Jurnal Informasi dan Teknologi

https://jidt.org/jidt

2023 Vol. 5 No. 1 Hal: 88-91 e-ISSN: 2714-9730

Perbandingan Kinerja Algoritma Kmeans dengan Kmeans Median pada Deteksi Kanker Payudara

Siti Mutrofin[1], Toni Wicaksono[2]✉, Ali Murtadho[3]

[1]Program Studi Sistem dan Teknologi Informasi, Universitas 17 Agustus 1945 Surabaya, Indonesia
[2,3]Program Stud Sistem Informasi, Universitas Pesantren Tinggi Darul Ulum, Indonesia

toniwicaksono2411@gmail.com

Abstrak

Clustering adalah salah satu tugas dari *Data Mining* berbasis *unsupervised learning*. *Clustering* dapat digunakan untuk permasalahan di berbagai bidang, seperti pendidikan, kesehatan, ekonomi, pertanian, hiburan, olah raga, dll. Salah satu algoritma *clustering* yang sederhana dan umum digunakan pada tipe data numerik adalah Kmeans. Terlepas dari segala kelebihan Kmeans, Kmeans memiliki permasalahan berupa pemilihan pusat awal *cluster* atau *centroid* yang dipilih secara acak (*random*). Karena hasil akhir *cluster* dari Kmeans sangat sensitif pada pemilihan awal *cluster* yang dapat menyebabkan konvergensi yang tidak optimal. Pada penelitian ini dilakukan perbandingan antara Kmeans dengan Kmeans Median dalam menentukan *centroid* awal yang bertujuan untuk mengetahui kinerja kedua algoritma yang paling optimal. Pada Kmeans Median, pemilihan *centroid* awal dipilih dengan menggunakan nilai median yang diambil dari kelompok data dan akan dicari nilai mediannya. *Dataset* yang digunakan dalam penelitian ini adalah *Breast Cancer Coimbra* yang dapat diakses di UCI *Machine Learning Repository*. Perhitungan jarak untuk kedua algoritma menggunakan *euclidean distance*. Pengujian pada penelitian ini untuk mengetahui kinerja kedua algoritma menggunakan *Davies-Bouldin Index* (DBI). Pengujian dilakukan empat kali pada masing-masing algoritma. Dari keempat pengujian tersebut, Kmeans mendapatkan DBI terbaik 0,47 dan terburuk 0,4796. Sedangkan pada Kmeans Median memiliki DBI 0,47 pada keempat pengujiannya. Kinerja Kmeans berdasarkan iterasi antara 3, 5, dan 6. Sedangkan Kmeans Median iterasinya konsisten hanya 4. Berdasarkan komputasinya, Kmeans lebih unggul karena algoritmanya lebih sederhana, hal itu terlihat dari waktu yang dibutuhkan lebih sedikit dibandingkan Kmeans Median.

Kata Kunci: *Centroid, Cluster, Davies-Bouldin Index*, Kmeans, Kmeans Median.

1. Pendahuluan

Kmeans adalah salah satu algoritma pengelompokan atau *clustering*. Kmeans adalah dalam mengelompokkan data berdasarkan *similarity*[1]. Kmeans memiliki kelebihan, diantaranya: 1) Sederhana dan konvergensi yang cepat [2], [3]; 2) Banyak digunakan dalam *dataset* besar karena efisien secara komputasi [4]; 3) Mudah diimplementasikan dan hasilnya yang mudah diinterpretasi [3], dll. Kmeans juga memiliki kekurangan, salah satu diantaranya adalah peka terhadap inisialisasi yang berarti bahwa hasilnya sangat bergantung pada pusat *cluster* (*centroid*) awal yang dipilih secara acak (*random*)[5], [6].

