Komparasi Farthest First dan K-Mean pada Clustering Huruf Alphabet

Dian Nursantika*, Eddie Krishna Putra
Jurusan Informatika, Fakultas MIPA
Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirman, Cimahi
dianursantika@lecture.unjani.ac.id, eddiekrishnaputra@lecture.unjani.ac.id

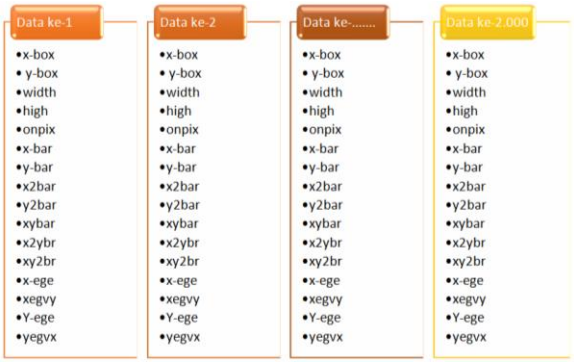
***Abstrak*—Clustering merupakan metode dalam mengenali pola tertentu dari sebuah dataset, dengan tujuan membagi dataset sesuai dengan sifat yang dimiliki oleh dataset tersebut. Pada penelitian ini, dilakukan clustering terhadap 2.000 dataset huruf alphabet yang masing-masing data set memiliki nilai 16 atribut. Penelitian ini melakukan komparasi hasil dari dua metode, yaitu Farthest First Clustering dan K-Mean Clustering. Hasil komparasi dari penelitian ini dengan menggunakan dataset yang sama pada ke dua metode, menunjukkan bahwa metode Farthest First Clustering lebih unggul diabandingkan dengan K-Mean Clustering. Hal tersebut dapat dilihat dari hasil cluster centroid Farthest First Clustering yaitu 78% dan waktu tempuh clustering yaitu 0,04 detik.**

Kata kunci—Clustering; Farthest First; K-Mean; huruf alphabet.

I. PENDAHULUAN

Penelitian mengenai pemrosesan citra digital tidak ada hentinya dilakukan demi mengembangkan ilmu pengetahuan berbasis komputasi terutama dalam sebuah citra. Penelitian yang ini lakukan merupakan penelitian yang berbasis citra dengan fokus penelitian pada letak koordinat citra, terdapat 16 atribut sebagai koordinat citra yang digunakan sebagai input dalam penelitian ini. Penggunaan koordinat citra bukan tanpa alasan, hal tersebut dilatarbelakangi oleh[1] yang mengusung penulangan atau skeleton terhadap citra digital. Skeleton dalam citra digital erat kaitannya dengan koordinat citra, dikarenakan penggunaan relasi piksel antar titik koordinat dari citra tersebut.

Penelitian lain yang menjadi pemicu penelitian ini yaitu[2], penelitian tersebut melakukan clustering terhadap citra huruf arab sebagai objek penelitiannya, sedangkan objek penelitian pada penelitian ini menggunakan citra huruf alphabet untuk dilakukan clustering. Huruf alphabet tersebut terdiri dari 16 atribut titik koordinat dan 2.000 data untuk setiap 16 atribut. Ilustrasi penggunaan data pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Data set huruf alphabeth

Jumlah atribut yang digunakan dalam penelitian ini sebanyak 16 atribut yang terdiri dari titik koordinat x-box, y-box, width, high, onpix, x-bar, y-bar, x2bar, y2bar, xybar, x2ybr, xy2br, x-ege, xegvy, y-ege, yegvx.

II. METODE

A. *Farthest First Clustering*

Berdasarkan penelitian[4] menunjukkan bahwa FF clustering merupakan kemampuan yang handal dalam memproses dataset yang besar dikarenakan adanya penentuan centre pada setiap cluster yang ditentukan. FF clustering merupakan metode tercepat dibandingkan dengan metode clustering lainnya dalam penelitian[5]. Memiliki tahapan yang serupa dengan K-Mean clustering yaitu adanya pemilihan center dan penentuan elemen-elemen yang berada pada centre[6]. Adapun algoritma dari FF clustering yaitu[7]:

Input: k //Jumlah cluster yang diinginkan
D = {x_1, x_2, …, x_n} //set elemen
Output: K = {C_1, C_2, …, C_k} //set cluster (k)

K-Mean Algorithm
Tetapkan nilai awal untuk menentukan titik tengah
Repeat
Tetapkan setiap item xi ke cluster yang paling dekat
Hitung mean baru untuk setiap cluster

