Evaluasi K-Means dan K-Medoids pada Dataset Kecil

Rima Dias Ramadhani*, Dwi Januarita AK
Program Studi Informatika
Sekolah Tinggi Teknologi Telematika Telkom
Jl. DI Panjaitan 128, Purwokerto
rimadias@st3telkom.ac.id*, dwijanuarita@st3telkom.ac.id

***Abstrak*— Klastering merupakan pengelompokan beberapa *records*, observasi, atau kasus lain dalam sebuah kelas yang dibedakan berdasarkan objek yang saling serupa. K-Means merupakan algoritma yang dapat melakukan pengelompokan berdasarkan jarak terdekat antara data dengan pusat data pada klaster. K-Medoids merupakan metode yang digunakan untuk mengelompokkan sekumpulan objek menjadi sebuah kelas. Pada penelitian ini akan membahas mengenai komparasi hasil evaluasi pada K-Means dan K-Medoids menggunakan *dataset* yang berukuran kecil yaitu *dataset* Iris dan Wine. Davies Bouldin Index digunakan pada penelitian ini untuk mengukur kesamaan dari ukuran klaster berdasarkan penyebaran data pada klaster dan ketidaksamaan ukuran klaster. Berdasarkan hasil eksperimen, K-Means menunjukkan hasil evaluasi yang lebih baik dibandingkan dengan K-Medoids dalam menangani *dataset* dengan ukuran kecil. Hal ini ditunjukkan dengan hasil evaluasi pada *dataset* Iris dengan menggunakan K-Means yaitu sebesar 0.662 dan pada *dataset* Wine menunjukkan hasil evaluasi sebesar 0.534.**

Kata kunci—K-Means; K-Medoids; davies bouldin index; dataset; klaster.

I. PENDAHULUAN

Menurut Velmurugan[1] *data mining* merupakan proses untuk menemukan pola maupun korelasi yang mempunyai makna dengan sebuah pengetahuan baru dengan cara memilah sejumlah data yang besar dengan menggunakan pengenalan pola, statistik dan matematis. Terdapat salah satu metode yang digunakan dalam *data mining* yaitu klastering. Menurut Anil K. Jain[2] klastering merupakan pengelompokan beberapa *records*, observasi, atau kasus lain dalam sebuah kelas yang dibedakan berdasarkan objek yang saling serupa.

Terdapat beberapa algoritma klastering di antaranya adalah *K*-Means dan *K*-Medoids.Anil K. Jain[2] mendefinisikan *K*-Means sebagai algoritma yang dapat melakukan pengelompokan berdasarkan jarak terdekat antara data dengan pusat klaster. Menurut beberapa penelitian[3],[4], *K*-Means dapat mengelompokkan objek ke masing-masing klaster dengan cepat dan sederhana, sedangkan K-Medoids menurut Jiawei Han dan Micheline Kamber[5] merupakan metode yang digunakan untuk mengelompokkan sekumpulan n objek menjadi sejumlah k klaster. Lebih lanjut pada penelitiannya, *K-Medoids* memiliki kehandalan dalam mengatasi *noise*.

Dengan adanya pendekatan dan kelebihan yang berbeda dari beberapa algoritma sering kali menyebabkan hasil yang berbeda. Termasuk algoritma klastering yang sering digunakan oleh peneliti pada berbagai aplikasi di bidang yang berbeda-beda, sehingga, pada penelitian ini akan melakukan komparasi berdasarkan dua algoritma klastering yaitu *K*-Means dan *K*-Medoids untuk melakukan perbandingan tingkat evaluasi dengan melakukan uji eksperimen pada dua *dataset* yang berbeda.

Dataset yang digunakan adalah *dataset* yang berukuran kecil yaitu Iris dan Wine. *Dataset* Iris merupakan merupakan *database* paling dikenal dapat ditemukan dalam literatur *pattern recognition* yang dibuat berdasarkan penelitian dari R. A Fisher yang memiliki 4 atribut dan 150 data, sedangkan *dataset* Wine merupakan hasil analisis kimia dari anggur yang tumbuh di daerah yang sama di Italia, tetapi berasal dari tiga kultivar yang berbeda yang memiliki 13 atribut dan 178 data.

