Jurnal Manajemen Informatika Jayakarta
Volume 7, Nomor 1, Februari 2023: halaman 52-56

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i1.1001

Analisis Data Mining Untuk Clustering DataFilm Dengan Menggunakan Algoritma K- Means

[1]Zacky Yaser Malik Gumiwang, [2] Ahmad Fahry Hamidy, [3]Eva Yulia Puspaningrum

1,2,3Informatika, Ilmu Komputer, UPN 'Veteran' Jawa Timur
Jl. Rungkut Madya No.1, Gn. Anyar, Kec. Gn. Anyar, Kota Surabaya, Jawa Timur 60294

e-mail: zackyyaser50@gmail.com

Received: 23 Desember 2022, **Revised:** 8 Januari 2023, **Accepted:** 10 Januari 2023

Abstrak

Pengelompokan data adalah teknik pengambilan data yang penting dengan banyak aplikasi dalam penambangan data. K-means adalah salah satu metode penambangan data yang paling terkenal, yang membagi kumpulan data menjadi kelompok-kelompok sampel, yang disebut kelipatan. Metode telah diusulkan untuk meningkatkan efisiensi algoritma K-Means. Standarisasi adalah kuncinya langkah preprocessing dalam data mining untuk membakukan fitur atau nilai atribut dari rentang dinamis yang berbeda dan terdapat di area spesifik. Pada artikel ini, kami menganalisis kinerja metode standardisasi algoritma K-means tradisional. Membandingkan hasil dataset film, ditemukan bahwa hasil yang diperoleh dengan metode elbow.

Kata kunci: *Clustering, Data Mining, Elbow Method, K-means*

Abstract

Data clustering is an important data capture technique with many applications in data mining. K-means is one of the most well-known data mining methods, which divides a data set into groups of samples, which are called multiples. Methods have been proposed to increase the efficiency of the K-Means algorithm. Standardization is the key preprocessing step in data mining to standardize features or attribute values from different dynamic ranges that are present in specific areas. In this article, we analyze the performance of the traditional K-means algorithm standardization method. Comparing the results of the film dataset, it was found that the results were obtained by the elbow standardization method.

Keywords: *Clustering, Data Mining, Elbow Method, K-means*

1 Pendahuluan (or Introduction)

Salah satu yang paling sederhana dan paling umum digunakan sebuah teknik yang bertujuan untuk menghasilkan pengelompokan melalui optimasi kriteria kelayakan bekerja dengan tepat ditentukan di seluruh dunia (desain keseluruhan) atau lokal (untuk beberapa model), adalah K teknologi buatan. Pengelompokan K-means adalah salah satu prediktor tertua dari n pengamatan dalam ruang d-dimensi (integer d) diberikan dan tugasnya adalah menentukan himpunan titik c untuk meminimalkan jarak kuadrat rata-rata untuk semua data yang menunjukkan pusat terdekat di mana setiap pengamatan berada terobsesi. Tidak ada algoritma waktu polinomial yang tepat dikenal untuk masalah ini. Masalahnya bisa berupa masalah pemrograman bilangan bulat tetapi selesaikan program integer dengan sejumlah besar variable memakan waktu, cluster sering dihitung menggunakan a cepat, Metode heuristik yang biasanya memberikan hasil yang baik (tapi belum tentu optimal).

Jurnal Manajemen Informatika Jayakarta
Volume 7, Nomor 1, Februari 2023: halaman 52-56
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v3i1.1001

Algoritma K-Means adalah metode di mana Pengelompokan membutuhkan lebih sedikit usaha. Pertama nomornya cluster c juga menentukan hub ini diterima. Objek acak apa pun dengan inisial centroid dapat diambil atau k objek pertama berturut-turut juga dapat berfungsi sebagai pusat peluncuran. Namun, jika ada beberapa fitur yang besar atau variable fungsi seperti itu memiliki efek yang kuat pada pengelompokan hasil dalam hal ini, standardisasi data akan menjadi tugas preprocessing yang penting dalam penskalaan atau control variabilitas data. Tujuan pengelompokan adalah untuk mencari tahu kesamaan dan desain dari kumpulan data besar dengan membagi data menjadi beberapa kelompok. Karena diasumsikan bahwa set data tidak berlabel, Pengelompokan sering terjadi dianggap sebagai pembelajaran tanpa pengawasan yang paling berharga.

