JITE, 3 (2) Januari 2020 ISSN 2549-6247 (Print) ISSN 2549-6255 (Online)

JITE (Journal of Informatics and Telecommunication Engineering)

Available online http://ojs.uma.ac.id/index.php/jite **DOI : 10.31289/jite.v3i2.3211**

Initial Centroid Optimization of K-Means Algorithm Using Cosine Similarity

Fadhillah Azmi[1)*], Kevin Utama[1)], Oki Thomas Gurning[1)], & Syukurmn Ndraha[1)]

1) Prodi Teknik Informatika, Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indonesia, Indonesia

*Coresponding Email: azmi.fadhillah007@gmail.com

Abstrak

Clustering salah satu metode yang sering digunakan di berbagai bidang yang melakukan analisis data, termasuk penggalian data, pengambilan dokumen, segmentasi gambar, dan klasifikasi pola. Adapaun tujuan dari metode tersebut adalah untuk mengelompokkan data ke dalam suatu *cluster* sehingga kesamaan antara anggota data dalam suatu data informasi yang telah di-*cluster* yang sama adalah maksimal, di sisi lain untuk kesamaan di antara anggota data yang lain berbeda *cluster* minimal. Ada beberapa pendekatan metode untuk mengurangi kesalahan pada saat *centroid* awal yang dipilih selama proses pengelompokan berlangsung. Disini data yang digunakan adalah data acak yang dibuat secara manual yatu 30 data dan 5 atribut, sehingga diperoleh hasil akurasi clustering dalam centroid dengan menggunakan metode K-Means memiliki signifikan 86.67%, sedangkan menggunakan K-Means dengan cosine similarity tidak jauh berbeda yaitu sebesar 89.7%, maka dari itu hasilnya cukup baik.
Kata Kunci: Optimalisasi, K-Means, Cosine Similarity, Initial Centroid.

Abstract

Clustering is a method that is often used in various fields that conduct data analysis, including data mining, document retrieval, image segmentation, and pattern classification. The purpose of the method is to group data into a cluster so that the similarity between data members in a data that has been in the same cluster is maximum, on the other hand for similarity among other data members different minimum clusters. There are several method approaches to reduce errors during the initial centroid selected during the grouping process. Here the data used are random data created manually by 30 data and 5 attributes, so that the accuracy of clustering in centroids obtained by using the K-Means method has a significant 86.67%, while using K-Means with cosine similarity is not much different that is equal to 89.7 %, therefore the results are quite good.
Keywords: Optimization, K-Means, Cosine Similarity, Initial Centroid.

How to Cite: Azmi, F. Utama, K. Gurning, O.T. & Ndraha. (2020). Optimalisasi *Centroid* Awal Algoritma K-Means dengan *Cosine Similarity*. *JITE (Journal of Informatics and Telecommunication Engineering)*. 3 (2): 224-231

I. PENDAHULUAN

Salah satu metode cluster non-hierarkis yang ada adalah K-Means yang membagi data yang ada menjadi satu atau lebih kelompok. Penting untuk dicatat bahwa algoritma K-means adalah sangat sensitif terhadap *outlier*. Pencilan adalah data yang jauh dari sebagian besar data lain, dan dengan demikian tidak dapat diterapkan ketika dimasukkan ke dalam sebuah cluster. Jenis data ini dapat mendistorsi nilai rata-rata cluster berlebihan Karena keterbatasan waktu, penelitian ini mengasumsikan outlier tidak signifikan terhadap hasil penelitian. (Usino, 2019).