Diharapkan, dengan adanya hasil evaluasi kinerja dari K-Means dan K-Medoids dapat menjadi referensi bagi para peneliti dalam melakukan pengolahan data dengan *dataset* berukuran kecil. Makalah ini terdiri dari empat bagian yaitu yang pertama adalah pendahuluan, ke dua adalah metode, selanjutnya adalah hasil dan diskusi, dan yang terakhir adalah kesimpulan.

II. METODE

Klastering merupakan metode yang menggunakan teknik *unsupervised learning* di mana pada prosesnya tidak membutuhkan label pada *dataset* seperti pada *data training*. Hal tersebut dijelaskan pada penelitian Tiwari dan Singh[6]. Terdapat beberapa penelitian yang membahas mengenai evaluasi pada algoritma klastering. Velmurugan dalam penelitiannya[7],[8] melakukan penerapan dan analisis klastering dengan menggunakan K-Means dan K-Medoids berdasarkan *dataset private* dengan kesimpulan bahwa *K*-Means lebih efektif pada *dataset* yang berukuran kecil sedangkan K-Medoids memiliki performa yang lebih baik untuk *dataset* yang berukuran besar.

Menurut Tiwari dan Singh[6] dengan menggunakan *dataset* berukuran kecil hingga medium menunjukkan bahwa K-Means memiliki performa yang lebih cepat namun sensitif pada *noise*. Sedangkan K-Medoids memiliki waktu komputasi yang lama dan handal dalam mengatasi *noise*. Pada penelitian ini akan diterapkan dua *dataset* yang berukuran kecil yang selanjutnya akan dilakukan eksperimen dengan menggunakan K-Means dan K-Medoids.

A. *K-Means*

K-Means merupakan algoritma klastering yang melakukan pengelompokan objek berdasarkan jarak terdekat dengan pusat klaster ke kelompok yang memiliki kesamaan satu sama lain. Tahapan-tahapan dari algoritma K-Means[9] adalah:

Tahap 1: Menentukan jumlah partisi/klaster dari *dataset* yang akan dibentuk.

Tahap 2: Memilih objek dari masing-masing secara acak untuk menjadi lokasi pusat klaster awal/*centroid*

Tahap 3: Menentukan objek berdasarkan jarak terdekat dengan *centroid.*

Tahap 4: Menghitung kembali *centroid* masing-masing klaster yang terbentuk untuk memperbarui *centroid* baru.

Tahap 5: Ulangi langkah ke 3 sampai 5 hingga tidak ada data yang berpindah ke klaster lainnya.

B. *K-Medoids*

K-Medoids menggunakan objek representatif sebagai titik acuan, bukan mengambil nilai rata-rata dari objek dalam setiap klaster. Algoritma mengambil parameter *input k*, jumlah klaster yang akan dipartisi di antara satu *set* n objek. Tahapan dari K-medoids adalah*:*

Tahap 1: Pilih titik awal K. Titik ini adalah *medoid* yang berfungsi sebagai kandidat dan dimaksudkan untuk menjadi poin paling sentral dari kelompok tersebut.

Tahap 2: Kemudian setiap objek yang tersisa dikelompokkan dengan objek representatif yang paling mirip.

Tahap 3: Kemudian mengganti salah satu *medoid* dengan *medoid* lain yang secara kualitas lebih baik dan dilakukan secara *iterative*.

