

AKURASI K-MEANS DENGAN MENGGUNAKAN CLUSTER DAN TITIK GRID TERBAIK PADA PEMETAAN GRID INTERATIF K-MEANS

¹Johanes Terang Kita Perangin Angin✉, ²Ari Rizkita, ¹Robet, ¹Octara Pribadi

¹STMIK TIME, Medan, Indonesia

²Universitas Efarina, Pematangsiantar, Indonesia

Email: timejohanes@gmail.com

DOI: <https://doi.org/10.46880/jmika.Vol9No1.pp127-129>

ABSTRACT

Traditional K-Means face 2 (two) main problems, namely: Determination of Initial Centroid and poor initial cluster. Determining the initial centroid using random numbers is one of the main problems in classical K-Means which results in low accuracy and long computation time. Likewise, determining the good centroid of each cluster without being accompanied by a process of paying attention to the performance of each cluster can also cause the accuracy value obtained is not good. This study will contribute to how the performance obtained by determining a good initial centroid is combined with the use of a good cluster. Determination of a good initial centroid is done by using the K-Means Grid Mapping which divides the determination of the centroid into several Grid Points. The result of this research is a combination of Iterative K-Means with Grid Mapping K-Means to become Iterative Grid Mapping K-Means which will get a good initial centroid and also a good cluster shown in the table of iris and abalone, comparison of the variables in the iris and abalone affecting the best cluster as a result.

Keyword: K-Means, Centroid, Grid Mapping K-Means, Iterative K-Means, Iterative Grid Mapping K-Means.

ABSTRAK

K-Means tradisional menghadapi 2 (dua) permasalahan utama, yaitu: Penentuan Centroid Awal dan Cluster Awal yang kurang baik. Penentuan centroid awal dengan menggunakan bilangan acak merupakan salah satu permasalahan utama pada K-Means klasik yang mengakibatkan akurasi rendah dan waktu komputasi yang lama. Begitu pula penentuan centroid yang baik dari setiap cluster tanpa disertai proses memperhatikan performa dari setiap cluster juga dapat menyebabkan nilai akurasi yang diperoleh kurang baik. Penelitian ini akan memberikan kontribusi tentang bagaimana performa yang diperoleh dari penentuan centroid awal yang baik dikombinasikan dengan penggunaan cluster yang baik. Penentuan centroid awal yang baik dilakukan dengan menggunakan K-Means Grid Mapping yang membagi penentuan centroid ke dalam beberapa Grid Point. Hasil dari penelitian ini adalah gabungan antara Iterative K-Means dengan Grid Mapping K-Means sehingga menjadi Iterative Grid Mapping K-Means yang akan mendapatkan centroid awal yang baik dan juga cluster yang baik seperti yang ditunjukkan pada tabel iris dan abalone, perbandingan variabel pada iris dan abalone mempengaruhi cluster yang paling baik sebagai hasilnya.

Kata Kunci: K-Means, Centroid, Grid Mapping K-Means, Iterative K-Means, Iterative Grid Mapping K-Means.

PENDAHULUAN

Clustering merupakan proses pencarian dan pengelompokan data yang pada dasarnya mempunyai tingkat kesamaan karakteristik (similarity) antara satu data dengan data lainnya. Clustering banyak dimanfaatkan dalam berbagai bidang seperti analisis jaringan sosial, rekayasa perangkat lunak, dan pendeteksian kejahatan. Ada beberapa algoritma clustering yang dapat digunakan, namun algoritma K-Means dan Fuzzy C-Means yang umum digunakan karena cukup sederhana. Clustering merupakan salah satu pengelompokan data mining. Pada algoritma K-

Means, penetapan jumlah cluster dan penetapan centroid cukup sulit dilakukan. Penetapan dari jumlah cluster dan penentuan centroid secara langsung mempengaruhi kualitas proses clustering (Sajidha et al., 2021).

Penentuan centroid awal yang baik akan mampu meningkatkan kemampuan K-Means dalam mencapai kinerja yang diinginkan. Akan tetapi, meskipun penentuan centroid awal yang dilakukan oleh beberapa metode telah berhasil menentukan centroid awal dengan baik, namun akurasi yang diberikan bisa jadi kurang baik apabila jumlah cluster yang digunakan terlalu banyak (Celebi et al., 2013).

Penelitian sejumlah pakar telah mengaitkan penentuan cluster dengan kinerja K-Means dan sebaliknya penentuan centroid awal dengan kinerja K-Means. Penelitian yang dilakukan oleh (Ismkhan, 2018) telah menghasilkan metode penentuan cluster yang baik untuk K-Means yang terdiri dari beberapa cluster, sehingga untuk cluster yang hasilnya buruk akan dihilangkan dan untuk cluster yang hasilnya baik akan dibagi menjadi beberapa cluster yang lain. Kelemahan metode ini adalah hanya menggunakan satu titik pusat untuk setiap cluster. Kelebihannya adalah metode ini telah mempertimbangkan penggunaan cluster yang terbaik.