Beberapa peneliti telah mencoba melakukan perbaikan terhadap kelemahan algoritma Kmeans yang berfokus pada penentuan *centroid* awal, diantaranya: 1) Median [7]; 2) Kmeans++ [8]; 3) *Principal component analysis* (PCA)[9]; 4) *Linear time-complexity, loglinear time-complexity, quadratic-complexity, other initialization*, dan *linear vs. superlinear initialization* [10], dll. Penelitian terkait Kmeans Median masih jarang dikaji, sehingga pada penelitian ini akan dilakukan perbandingan kinerja antara Kmeans dengan Kmeans berbasis median dari sisi penentuan *centroid* awal.

2. Metode Penelitian

2.1. *Dataset*

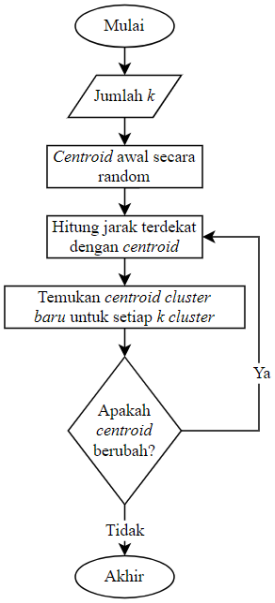
Penelitian ini akan menggunakan *dataset Breast Cancer Coimbra* yang dapat diakses di UCI *Machine Learning Repository*. Atribut dari *dataset Breast Cancer Coimbra* berjumlah 9, diantaranya: *Age, BMI, glucose, insulin, HOMA, leptin, adiponectin*, dan *resistin*. Atribut memiliki tipe data integer, jumlah data 116, memiliki 2 label berupa binominal (*healthy controls* dan *patients*), dan tidak ada *missing values* [11].

Diterima: 27-02-2023 | Revisi: 28-02-2023 | Diterbitkan: 28-02-2023 | doi: 10.37034/jidt.v5i1.274

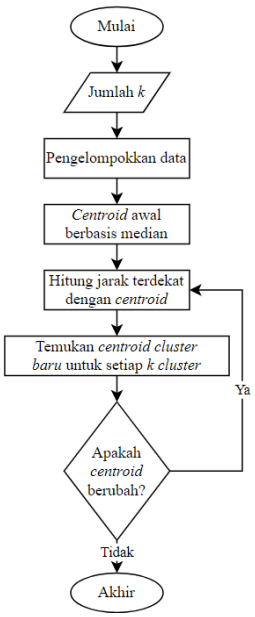
2.2. Kmeans

Algoritma Kmeans menurut MacQueen [12] adalah algoritma yang mudah dan efektif untuk menemukan *cluster* dalam data. Langkah-langkah Kmeans diilustrasikan seperti pada Gambar 1 dengan penjelasan sebagai berikut [12]:

a. Tentukan jumlah *k* dari kumpulan data yang akan dikelompokkan.
b. Tetapkan *k* data secara *random* untuk menjadi lokasi pusat *cluster* awal.
c. Untuk setiap data, temukan pusat *cluster* terdekat. Setiap pusat *cluster* "memiliki" subset dari data, sehingga mewakili partisi dari kumpulan data. Oleh karena, itu kami memiliki *k cluster*, $C_1, C_2, \dots C_k$.
d. Untuk setiap *k cluster*, temukan *centroid cluster*, dan perbarui lokasi setiap pusat *cluster* ke nilai *centroid* yang baru.
e. Ulangi langkah 3 sampai 5 sampai konvergensi.



Gambar 1. *Flowchart* Kmeans [12]



Gambar 2. *Flowchart* Kmeans berbasis median [7]

2.3. Kmeans Median

Kmeans berbasis median memiliki cara kerja yang sama Kmeans. Hanya saja, Kmeans median pemilihan awal pusat *cluster* atau *centroid* dengan cara mengambil dari *data set* menggunakan rumus median. Penentuan nilai *centroid* awal sudah dilakukan pegelompokan data (Persamaan 1) terlebih dahulu dengan membagi data ke dalam beberapa kelompok sesuai dengan jumlah *cluster* yang sudah ditentukan [7]. Selanjutnya setelah dibagi dalam

beberapa kelompok, maka dicari nilai median (Persamaan 2) untuk ditentukan sebagai *centroid* awal. Ilustrasi Kmeans berbasis median ditunjukkan pada Gambar 2.

$$\text{penglompokan data} = \frac{n}{k} \quad (1)$$

$$\text{median} = \frac{\text{bilangan terkecil+bilangan terbesar}}{2} \quad (2)$$

di mana *n* adalah banyaknya data dan *k* adalah jumlah cluster.