C. *Euclidean Distance*

Dalam mencari setiap objek yang saling memiliki kemiripan baik *K-Means* maupun *K-medoids* pada penelitian ini menggunakan pengukuran jarak menggunakan *Euclidean distance* yang ditunjukkan pada Persamaan 1:

$$d_{euclidean}(x,y) = \sqrt{\sum_i (xi - xy)^2} \quad (1)$$

dimana $x = x1, x2, \ldots, xm$ dan $y = y1, y2, \ldots, ym$ merupakan atribut m dari 2 *records.* Pada tahap ke empat, untuk menghitung kembali *centroid* baru misalnya data n berisi $(a_1, b_1, c_1), (a_2, b_2, c_2)$,…….. (a_k, b_k, c_k) merupakan *centroid* yang akan dihitung kembali berdasarkan $\sum ai/n, \sum bi/n, \sum ci/n$.

III. HASIL DAN DISKUSI

A. *Dataset*

Penelitian ini menggunakan *dataset* publik yaitu Iris dan Wine yang dapat diakses pada https://archive.ics.uci.edu/ml/datasets.html. Identifikasi masing-masing *dataset* yang digunakan pada penelitian ini menampilkan kelas, atribut numerik, jumlah *record* dan *missing value* seperti pada Tabel 1.

TABEL 1. IDENTIFIKASI *DATASET*

Dataset	Kelas	Atribut Numerik	Jumlah *Record*	*Missing Value?*
Iris	3	4	150	0
Wine	3	13	178	0

Dataset Iris merupakan *database* paling dikenal dapat ditemukan dalam literatur *pattern recognition* yang dibuat berdasarkan penelitian dari R. A Fisher dan *dataset* Iris masih sering dirujuk hingga hari ini. Tabel 2 menunjukkan *dataset* Iris yang berjumlah 150 *record*, mempunyai 4 atribut numerik yaitu *sepal length, sepal width, petal length,* dan *petal width* serta berisi 3 kelas yaitu Iris Sentosa, Iris Versi *colour*, dan Iris Virginica.

TABEL 2. *DATASET* IRIS

Sepal Length	Sepal Width	Petal Length	Petal Width	id	Kelas
5,1	3,5	1,4	0,2	id_1	Iris-setosa
4,7	3,2	1,3	0,2	id_3	Iris-setosa
7,0	3,2	4,7	1,4	id_51	Iris-versicolor
6,9	3,1	4,9	1,5	id_53	Iris-versicolor
6,3	3,3	6,0	2,5	id_101	Iris-virginica
6,3	2,9	5,6	1,8	id_104	Iris-virginica

Dataset Wine merupakan hasil analisis kimia dari anggur yang tumbuh di daerah yang sama di Italia, tetapi berasal dari tiga kultivar yang berbeda. Tabel 3 menunjukkan dataset Wine yang berjumlah 178 *record*, mempunyai 13 atribut numerik yaitu alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, dan proline serta berisi 3 kelas yaitu kelas 1, kelas 2, dan kelas 3.

TABEL 2. *DATASET* IRIS

a1	a2	a3	...	...	a13	id	Kelas
14,23	1,71	2,43	...	...	1065	id_1	1
14,37	1.95	2,5	...	...	1480	id_4	1
12,37	0,94	1,36	...	...	520	id_60	2
12,33	1,1	2,28	...	...	680	id_61	2
12,86	1,35	2,32	...	...	630	id_72	3
12,88	2,99	2,4	...	...	530	id_73	3

B. Davies Bouldin Index

Davies-Bouldin Index[20] didasarkan pada kesamaan dari ukuran klaster yang berdasarkan pada penyebaran klaster dan ketidaksamaan ukuran klaster. Pendekatan ini adalah untuk memaksimalkan jarak *inter cluster* dan meminimalkan jarak *intra cluster* yang dapat pada Persamaan 2:

$$Si = \frac{1}{|Ci|}\sum_{x \in Ci}\{\|x - zi\|\} \quad (2)$$

Dimana Ci sebagai banyaknya titik yang masuk ke dalam klaster *i*, x adalah data, dan Zi *centroid* dari klaster *i.* Sedangkan jarak antara klaster didefinisikan pada Persamaan 2:

$$d_{ij} = \|zi - zj\| \quad (3)$$

Dimana zi *centroid* dari klaster *i* dan zj *centroid* dari klaster *j*. Perhitungan jarak d_{ij} dapat mengunakan e*uclidean.* Selanjutnya akan mendefinisikan Ri, qt untuk klaster Ci pada Persamaan 4

$$Ri, qt = \max_{j,j \neq i}\left\{\frac{s_{i,q}+s_{j,q}}{d_{ij,t}}\right\} \quad (4)$$