Ada beberapa jenis algoritma yang dilakukan untuk mengelompokan kelas dari suatu data informasi dengan menggunakan metode berbasis jarak. Metode berbasis jarak tersebut sangat popular di terapkan untuk kasus yang serupa dengan ini, yaitu algoritma pengelompokan bagian (*partition*) dan dilakukan secara hierarkis (*heirarcy*), yaitu algoritma K-Means merupakan salah satu metode yang paling banyak digunakan dalam kasus seperti ini yang mana memisahkan data menjadi k yang saling berkaitan antara satu dengan yang lainnya dalam sekumpuan data informasi yang akan di-cluster. (Budiman, 2012). Karena kemampuan algoritma K-Means untuk mengelompok-kan sekumpulan data yang sangat besar dan mudah diterapkan, sehingga metode ini paling sering digunakan. Namun, ada kelemahan metode ini yang cukup sensitif terhadap *cluster*, yaitu pemilihan *centroid* awal yang mana tidak menjamin pengelompokan terjadi secara signifikan karena hasil yang diperoleh akan berbeda yang apabila dilakukan pemilihan cen*troid* awal secara acak. Sehingga pada saat hasil akhir c*luster* kemungkinan *centroid* yang diperoleh bukan optimal karena algoritma K-Means akan menjadi penyelesaian optimal lokal. Hal tersebut akan mempengaruhi kualitas algoritma K-Means, maka pemilihan *centroid* awal sangatlah penting di dalam algoritma K-Means. Ada beberapa pendekatan metode untuk mengurangi kesalahan pada saat *centroid* awal yang dipilih selama proses pengelompokan berlangsung, salah satunya dengan menggunakan metode *cosine similarity*. (Mirza, 2008).

Beberapa penelitian yang telah dilakukan yaitu teknik pengelompokan teks secara hybrid dikembangkann untuk mengategorikan teks dalam domain yang diberikan. Dalam pengelompokkan teks dan kategorisasi teks, sumber daya konsumsi adalah masalah utama. Oleh karena itu dilakukan ekstraksi fitur. Teknik ini digunakan untuk mengurangi

jumlah teks. Teks yang direduksi ini mewakili seluruh dokumen teks. (Kazmierska, 2008).

Selanjutnya pada penelitian, metode Euclidian pada K-Means *clustering* menggunakan pendekatan berbasis jarak yang mana dilakukan untuk menemukan kesamaan dalam metode klasifikasi teks dari kumpulan dokumen yang tidak dikenal atau pengambilan informasi secara acak. (Singh, 2016).

Dalam beberapa tahun terakhir seorang penulis mengusulkan metode pengelompokan berbasis partisi. K-means yang mana *centroid* awal diambil secara acak, selanjutnya menghitung jarak antara elemen lainnya dengan membentuk kluster elemen-elemen yang jaraknya sangat kurang dari pusat. K-Means adalah metode yang sederhana, fleksibel dan mudah dimengerti serta mengimplemen-tasikan yang mana bekerja untuk dataset numerik. (Agusta, Y. 2007).

Pada 2014, diusulkan metode yang membagi jarak akar kuadrat dengan standar deviasi yang menghasilkan peningkatan kinerja rata-rata k jauh lebih baik daripada metode k dan mean-k dan itu membutuhkan waktu lebih sedikit untuk merumuskan *cluster*. Apalagi itu tidak hanya berlaku untuk dataset kecil tetapi juga berfungsi sangat baik efisien untuk dataset yang sangat besar. (Vishwanath, 2014).

Pendekatan metode selanjutnya yang digunakan adalah dengan menggunakan *cosine similarity* sebagai penentu jarak untuk centroid awal pada K-Means.

II. STUDI PUSTAKA

Clustering

Clustering adalah salah satu metode yang sering digunakan di berbagai bidang yang berfungsi untuk analisis data, penggalian data, pengambilan dokumen, segmentasi gambar, dan klasifikasi pola. Metode tersebut juga digunakan untuk mengelompokkan data ke dalam suatu *cluster* dari sekelompok data informasi yang mana memiliki kesamaan di antara anggota data dalam suatu *cluster* diperoleh secara maksimal, di sisi lain kesamaan di antara anggota data dari *cluster* yang berbeda adalah minimal. (Nurjayanti, 2011).