Sebuah aplikasi utama dari ukuran geometris (jarak) ke fitur yang memiliki rentang besar secara implisit menetapkan upaya yang lebih besar dalam metrik dibandingkan ke aplikasi dengan fitur yang memiliki rentang lebih kecil. Selain itu, fitur harus tidak berdimensi karena nilai numerik dari rentang dimensi fitur bergantung pada unit pengukuran dan batas, Pemilihan unit pengukuran mungkin secara signifikan mengubah hasil pengelompokan. Oleh karena itu, seseorang tidak boleh menggunakan ukuran jarak seperti jarak Euclidean tanpa memiliki normalisasi set data.

Preprocessing sebenarnya penting sebelum menggunakan algoritme eksplorasi data apa pun untuk meningkatkan kinerja hasil. Normalisasi dari dataset adalah salah satu proses preprocessing dalam data eksplorasi, di mana data atribut diskalakan untuk jatuh dalam rentang tertentu kecil. Normalisasi sebelumnya pengelompokan secara khusus diperlukan untuk metrik jarak, seperti jarak Euclidian yang peka terhadap variasi dalam besaran atau skala dari atribut. Dalam aplikasi sebenarnya, karena variasi dalam pemilihan nilai atribut, satu atribut mungkin mengalahkan yang lain. Normalisasi mencegah melebihi fitur memiliki sejumlah besar atas fitur dengan angka yang lebih kecil. Tujuannya adalah untuk menyamakan dimensi atau besaran dan juga variabilitas fitur tersebut.

Teknik preprocessing data diterapkan pada data mentah untuk membuat data bersih, bebas noise dan konsisten. Normalisasi Data membakukan data mentah dengan mengubahnya menjadi rentang tertentu menggunakan transformasi linier yang bisa menghasilkan cluster berkualitas baik dan meningkatkan keakuratan algoritma pengelompokan.

2 Metode Penelitian (or Research Method)

2.1 Elbow Method:

Metode dan analisis ini digunakan untuk memilih jumlah cluster atau kelompok optimal. Konsekuensi algoritma deep-elbow disajikan menunjukkan jumlah kelompok dibentuk. Itu tergantung pada kerumunan kesalahan kuadrat. K adalah jumlah kelompok digunakan dalam algoritma K-MeansXi adalah dataset dan Ck adalah jumlah cluster di cluster c.

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in S_k} \left\| X_i - C_k \right\|^2$$

Gambar 1. Metode Elbow

2.2 K-means Clustering:

Diberikan satu set pengamatan (x1, x2, …, xn), di mana setiap observasi adalah d-dimensi vector nyata, pengelompokan K-means bertujuan untuk mempartisi n pengamatan ke dalam k himpunan (k≤n) S = {S1, S2, …, Sk} sehingga menjadi untuk meminimalkan Jumlah Kuadrat Dalam-Cluster.

Jurnal Manajemen Informatika Jayakarta
Volume 7, Nomor 1, Februari 2023: halaman 52-56

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i1.1001

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Gambar 2. K-means Clustering

2.3 Langkah – langkah analisis data

Fase penelitian mendalam penelitian ini berbunyi sebagai berikut:

1. Analisis deskriptif untuk mengetahuinya deskripsi data film
2. Masukkan jumlah cluster dalam penelitian Ini adalah jumlah yang terbentuk sebanyak 5 cluster
3. Lakukan analisis cluster dengan Metode K-Means Cluster cocok dengan langkah 1.
4. Alokasi anggota di setiap klaster
5. Jelaskan karakteristik cluster diperoleh pada langkah 2
6. Buat kesimpulan tentang topik tersebut analisis dilakukan