KAJIAN LITERATUR

Penelitian yang dilakukan oleh (Zhu & Ma, 2018) telah mengemukakan bahwa dalam menentukan centroid awal dilakukan dengan cara membagi penentuan centroid tersebut ke dalam beberapa Grid Point, maksud Grid Point disini adalah setiap cluster dipecah lagi menjadi beberapa titik centroid. Kelemahannya adalah, metode ini masih menggunakan cluster-cluster dengan hasil yang kurang baik untuk dilanjutkan ke tahap selanjutnya. Kelebihannya adalah metode ini sudah menggunakan beberapa titik centroid untuk setiap clusternya.

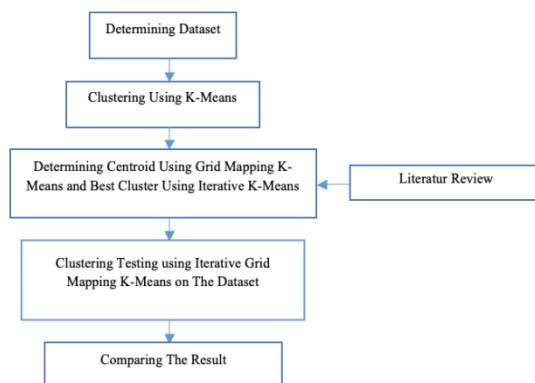
Kemudian penelitian penentuan cluster yang baik seperti yang dikemukakan oleh (Ismkhan, 2018) dengan mengeliminasi cluster yang buruk dikombinasikan dengan penentuan centroid awal yang berasal dari centroid setiap titik grid setiap cluster seperti yang dikemukakan oleh (Zhu & Ma, 2018). Ide dasarnya adalah menentukan centroid awal dari centroid setiap titik grid setiap cluster yang digunakan dalam penentuan centroid awal, namun titik grid setiap cluster yang kinerjanya buruk tidak akan dimasukkan dalam tahap selanjutnya.

K-Means adalah salah satu algoritma clustering yang bekerja dengan cara mengelompokkan sejumlah instance ke dalam suatu kelas berdasarkan kedekatan setiap atribut terhadap centroid masing-masing kelas. Setiap instance dikelompokkan ke dalam centroid dengan jarak terdekat (Esnault et al., n.d.). Penentuan centroid awal menggunakan bilangan acak merupakan salah satu permasalahan utama dalam K-Means klasik yang mengakibatkan rendahnya akurasi dan waktu komputasi yang lama (Ma & Chow, 2004). Sehingga hasil dari K-Means sangat bergantung pada inisialisasi yang baik. Inisialisasi yang kurang baik dapat berdampak pada hasil K-Means menjadi stuck pada minimum lokal (Angin et al., 2020). Sedangkan (Xu et al., 2017) mengusulkan konsep Hierarchical K-Means

yang menggunakan Struktur Hirarkis dataset dan dapat mengurangi waktu komputasi.

METODE PENELITIAN

Tahapan penelitian dapat dilihat pada Gambar 1.

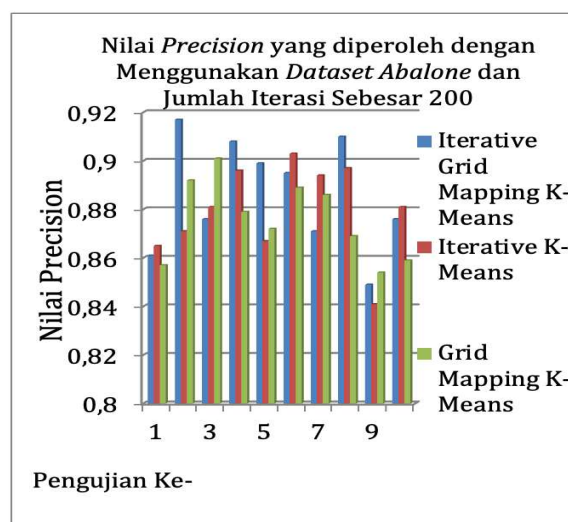


Gambar 1. Kerangka Penelitian

Berdasarkan Gambar 1, dapat diketahui bahwa penelitian ini akan melakukan penentuan cluster yang baik dengan mengeliminasi cluster yang buruk yang dikombinasikan dengan penentuan centroid awal yang berasal dari centroid setiap titik grid setiap cluster. Ide dasarnya adalah menentukan centroid awal dari centroid setiap titik grid setiap cluster yang digunakan dalam penentuan centroid awal, namun titik grid setiap cluster dengan kinerja yang buruk tidak akan dimasukkan dalam tahap selanjutnya.