Pada proses ini dilakukan pengelompokkan objek ke dalam bentuk subset yang memiliki arti dalam konsep masalah tertentu itulah yang disebut dengan analisis *clustering*. *Clustering* tidak sama dengan klasifikasi yang mana *clustering* tidak berdasarkan pada kelas yang sudah ada. Metode yang mempelajari tentang *unsupervised* karena tidak adanya informasi yang disajikan dalam bentuk

pernyataan benar untuk objek apa pun, hal tersebut disebut dengan *clustering*. Sehingga dapat ditemukan hubungan dari sebelumnya yang tidak diketahui di dalam suatu *data set* yang kompleks. (Larose, 2005).

Analisis *clustering* merupakan teknik analisis yang multivariasi dimana dilakukan untuk mencari dan mengorganisir suatu informasi tentang variabel, sehingga secara relatif dapat ditentukan kelompokmya ke dalam bentuk kelompok yang homogen atau ter-*cluster*. *Cluster* yang dibentuk merupakan metode kedekatan yang secara internal harus berupa homogen yang mana anggota sama dengan anggota yang lain dan secara eksternal tidak sejenis yang mana anggota tidak sama dengan anggota yang lain. (Huliman. 2013).

Algoritma K-Means

Algoritma K-Means pertama kali diperkenalkan oleh Mac Queen JB pada tahun 1976 yang mana algoritma ini merupakan satu algoritma yang *non hierarchi* secara umum digunakan. Algoritma K-Means dimulai dari memilih jumlah k cluster, menugaskan setiap titik data ke pusat cluster terdekat, dan memindahkan setiap pusat cluster ke data rata-rata dan terakhir poin. Langkah-langkah ini diulang beberapa kali untuk mencapai konvergensi. Hasil akhir dari algoritma K-means adalah jumlah cluster yang sesuai. Menciptakan jumlah cluster sebelum mengimplementasikan algoritma dianggap sebagai tidak praktis. Ini juga membutuhkan pengetahuan yang mendalam tentang bidang pengelompokan. Sebelum menerapkan model ruang vektor ke teks dokumen, pengambilan informasi dilakukan melalui suatu proses. Proses input adalah dokumen teks biasa dan outputnya adalah serangkaian token yang digunakan dalam model vektor. Algoritma ini juga merupakan bentuk teknik pembagian kelompok atau partisi (*partition*) yang membagi dan memisahkan suatu data ke suatu data tertentu ke dalam data terpisah yang sesuai dengan informasi data yang ditentukan. Di dalam algoritma K-Means, pada setiap objek yang telah dimasukkan ke dalam suatu kelompok tertentu, selanjutnya akan dilanjutkan ke satu tahapan objek berikutnya yang akan berpindah ke kelompok lain. (Oscar, 2013)

Metode ini juga dapat dilakukan untuk membentuk suatu partisi (*partition*) data ke dalam bentuk satu atau lebih *cluster* yang mana dapat membagi objek ke dalam *cluster* yang berbeda, sehingga data yang mempunyai karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama berdasarkan data informasinya dan

di sisi lain data yang tidak mempunyai karakteristik sama atau berbeda, maka dikelompokkan pada kelompok lain yang mempunyai kesaaman dengan kelompok yang berbeda itu juga. (Giyanto dan Heribertus, 2008).

Adapun proses *clustering* dimulai dengan melakukan identifikasi pada data yang akan dianalisa secara *cluster*, X_{ij} (I = 1, … , n; j = 1, … , m), dimana n adalah jumlah data yang akan di-*cluster* dan m adalah jumlah variabel. Pada iterasi awal, *centroid* pada setiap *cluster* ditetapkan secara bebas (*random*), c_{kj} (k = 1, … , k; j = 1, … , m). Selanjutnya, menghitung jarak pada data ke – I (x_i), *centroid cluster* pada data ke – k (c_k), disebut dengan (d_{ik}), sehingga formula yang digunakan adalah formula Euclidean sebagai berikut:

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(c_{ij} - c_{kj})^2} \quad \ldots\ldots\ldots\ldots\ldots \quad (1)$$

Dimana

c_{ij} = data yang akan di-*cluster*; k = 1, … , k; j = 1, … , m.

c_{kj} = *cluster* yang dilakukan secara *random*;

k = 1, … , k; j = 1, … , m.