Selanjutnya *Davies-Bouldin Index* didefinisikan pada Persamaan 5:

$$DB = \frac{1}{K}\sum_{i=1}^{K} Ri, qt \quad (5)$$

C. Hasil dan Diskusi

Pada sesi ini akan dibahas mengenai tahapan yang dilakukan untuk mengelompokkan *dataset* Iris dan Wine dengan menggunakan K-Means dan K-Medoids. Setelah didapatkan hasil, maka langkah selanjutnya adalah menghitung dan membandingkan nilai evaluasi dari ke dua algoritma tersdatasetebut. Tahapannya adalah:

Tahap 1: Memilih algoritma yang akan digunakan

Tahap 2: Menentukan jumlah partisi/klaster dari *dataset* yang akan dibentuk. Kelas yang terbentuk dari masing-masing *dataset* adalah sebanyak 3 kelas dengan jumlah data pada *dataset* Iris sebanyak 150 data sedangkan Wine sebanyak 178 data.

Tahap 3: Memilih *centroid* yang merupakan merupakan titik pusat pada suatu kelas secara acak.

Tahap 4: Selanjutnya adalah menghitung jarak terdekat antara objek dengan titik pusat. Pada penelitian ini perhitungan jarang dilakukan dengan *Euclidean Distance.*

Tahap 5: Kemudian setiap objek yang tersisa dikelompokkan dengan objek yang memiliki kemiripan data.

Tahap 6: Menghitung kembali *centroid* masing-masing klaster yang terbentuk untuk memperbarui *centroid* baru.

Tahap 7: Ulangi langkah ke 3 sampai 6 hingga tidak ada data yang berpindah dari satu klaster ke klaster lainnya.

Tahap 8: Selanjutnya, jika tidak ada data yang berpindah dari satu klaster ke klaster lainnya maka akan dilakukan proses evaluasi menggunakan *Davies Bouldin Index* untuk mengukur kesamaan dari ukuran *cluster* yang berdasarkan pada penyebaran klaster dan ketidaksamaan ukuran klaster.

Setelah melakukan seluruh tahapan di atas maka selanjutnya adalah melakukan perbandingan hasil evaluasi menggunakan *Davies Bouldin Index* dari ke dua algoritma tersebut.

Pada penelitian ini baik K-Means maupun K-Medoids ditentukan sebanyak 3 kelas, yang selanjutnya adalah menentukan *centroid* atau titik pusat pada data dari masing-masing kelas yang ditunjukkan pada Tabel 4 untuk *dataset* Iris dan Tabel 5 untuk *dataset* Wine.

TABEL 3. CENTROID IRIS MENGGUNAKAN *K-MEANS*

Atribut	Centroid 1	Centroid 2	Centroid 3
A1	5,902	6,850	5,006
A2	2,748	3,074	3,418
A3	4,394	5,742	1,464
A4	1,434	2,071	0,244

TABEL 4. CENTROID WINE MENGGUNAKAN *K-MEANS*

Atribut	Centroid 1	Centroid 2	Centroid 3
A1	12,930	12,517	13,804
A2	2,504	2,494	1,883
A3	2,408	2,289	2,426
....	....	....	....
....	....	....	....
A13	728,339	458,232	1195,149

Berdasarkan penentuan *centroid* seperti yang telah ditunjukkan pada pada Tabel 4 dan Tabel 5 selanjutnya adalah menghitung hasil evaluasi dari algoritma K-Means menggunakan *Davies Bouldin Index* seperti yang ditunjukkan pada Tabel 6. Pada *dataset* Iris, kelas yang terbentuk adalah sebanyak 3 kelas. Kelas pertama memiliki data sebanyak 62, kelas ke dua sebanyak 38 data, dan kelas ketiga adalah sebanyak 50 data dengan hasil evaluasi adalah sebesar 0,662. *Dataset* Wine kelas yang terbentuk adalah sebanyak 3 kelas. Kelas pertama memiliki data sebanyak 62, kelas ke dua sebanyak 69 data, dan kelas ke tiga sebanyak 47 data dengan hasil evaluasi adalah sebesar 0,534.