HASIL DAN PEMBAHASAN

Dataset yang dipakai pada penelitian ini bersumber dari KEEL Repository. Dataset yang dipakai pada penelitian ini dapat dilihat pada Gambar 2 berikut:



Gambar 2. Grafik Deskripsi Dataset

Hasil Akurasi K-Means dengan Menggunakan Cluster dan Titik Grid Terbaik pada Iterative Grid Mapping K-Means dengan menggunakan dataset yaitu:

@relation abalone
@attribute Sepal_Length real [4.3, 7.9]
@attribute Sepal_Width real [2.0, 4.4]
@attribute Petal_Length real [1.0, 6.9]
@attribute Petal_Width real [0.1, 2.5]
@attribute Class {positive, negative}
@inputs Sepal_Length, Sepal_Width, Petal_Length, Petal_Width
@outputs Class
@data

Tabel 2. Deskripsi Abalone

Atts 1	Atts 3	Atts 5	Atts 7	Atts 9	Result
0.4	0.305	0.1	0.3415	0.0865	7
0.635	0.5	0.15	1.376	0.31	10
0.37	0.27	0.09	0.1855	0.065	7
0.68	0.54	0.155	1.534	0.384	10
0.375	0.285	0.09	0.2545	0.0675	6
0.58	0.475	0.155	0.974	0.285	10
0.525	0.38	0.14	0.6065	0.21	14
0.63	0.495	0.19	1.1655	0.1625	10

Perbandingan kelima variabel pada abalon mempengaruhi klaster terbaik sebagai hasilnya.

KESIMPULAN

Dalam penelitian ini, kontribusi yang diberikan adalah sebagai berikut, hasil pengujian menunjukkan bahwa Iterative Grid Mapping K-Means memberikan hasil yang lebih baik dibandingkan dengan Iterative K-Means dan Iterative Grid Mapping K-Means. Peningkatan jumlah iterasi dapat meningkatkan kinerja masing-masing metode, baik dari segi nilai Average Precision maupun Best Precision. Jumlah instance dan atribut dalam dataset dapat mempengaruhi kinerja masing-masing metode yang menunjukkan bahwa kinerja masing-masing metode lebih baik pada dataset Iris dibandingkan dengan dataset Abalone.

DISEMINASI

Artikel ini telah diseminasikan pada Seminar Nasional Teknologi Informasi dan Komunikasi (SEMNASITIK) APTIKOM Tahun 2024 yang diselenggarakan oleh Universitas Methodist Indonesia pada tanggal 24-26 Oktober 2024.

DAFTAR PUSTAKA

Angin, J. T. K. P., Johan, Sukiman, Sugianto, Simarmata, B. R., & Suharjito. (2020). Face Recognition Application with the Complete Kernel Fisher Discriminant (CKFD) Method.

MECnIT 2020 - International Conference on Mechanical, Electronics, Computer, and Industrial Technology, 359–364.

<https://doi.org/10.1109/MECNIT48290.2020.9166682>

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1), 200–210. <https://doi.org/10.1016/j.eswa.2012.07.021>

Esnault, C., Rollot, M., Guilmin, P., & Zucker, J.-D. (n.d.). *Qluster: An easy-to-implement generic workflow for robust clustering of health data*.

Ismkhan, H. (2018). I-k-means+: An iterative clustering algorithm based on an enhanced version of the k-means. *Pattern Recognition*, 79, 402–413.

<https://doi.org/10.1016/J.PATCOG.2018.02.015>

Ma, E. W. M., & Chow, T. W. S. (2004). A new shifting grid clustering algorithm. *Pattern Recognition*, 37(3), 503–514.

<https://doi.org/10.1016/J.PATCOG.2003.08.014>

Sajidha, S. A., Chodnekar, S. P., & Desikan, K. (2021). Initial seed selection for K-modes clustering – A distance and density-based approach. *Journal of King Saud University - Computer and Information Sciences*, 33(6), 693–701.

<https://doi.org/10.1016/j.jksuci.2018.04.013>

Xu, T. S., Chiang, H. D., Liu, G. Y., & Tan, C. W. (2017). Hierarchical K-means Method for Clustering Large-Scale Advanced Metering Infrastructure Data. *IEEE Transactions on Power Delivery*, 32(2), 609–616.

<https://doi.org/10.1109/TPWRD.2015.2479941>

Zhu, E., & Ma, R. (2018). An effective partitional clustering algorithm based on new clustering validity index. *Applied Soft Computing Journal*, 71, 608–621.

<https://doi.org/10.1016/J.ASOC.2018.07.026>