B. *K-Mean Clustering*

K-Mean Clustering adalah sebuah metode Clustering Analysis yang digunakan untuk mempartisi n ke dalam K Cluster[8], dengan menggunakan nilai *mean* terdekat. Berdasarkan [9], tujuan dari K-Mean Clustering yaitu membagi M titik yang memiliki N dimensi ke dalam K Cluster, sehingga memiliki nilai *sum of square* yang kecil. Penelitian lain mengenai K-Mean Clustering telah dilakukan oleh[10] dalam melakukan *clustering* terhadap citra buah, dengan hasil *clustering* yang sangat menjanjikan. Algoritma K-Mean Clustering berdasarkan[7] adalah:

Input: k //Jumlah cluster yang diinginkan
D = {x_1, x_2, …, x_n} //set elemen
Output: K = {C_1, C_2, …, C_k} //set cluster (k)

Farthest First Algorithm
Pilih cluster pertama secara acak
For(i=2, …k)
For each titik yang tersisa
Hitung jarak ke titik pusat
Pilih titik yang memiliki jarak maksimum
sebagai pusat baru
for each titik yang tersisa
hitung jarak ke masing-masing pusat cluster
simpan di *cluster* dengan jarak yang minimum

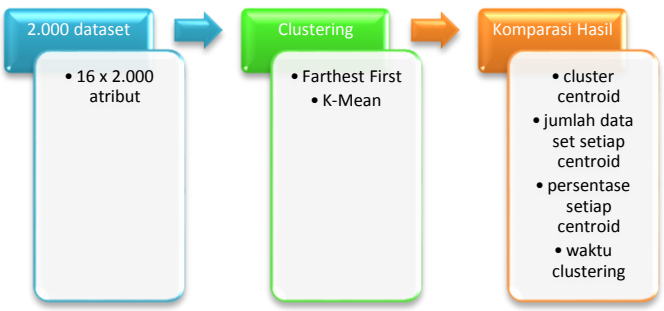
C. *Perhitungan Jarak Manhattan Distance*

Metode yang ini gunakan dalam perhitungan jarak pada proses *clustering* yaitu dengan menggunakan metode Manhattan Distance[11], lihat Persamaan 1.

$$MD_{(x,y)} = \sum_{i=1}^{n} |x_i - y_i| \qquad (1)$$

Dimana n adalah jumlah dari variabel atau data, sedangkan x_i dan y_i adalah nilai dari jumlah atribut yang dibandingkan.

Proses yang dilakukan dalam penelitian ini yaitu menyiapkan 2.000 dataset (Gambar 1), data set tersebut kemudian di clustering dengan menggunakan dua metode yaitu Farthest First Clustering dan K-Mean Clustering. Hasil dari kedua metode tersebut kemudian dibandingkan dari berbagai analisis yang akan dikemukakan di bagian hasil penelitian. Adapun kinerja dari proses penelitian ini dapat dilihat pada Gambar 2.



Gambar 2. Diagram sistem penelitian

III. HASIL DAN DISKUSI

Input huruf *alphabet* yang terdiri dari 16 atribut (x-box, y-box, width, high, onpix, x-bar, y-bar, x2bar, y2bar, xybar, x2ybr, xy2br, x-ege, xegvy, Y-ege, yegvx. Jumlah keseluruhan data yang akan di cluster yaitu 2.000 data. Tabel dataset 2.000 alphabet, dengan 16 atribut untuk setiap database. Tabel 1. dataset 2.000 alphabet, dengan 16 atribut untuk setiap dataset.

TABEL 1. DATASET ALPHABET (16 ATRIBUT)

Atribut	Dataset			
	ke-1	ke-2	ke-……	ke-2000
x-box	2	5	…	4
y-box	8	12	…	9
Width	3	3	…	6
High	5	7	…	6
Onpix	1	2	…	2
x-bar	8	10	…	9
y-bar	13	5	…	5
X2bar	0	5	…	3
Y2bar	6	4	…	1
Xybar	6	13	…	8
X2ybr	10	3	…	1
Xy2br	8	9	…	8
x-ege	0	2	…	2
Xegvy	8	8	…	7
y-ege	0	4	…	2
Yegvx	8	10	…	8

Jumlah cluster yang digunakan dalam proses clustering penelitian ini yaitu 2 cluster, yang diasumsikan kedalam cluster 0 dan cluster 1 baik pada Farthest First clustering maupun K-Mean Clustering.

Cluster centroid dari metode Farthest First Clustering dapat dilihat pada Tabel 2, yang menunjukkan data atribut dari *cluster centroid* pertama dan ke dua. Jumlah data set pada *centroid* pertama adalah 15.696 data dan jumlah data set pada *centroid* ke dua adalah 4.303 data. Waktu yang dibutuhkan untuk *clustering* dengan metode Farthest First Clustering yaitu 0,04 detik dengan persentase di *cluster* pertama 78% dan *cluster* ke dua =22%.