3 Hasil dan Pembahasan

Pada bagian ini, rincian hasil keseluruhan telah dibahas. Program lengkap menggunakan Jyupiter Noteboook dikembangkan untuk mencari solusi optimal. Sedikit percobaan telah dilakukan pada tiga prosedur standardisasi dan membandingkannya kinerja pada K-means clustering algoritma dengan dataset film yang memiliki 10 objek data dan 9 atribut seperti yang ditunjukkan pada Tabel 1. Untuk atribut 1 ke atribut 9 masing-masing digunakan untuk menguji kinerja metode standardisasi pada K-means clustering.

Tabel.1 Dataset film dengan 10 data dan 9 atribut

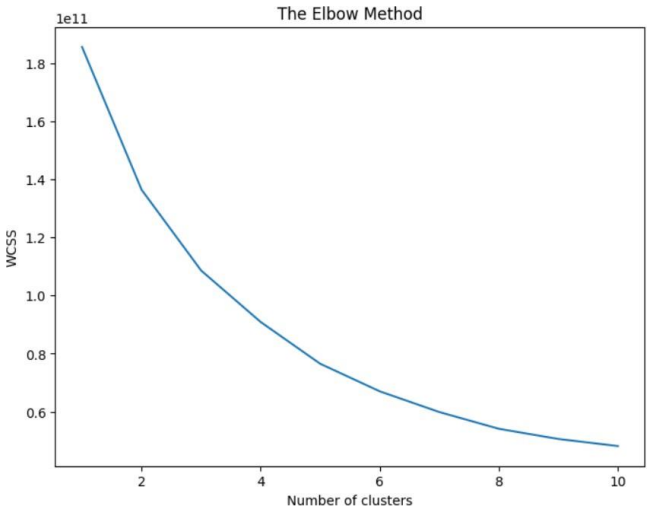
	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
0	Blood Red Sky	(2021)	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\n| \n Star...	21,062	121.0	NaN
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870	25.0	NaN
2	The Walking Dead	(2010–2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	885,805	44.0	NaN
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	414,849	23.0	NaN
4	Army of Thieves	(2021)	\nAction, Crime, Horror	NaN	\nA prequel, set before the events of Army of ...	\n Director:\nMatthias Schweighöfer\n| \n ...	NaN	NaN	NaN
5	Outer Banks	(2020–)	\nAction, Crime, Drama	7.6	\nA group of teenagers from the wrong side of ...	\n \n Stars:\nChase Stokes, \nMa...	25,858	50.0	NaN
6	The Last Letter from Your Lover	(2021)	\nDrama, Romance	6.8	\nA pair of interwoven stories set in the past...	\n Director:\nAugustine Frizzell\n| \n S...	5,283	110.0	NaN
7	Dexter	(2006–2013)	\nCrime, Drama, Mystery	8.6	\nBy day, mild-mannered Dexter is a blood-spat...	\n \n Stars:\nMichael C. Hall, \...	665,387	53.0	NaN
8	Never Have I Ever	(2020–)	\nComedy	7.9	\nThe complicated life of a modern-day first g...	\n \n Stars:\nMaitreyi Ramakrish...	34,530	30.0	NaN
9	Virgin River	(2019–)	\nDrama, Romance	7.4	\nSeeking a fresh start, nurse practitioner Me...	\n \n Stars:\nAlexandra Breckenr...	27,279	44.0	NaN

Jurnal Manajemen Informatika Jayakarta
Volume 7, Nomor 1, Februari 2023: halaman 52-56
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v3i1.1001

3.1 Analisis K-means cluster indikator kualitas film

Sebelum analisis K-Means Cluster, diperlukan pertama menentukan jumlah cluster. Dalam penelitian ini, jumlah cluster hingga 5 cluster ditentukan. dengan menggunakan algoritma k-means clustering hingga 5 cluster yang merupakan cluster optimal. Untuk melihat lebih jelas pada gambar berikut.



