakan pemerintah. Penerapan SIG berbasis K-Means ini memungkinkan visualisasi data yang lebih baik dan membantu pengambil kebijakan dalam membuat keputusan yang lebih informasional.

Dari hasil penelitian di atas, dapat disimpulkan bahwa metode K-Means memiliki aplikasi yang luas dan efektif dalam berbagai domain, mulai dari pendidikan hingga kesejahteraan sosial. Meskipun demikian, tantangan seperti penentuan jumlah klaster yang optimal dan sensitivitas terhadap outlier masih perlu diatasi untuk meningkatkan akurasi dan reliabilitas hasil klasterisasi. Studi-studi ini menunjukkan bahwa dengan pendekatan yang tepat, K-Means dapat digunakan untuk memberikan wawasan yang berharga dan membantu dalam pengambilan keputusan yang lebih baik dan lebih informatif di berbagai bidang.

4. SIMPULAN

Penelitian ini telah meninjau berbagai aplikasi metode K-Means dalam klasterisasi data mining dan menunjukkan efektivitasnya di berbagai domain, termasuk pendidikan, kesejahteraan sosial, dan sistem informasi geografis. Meskipun metode K-Means memiliki keunggulan dalam kesederhanaan dan efisiensi, terdapat beberapa tantangan utama yang perlu diatasi, seperti penentuan jumlah klaster yang optimal dan sensitivitas terhadap outlier. Studi-studi terbaru telah menunjukkan berbagai solusi dan inovasi untuk mengatasi tantangan ini, seperti penggunaan metode K-Means++ dan penggabungan dengan algoritma lain untuk meningkatkan kinerja klasterisasi.

Penerapan metode K-Means dalam berbagai penelitian, seperti yang dilakukan oleh Zahroh (2024), Maryam (2024), dan Paridah (2024), menunjukkan bahwa algoritma ini dapat digunakan untuk mengelompokkan data dengan akurasi yang tinggi dan membantu dalam pengambilan keputusan yang lebih baik. Pengembangan

sistem informasi berbasis K-Means, seperti yang ditunjukkan oleh Rivaldi (2024), juga memperlihatkan bagaimana metode ini dapat digunakan untuk visualisasi data yang lebih baik dan perencanaan kebijakan yang lebih efektif.

Secara keseluruhan, meskipun masih ada tantangan yang perlu diatasi, metode K-Means tetap merupakan alat yang kuat dan serbaguna dalam klasterisasi data mining. Penelitian lebih lanjut diperlukan untuk mengatasi tantangan yang ada dan meningkatkan akurasi serta reliabilitas hasil klasterisasi. Dengan demikian, K-Means dapat terus memberikan kontribusi yang signifikan dalam berbagai aplikasi praktis di masa mendatang

PUSTAKA

Helia, A. (2024). Analisis pengelompokan daya tarik obyek wisata berdasarkan jenisnya menggunakan metode k-means pada data Pemprov Jabar. Jati (Jurnal Mahasiswa Teknik Informatika), 8(3), 2764-2772. https://doi.org/10.36040/jati.v8i3.8369

Hidayati, N., & Rahmah, S. A. (2022). Clasterization Of Zeeida Product Sales Using K-Means Method In Medan Distributors. *Jurnal Mantik*, 6(2), 1685-1692.

Maryam, S. (2024). Optimalisasi jumlah cluster data sekolah dasar (SD) menggunakan algoritma k-means clustering. Jati (Jurnal Mahasiswa Teknik Informatika), 7(6), 3640-3646. https://doi.org/10.36040/jati.v7i6.8246

Paridah, N. (2024). Klasterisasi penerima dana bantuan program keluarga harapan menggunakan metode k-means pada desa Gereba. Jati (Jurnal Mahasiswa Teknik Informatika), 8(1), 1036-1043. https://doi.org/10.36040/jati.v8i1.8873

Rivaldi, A. (2024). Sistem informasi geografis pemetaan sebaran keluarga miskin menggunakan metode k-means. Jati (Jurnal Mahasiswa Teknik Informatika), 7(4), 2436-2442. https://doi.org/10.36040/jati.v7i4.7541

Riyandoro, A. (2023). Implementasi data mining clustering k-means dalam menggolongkan beragam merek laptop. Jati (Jurnal Mahasiswa Teknik Informatika), 7(2), 1372-1377. https://doi.org/10.36040/jati.v7i2.6816

Samsudin, R. (2024). Optimalisasi stok barang melalui algoritma k-means clustering analisis untuk manajemen persediaan dalam konteks bisnis modern. Jati (Jurnal Mahasiswa Teknik Informatika), 8(3), 3572-3580. https://doi.org/10.36040/jati.v8i3.9742

Santana, S. (2024). Klastering kopi arabika menggunakan algoritma k-medoids. Jati (Jurnal Mahasiswa Teknik Informatika), 8(2), 2390-2398. https://doi.org/10.36040/jati.v8i2.9480

Sholekha, E. (2024). Analisis penjualan produk snack dan minuman menggunakan metode k-means pada dataset transaksi penjualan. Jati (Jurnal Mahasiswa Teknik Informatika), 8(3), 2533-2539. https://doi.org/10.36040/jati.v8i3.9310

Subur, M. (2024). Clustering penduduk miskin untuk penerima bantuan sosial menggunakan algoritma k-means. Jati (Jurnal Mahasiswa Teknik Informatika), 8(1), 789-795. https://doi.org/10.36040/jati.v8i1.8809

Zahroh, L. (2024). Klasterisasi data kegemaran membaca menggunakan algoritma k-means di SMA Al-Islam Cirebon. Jati (Jurnal Mahasiswa Teknik Informatika), 8(3), 2692-2698. https://doi.org/10.36040/jati.v8i3.9543