2.4. Davies Bouldin Index (DBI)

David L. Davies dan Donald W. Bouldin memperkenalkan *Davies-Bouldin Index* (DBI) yang digunakan untuk mengevaluasi *cluster* [13]. Evaluasi menggunakan DBI memiliki skema evaluasi internal *cluster*, di mana baik atau tidaknya hasil *cluster* dilihat dari kuantitas dan kedekatan antar data hasil *cluster*. Kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik *centroid* dari *cluster* yang diikuti. Separasi didasarkan pada jarak antar titik *centroid* terhadap klasternya[3].

3. Hasil dan Pembahasan

Berdasarkan hasil uji coba sebanyak empat kali untuk masing-masing algoritma seperti yang disajikan pada Tabel 1. Kinerja Kmeans memperoleh DBI cenderung konstan, tetapi pada pengujian kedua membutuhkan iterasi lebih banyak dan DBI lebih tinggi, tetapi komputasi lebih rendah. Kmeans Median untuk DBI dan iterasinya konstan, sedangkan komputasinya mengalami fluktuasi. Ketika pengujian kedua membutuhkan komputasi lebih rendah, sedangkan pengujian pertama membutuhkan komputasi lebih tinggi. Komputasi yang dibutuhkan Kmeans Median cenderung lebih tinggi karena adanya formula pengelompokan data. Sedangkan untuk kinerja Kmeans dalam melakukan pengelompokan berdasarkan DBI tidak terlalu ada banyak perbedaan. Semakin rendah nilai DBI maka kinerja algoritma dalam menghasilkan *cluster* semakin baik [14].

Tabel 1. Hasil pengujian

Algoritma	Pengujian	Iterasi	Waktu (detik)	DBI
Kmeans	1	5	1,0220	0,4700
	2	6	0,7279	**0,4796**
	3	3	0,7547	0,4700
	4	3	0,7447	0,4700
Kmeans Median	1	4	**1,4747**	0,4700
	2	4	**0,1283**	0,4700
	3	4	1,3973	0,4700
	4	4	1,2045	0,4700

Pada penelitian selanjutnya perlu dilakukan skenario uji coba yang lebih kompleks, misalkan data yang lebih bervariasi dari sisi jumlah datanya, jumlah atributnya, permasalahan *missing value*[15], *imbalanced data*[16], dll. Termasuk beberapa teknik evaluasi perlu dicoba, misalkan *Silhouette Coefficient*[14], [17], *Calinski-Harabasz Index*[17], [18], *Rand Index*[19], [20], dll.

4. Kesimpulan

Penelitian ini bertujuan untuk melihat kinerja algoritma Kmeans dengan Kmeans Median berdasarkan penentuan *centroid* awal. Berdasarkan hasil pengujian didapatkan Kmeans berpotensi mendapatkan nilai DBI tinggi yang artinya kurang bagus hasil pengelompokannya, tetapi komputasi cenderung rendah. Sedangkan Kmeans median berpotensi memiliki kinerja bagus dalam pengelompokan tetapi tidak terlalu banyak perbedaan. Pada penelitian di masa mendatang disarankan agar lebih menggali lebih dalam terkait Kmeans Median untuk mengetahui secara detail kelebihan dan kekurangannya dibandingkan dengan Kmeans atau Kmeans turunannya yang lain.

Daftar Rujukan

[1] R. Kurniawan, S. Defit, and S. Sumijan, "Prediksi Tingkat Kerugian Peternak Akibat Penyakit pada Sapi Menggunakan Algoritma K-Means Clustering," *J. Inf. dan Teknol.*, vol. 3, no. 1, pp. 29–35, 2021, doi: 10.37034/jidt.v3i1.87.