TABEL 5. EVALUASI *K-MEANS*

Dataset	Kelas	Atribut Numerik	Nilai Evaluasi
Iris	3	4	0,662
Wine	3	13	0,534

Pengelompokan data berdasarkan algoritma K-Medoids proses yang dilakukan adalah sama dengan K-Means yaitu terlebih dahulu menentukan kelas sebanyak 3 dan menentukan *centroid* atau titik pusat pada masing-masing *dataset*. *Dataset* Iris ditunjukkan padaTabel 7 dan *dataset* Wine ditunjukkan pada Tabel 8.

TABEL 6. *CENTROID* IRIS MENGGUNAKAN *K-MEDOIDS*

Atribut	*Centroid* 1	*Centroid* 2	*Centroid* 3
A1	5,9	5,0	5,1
A2	3,0	3,3	2,5
A3	5,1	1,4	3,0
A4	1,8	0,2	1,1

TABEL 7. *CENTROID* WINE MENGGUNAKAN *K-MEDOIDS*

Atribut	*Centroid* 1	*Centroid* 2	*Centroid* 3
A1	13,72	13,17	14,13
A2	1,43	2,59	4,1
A3	2,5	2,37	2,74
....	....	....	....
....	....	....	....
A13	1285,0	840,0	560,0

Tabel 9 menunjukkan hasil evaluasi menggunakan algoritma K-Medoids. Pada *dataset* Iris, kelas yang terbentuk adalah sebanyak 3 kelas. Kelas pertama memiliki data sebanyak 85, kelas ke dua sebanyak 50 data, dan kelas ke tiga adalah sebanyak 15 data dengan hasil evaluasi adalah sebesar 0,748.

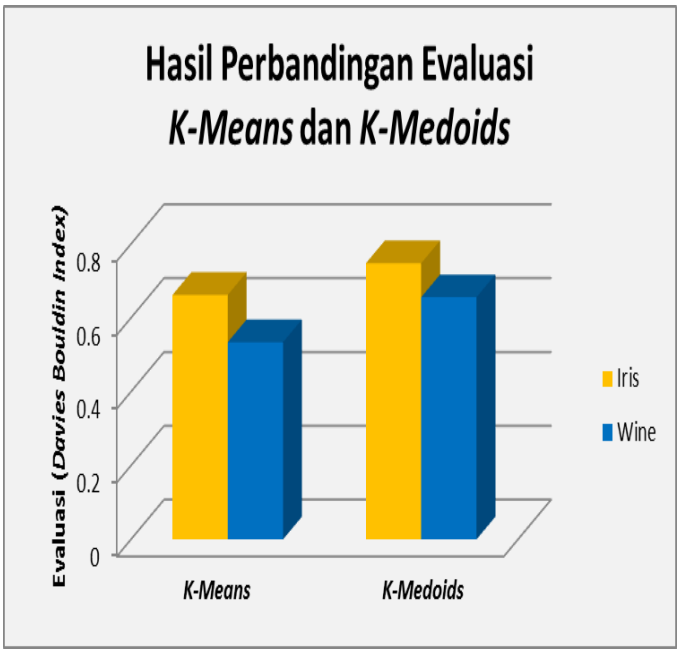
Dataset Wine kelas yang terbentuk adalah sebanyak 3 kelas. Kelas pertama memiliki data sebanyak 33, kelas ke dua sebanyak 45 data, dan kelas ke tiga sebanyak 100 data dengan hasil evaluasi adalah sebesar 0,656.