TABEL 2. CLUSTER CENTROID METODE FARTHEST FIRST CLUSTERING

Atribut	Pertama	Kedua
x-box	9	1
y-box	13	0
Width	7	1
High	8	0
Onpix	5	0
x-bar	8	2
y-bar	7	1
X2bar	5	6
Y2bar	5	4
Xybar	9	0
X2ybr	4	3
Xy2br	9	4
x-ege	7	0
Xegvy	5	8
y-ege	8	0
yegvx	11	8

Hasil penelitian yaitu *unsupervised learning* dari hasil *clustering*. Adapun metode K-Mean Clustering pada *centroid* pertama adalah 14.490 data dan jumlah data set pada *centroid* ke dua adalah 5.510 data, dengan persentase pada *centroid* pertama yaitu 46% dan pada *centroid* kedua yaitu 54% dengan menghabiskan waktu *clustering* sebanyak 1,71 detik. Tabel 3, merupakan tabel yang menunjukkan data yang digunakan sebagai *cluster centroid* pada metode K-Mean Clustering.

TABEL 3. CLUSTER CENTROID METODE K-MEAN CLUSTERING

Atribut	Pertama	Kedua
x-box	4	4
y-box	7	8
Width	5	5
High	6	6
Onpix	3	3
x-bar	7	8
y-bar	8	6
X2bar	5	3
Y2bar	5	6
Xybar	7	10
X2ybr	7	4
Xy2br	8	8
x-ege	3	2
Xegvy	8	7
y-ege	4	3
yegvx	8	8

IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, yaitu *clustering* dengan menggunakan dua metode yaitu metode Farthest First Clustering dan K-Mean Clustering. Maka dapat ditarik kesimpulan berupa nilai komparasi dari kedua metode tersebut. Dapat dibuktikan bahwa dengan menggunakan dataset yang sama yaitu 2.000 dataset dan jumlah penentuan *cluster* yang sama yaitu 2 *cluster* untuk masing-masing metode, menunjukkan bahwa nilai rata-rata *cluster centroid* Farthest First Clustering adalah 4,93 Dan nilai rata-rata *cluster centroid* K-Mean Clustering adalah 5,81. Persentase terbaik dari *cluster* pertema yaitu pada metode Farthest First Clustering sebanyak 78% dan persentase terbaik dari *cluster* ke dua yaitu pada metode K-Mean sebesar 54%, sedangkan waktu *clustering* terbaik yaitu 0,04 detik diunggulkan oleh metode Farthest First Clustering.

DAFTAR PUSTAKA

[1] G. Goyal dan R. Luthra, "Skeleton Generation for Digital Images Based on Performance Evaluation Parameters", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 9. No. 2 (2016), pp.47-58.

[2] A. N. De Roeck dan W. Al-Fares, "A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots", University of Essex, U.K.

[3] Pallavi dan S. Godara. "A Comparative Performance Analysis of Clustering Algorithm", International Journal of Engineering Research and Application, Vol. 1, Issue 3, pp.441-445.

[4] D. Asir Antony Gnana Singh, A. Escalin Fernando dan E. Jebamalar Leavline, "Performance Analysis on Clustering Approach for Gene Expression Data", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2016.

[5] G. Sehgal dan K. Garg, "Comparason of Various Clustering Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3074-3076.

[6] D. Ramya dan D. Rao, "Performance Evaluation of Learning by Example Techniques over Different Datasets", International Journal of Innovative Research in Science, Engineering and Technology, ISSN: 2319-8753.

[7] F. Kaya Gulagiz dan S. Sahin, "Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms", International Journal of Computer Engineering and Information Technology, Vol. 9, No. 1, January 2017, 6-14, E-ISSN 2412-8856.

[8] Sharmila dan R. C Mishra, "Performance Evaluation of Clustering Algorithm", International Journal of Engineering Trends and Technology, Vol. 4 Issue 7-July 2013.

[9] J. A. Hartigan dan M. A. Wong, "A K-Means Clustering Algorithm", Journal of the Royal Statistical Society.

[10] S. R. Dubey, P. Dixit, N. Singh dan J. P. Gipta, "Infected Fruit Part Detection using K-Means Clustering Segmentation Technique", International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, No. 2.

[11] K. M. Ponnmoli and S. Selvamuthukumaran, "Analysis of Face Recognition using Manhattan Distance Algorithm with Image Segmentation", International Journal of Computer Science and Mobile Computing, Vol. 3 Issue. 7, July-2014, 18-27.

ISBN: 978-602-50525-0-7