Gambar 6. Metode Elbow

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross	cluster
0	1139	338	60	6.1	1548	6057	1656	121.0	332	1
1	3719	339	112	5.0	7482	622	1110	25.0	332	0
2	6022	222	387	8.2	6371	305	3938	44.0	332	0
3	4621	255	123	9.2	2222	1476	2777	23.0	332	1
4	836	338	31	NaN	1108	5543	4129	NaN	332	1

Tabel 2. Indikator film pada setiap cluster

Dari Tabel 2 di atas terlihat tipe cluster 3, jumlah film sebanyak 4621 film dengan rating 9,2, dan jumlah vote sebanyak 2777 orang. Ciri-ciri tersebut menunjukkan bahwa Cluster 3 merupakan ciri film berkualitas baik dengan rating tinggi. Ketika rating film tinggi, kualitas film meningkat. Rating film cluster 0 rendah, sehingga karakteristik kualitas film Cluster 0 buruk. Dalam hal ini, Cluster 4 tidak memiliki klasifikasi film, sehingga karakteristik film Cluster 4 tidak jelas.

4 Kesimpulan

Berdasarkan hasil diskusi dari sini dapat disimpulkan bahwa kondisi indikator kualitas film masih memiliki ketidakseimbangan karena ada beberapa kategori film yang memilki kualitas buruk dan kategori film yang tidak jelas keterangannya. Dengan 5 cluster menggunakan metode clustering K-

Jurnal Manajemen Informatika Jayakarta
Volume 7, Nomor 1, Februari 2023: halaman 52-56
P-ISSN: 2746-5985
e-ISSN: 2797-0930
DOI: https://doi.org/10.52362/jmijayakarta.v3i1.1001

means, diperoleh cluster 0 dengan kualitas film yang buruk, cluster 1 dengan kualitas film sangat buruk, Cluster 2 dengan kualitas film baik, Cluster 3 dengan kualitas film sangat baik sedangkan Cluster 4 tidak jelas keterangannya.

Ucapan Terima Kasih

Ucapan terimakasih kami haturkan kepada program studi informatika UPN "Veteran" Jawa Timur atas terlaksananya penelitian kolaboratif dosen dan mahasiswa..

Daftar Pustaka

[1] Widodo. *Psikologi Belajar*. Jakarta:Rineka Cipta.2013.

[2] Prasetyo, Eko, *DATA MINING - Konsep dan Aplikasi Menggunakan MATLAB*, Nikodemus, Ed. Yogyakarta,Indonesia: Penerbit ANDI, 2012.

[3] Teguh Wibowo, *Penerapan Data Mining Pemilihan Siswa Kelas Unggulan dengan Metode K-Means Clustering di SMP N 02 Tasikmadu*, Program Studi Strata I pada Jurusan Informatika Fakultas Komunikasi dan Informatika. 2018.

[4] Himmah, Nofrida Rif'atul. *Implementasi Algoritma K-Means Untuk Pengelompokan Siswa Berdasarkan Nilai Akademik (Studi Kasus Mtsn Gresik).* Undergraduate Thesis, Universitas Muhammadiyah Gresik. 2019.

[5] Aniek Suryanti Kusuma. *Jurnal Sistem Informasi dan Komputer Terapan Indonesia (JSIKTI) Vol.1 (3).* SistemInformasi Akademik Serta Penentuan Kelas Unggulan Dengan Algoritama K-Means di SMP Negeri 3 Ubud, 2 Program Studi Teknik Informatika, STMIK STIKOM, Bali, Indonesia. 2019.

[6] Agusta Y. K-Means-Penerapan, Permasalahan dan Metode Terkait. Denpasar, Bali: *Jurnal Sistem dan Informatika (Februari 2007) Vol. 3: 47-60;* 2007.