TABEL 8. EVALUASI *K-MEDOIDS*

Dataset	Kelas	Atribut Numerik	Nilai Evaluasi
Iris	3	4	0,748
Wine	3	13	0,656

Hasil perbandingan evaluasi K-Means dan K-Medoids dapat dilihat pada Gambar 1 di mana dengan menggunakan K-Means pada *dataset* Iris menunjukkan hasil evaluasi sebesar 0,662 dan pada *dataset* Wine adalah sebesar 0,534.

Haasil evaluasi dengan menggunakan K-Medoids pada *dataset* Iris menunjukkan hasil evaluasi sebesar 0,748 dan pada *dataset* Wine adalah sebesar 0,656.



Gambar 1. Grafik perbandingan evaluasi

Berdasarkan hasil evaluasi *dataset* Iris menggunakan K-Means menunjukkan hasil yang lebih baik dibandingkan dengan hasil evaluasi *dataset* Iris dengan menggunakan K-Medoids dengan perbedaan hasil evaluasi sebesar 0,082. Ada pun hasil *dataset* Wine dengan menggunakan K-Means menunjukkan hasil yang lebih baik dibandingkan dengan K-Medoids dengan perbedaan hasil evaluasi sebesar 0,122. Pada penelitian ini juga menunjukkan perbedaan hasil evaluasi dalam menggunakan *dataset* Iris maupun Wine yang terlihat lebih memiliki hasil evaluasi yang lebih baik dengan menggunakan K-Means. Hal ini terbukti dengan kesimpulan pada penelitian Velmurugan[1] dan Tiwari dan Singh[6] yang menyatakan bahwa K-Means memiliki performa yang lebih handal dalam menangani *dataset* kecil.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat diketahui bahwa kedua algoritma pengelompokan yaitu K-Means dan K-Medoids dengan dilakukan eksperimen dengan *dataset* yang berukuran kecil, maka K-Means lebih efektif dalam menangani data dengan ukuran kecil. Pada *dataset* Iris menggunakan K-Means menunjukkan hasil evaluasi sebesar 0,662, sedangkan dengan menggunakan K-Medoids hasil evaluasi menunjukkan hasil sebesar 0,748.

Hasil evaluasi pada *dataset* Wine dengan menggunakan K-Means adalah sebesar 0,534, sedangkan dengan menggunakan K-Medoids adalah sebesar 0,656. Hal ini disebabkan karena K-Means memiliki performa komputasi yang rendah dibandingkan dengan K-Medoids. Untuk penelitian selanjutnya dapat dilakukan evaluasi pada kedua algoritma klastering di atas dengan dilakukan eksperimen dengan *dataset* berukuran besar.

ISBN: 978-602-50525-0-7

DAFTAR PUSTAKA

[1] V. T., "Performance based analysis between K-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data," Appl. Soft Comput., vol. 19, pp. 134–146, Jun. 2014.

[2] A. K. Jain, "Data clustering: 50 years beyond K-Means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010.

[3] X. Wu et al., "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.

[4] J. Macqueen, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Stat., vol. 233, no. 233, pp. 281–297, 1967.

[5] J. Han and M. Kamber, "Data mining: concepts and techniques," vol. 49, no. 6, pp. 49-3305-49–3305, Feb. 2012.

[6] M. Tiwari and R. Singh, "Comparative Investigation of K-Means and K-Medoid Algorithm on Iris Data," vol. 4, no. 8, pp. 69–72, 2012.

[7] T. Velmurugan, "Evaluation of k-Medoids and Fuzzy C-Means Clustering Algorithms for Clustering Telecommunication Data," pp. 115–120, 2012.

[8] T. Velmurugan, "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points," vol. 3, no. 5, pp. 1758–1764, 2012.

[9] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Technique Third Edition. 2012.

[10] D. L. Davies and D. W. Bouldin, "A cluster separation measure.," IEEE Trans. Pattern Anal. Mach. Intell., vol. 1, no. 2, pp. 224–227, 1979.