[2] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J-Multidisciplinary Sci. J.*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[3] S. Nawrin, M. R. Rahman, and S. Akhter, "Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 264–272, 2017, doi: 10.14569/ijacsa.2017.080337.

[4] T.-S. Xu, H.-D. Chiang, G.-Y. Liu, and C.-W. Tan, "Hierarchical K-means Method for Clustering Large-Scale Advanced Metering Infrastructure Data," *IEEE Trans. Power Deliv.*, vol. 32, no. 2, pp. 609–616, 2017, doi: 10.1109/TPWRD.2015.2479941.

[5] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Syst. Appl.*, vol. 67, pp. 12–18, 2017, doi: 10.1016/j.eswa.2016.09.025.
[6] X. Tong, M. Fanrong, and W. Zhixiao, "K-means for optimizing the initial cluster centers," *Comput. Eng. Des.*, vol. 32, no. 8, pp. 2721–2723, 2011.
[7] H. Sopian, P. Purwanto, and M. A. Soeleman, "Pemilihan Pusat Klaster Awal Pada Metode K-Means Berbasis Median," *J. Intake J. Penelit. Ilmu Tek. dan Terap.*, vol. 7, no. 2, pp. 93–104, 2016.
[8] Z. Min and D. Kai-Fei, "Improved Research to K-means Initial Cluster Centers," in *Proceedings - 2015 9th International Conference on Frontier of Computer Science and Technology, FCST 2015*, 2015, pp. 349–353. doi: 10.1109/FCST.2015.61.
[9] S. A. M. Anaraki, A. Haeri, and F. Moslehi, "A hybrid reciprocal model of PCA and K-means with an innovative approach of considering sub-datasets for the improvement of K-means initialization and step-by-step labeling to create clusters with high interpretability," *Pattern Anal. Appl.*, vol. 24, pp. 1387–1402, 2021, doi: 10.1007/s10044-021-00977-x.
[10] M. Z. Usman and T. Oktiarso, "Implementasi Algoritma Greedy Untuk Menyelesaikan Travelling Salesman Problem di Distributor PT. Z," *J. Integr. Syst.*, vol. 1, no. 2, pp. 216–229, Mar. 2018, doi: 10.28932/JIS.V1I2.1049.
[11] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, pp. 1–8, 2018, doi: 10.1186/s12885-017-3877-1.
[12] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. 2014. doi: 10.1002/0471687545.
[13] S. Rustam, "Analisa Clustering Phising dengan K-Means dalam Meningkatkan Keamanan Komputer," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 175–181, 2018, doi: 10.33096/ilkom.v10i2.309.175-181.
[14] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.
[15] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, pp. 1785–1792, 2022, doi: 10.1016/j.jksuci.2019.12.011.
[16] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.
[17] A. F. Khairati, A. . Adlina, G. . Hertono, and B. . Handari, "Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA," in *PRISMA, Prosiding Seminar Nasional Matematika*, 2019, vol. 2, pp. 161–170. [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/article/view/28906
[18] S. Lukasik, P. A. Kowalski, M. Charytanowicz, and P. Kulczycki, "Clustering using flower pollination algorithm and Calinski-Harabasz index," in *2016 IEEE Congress on Evolutionary Computation, CEC 2016*, 2016, pp. 2724–2728. doi: 10.1109/CEC.2016.7744132.
[19] S. Krey, U. Ligges, and F. Leisch, "Music and timbre segmentation by recursive constrained K-means clustering," *Comput. Stat.*, vol. 29, pp. 37–50, 2014, doi: 10.1007/s00180-012-0358-5.
[20] K. Laskhmaiah, S. M. Krishna, and B. E. Reddy, "An optimized k-means with density and distance-based clustering algorithm for multidimensional spatial databases," *Int. J. Comput. Netw. Inf. Secur.*, vol. 13, no. 6, pp. 70–82, 2021, doi: 10.5815/ijcnis.2021.06.06.