Di dalam suatu data yang akan menjadi anggota dari suatu *cluster* ke – k yang jika data tersebut ke *cluster centroid* pada data ke – k yang memiliki nilai paling kecil apabila dibandingkan dnegan jarak ke *cluster centroid* lainnya. Sehingga, kasus tersebut dapat dihitung dengan menggunakan persamaan berikut: (Nugraheni, 2011)

$$Min \sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j=1}^{m}(c_{ij} - c_{kj})^2} \quad \ldots\ldots\ldots\ldots \quad (2)$$

Nilai pada *cluster centroid* yang dihasilkan dapat dihitung dengan cara menghitung nilai rata-rata dari data-data tersebut yang merupakan anggota pada *cluster itu*:

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

Dimana:

x_{ij} = ϵ cluster data ke – k

p = jumlah anggota cluster data ke – k

Cosine Similarity

Konsep kesamaan penting dalam banyak bidang ilmiah seperti matematika, statistik, dan Ilmu Komputer. Ukuran kesamaan dapat menentukan relevansi antara keduanya item, dua pengguna, dua permintaan, dua artikel, dan banyak lagi. Ada banyak metode kesamaan diusulkan dalam literatur. Biasanya, mereka dikategorikan berdasarkan tipe data mereka seperti data numerik, data kategorikal, data deret waktu, data biner, dan data tipe campuran. Data numerik metode kesamaan biasanya menggunakan jarak Euclidean, jarak Manhattan,

d_{ik} = *centroid clu*

Minkowski jarak, dan jarak rata-rata. (Xiong, 2017).

Metode ini digunakan untuk pengukuran kesamaan antara dua vektor (dua dokumen di Ruang Vektor), untuk menghitung sudut *cosinus* antara dua dokumen. Prinsip kerjanya adalah evaluasi pengukuran orientasi. (Paulanda, 2012).

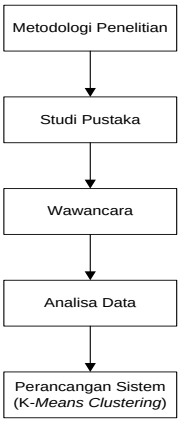
$$\text{cosSim}(d_j, q_k) = \frac{\sum_{i=1}^{n}(td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} tq_{ik}^2}} \quad (4)$$

Keterangan:

cosSim(dj, qk) = tingkat kesamaan dokumen dengan; query tertentu tdij = term ke –I dalam vektor untuk dokumen ke-j; tqik = term ke-i dalam vektor untuk wuey; ke-n = jumlah term yang unik dalam data set.

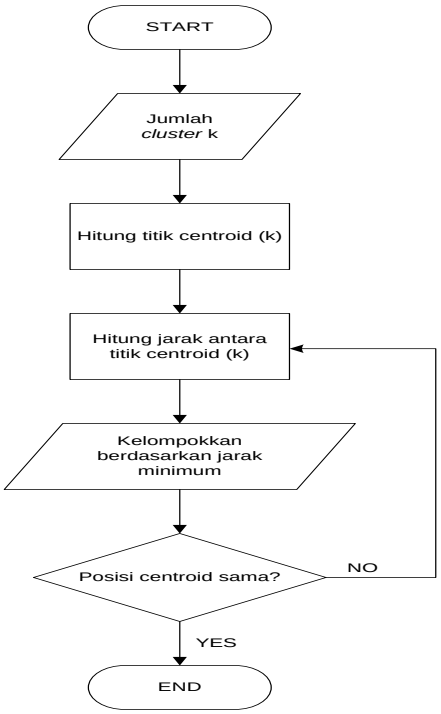
III. METODE PENELITIAN

Pada penelitian ini, data yang digunakan adalah data manual yang diambil sampel 30 data dan 5 atribut untuk pengujian *cluster*. Adapaun dalam penyelesaian masalah dilakukan beberapa metode penelitian adalah sebagai berikut:



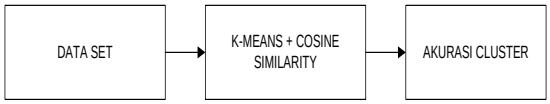
Gambar 1. Metodelogi Penelitian

Pada dasarnya, algoritma K-Means merupakan algoritma *clustering* yang secara umum dilakukkan secara beberapa tahap adalah sebagai berikut:



Gambar 2. Algoritma K-Means Secara Umum

1. Menentukan berapa jumlah anggota *cluster* (k) yang akan dibentuk.
2. Menentukan nilai k yaitu *centroid* (titik pusat dari suatu *cluster*) awal secara *random*.
3. Menghitung jarak pada setiap data ke dalam masing-masing *centroid*.
4. Menentukan posisi *centroid* yang baru dengan menghitung nilai rata-rata dari suatu data pada centroid yang sama.
5. Apabila posisi *centroid* yang baru dan *centroid* sebelumnya tidak sama, maka lakukan kembali tahap 3.



Gambar 3. Algoritma K-Means dan Cosine Similarity

IV. HASIL DAN PEMBAHASAN

Pada penelitian ini akan dicari nilai lulus dan tidak lulus dari hubungan kategori kelulusan dengan data calon pegawai. Data yang digunakan berdasarkan dari nilai tes yang dilakukan. Adapun atribut kategori kelulusan adalah sebagai berikut:

Tabel 1. Atribut Kategori Kelulusan

No	Atribut Kategori
1	Administrasi
2	Kesehatan Awal
3	Psikologi
4	Wawancara
5	Kesehatan Akhir

Tabel 2. Data Kriteria Kelulusan

Calon Pegawai	Kriteria Kelulusan				
	Administrasi	Kesehatan Awal	Psikologi	Wawancara	Kesehatan Akhir
1	5	5	6	7	6
2	6	7	7	7	6
3	7	7	8	7	8
4	8	8	9	9	9
5	5	5	6	7	6
6	5	5	6	7	6
7	6	7	7	7	6
8	7	7	8	7	8
9	8	8	9	8	9
10	5	5	5	7	6
11	5	5	6	7	6
12	6	7	7	7	6
13	7	7	8	7	8
14	8	8	9	8	9
15	5	5	6	7	6
16	5	5	6	7	6
17	6	7	7	7	6
19	8	8	9	8	9
20	5	5	6	7	6
21	5	5	6	7	6
22	6	7	7	7	6
23	7	7	8	7	8
24	8	8	9	8	9
25	5	5	6	7	6
26	5	5	6	7	6
27	6	7	7	7	6
28	7	7	8	7	8
29	8	8	9	8	9
30	5	5	6	7	6

Centroid awal pada metode K-Means dipilih secara acak, maka pada *cluster* 1 diperoleh:

$$d_{11} = \sqrt{(5-8)^2 + (5-8)^2 + (6-9)^2 + (7-9)^2 + (6-9)^2} = 6.32$$

$$d_{12} = \sqrt{(6-8)^2 + (7-8)^2 + (7-9)^2 + (7-9)^2 + (6-9)^2} = 4.69$$

$$d_{13} = \sqrt{(7-8)^2 + (7-8)^2 + (8-9)^2 + (7-9)^2 + (8-9)^2} = 2.83$$

$$d_{14} = \sqrt{(8-8)^2 + (8-8)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2} = 0$$

$$d_{15} = \sqrt{(5-8)^2 + (5-8)^2 + (6-9)^2 + (7-9)^2 + (6-9)^2} = 6.32$$

$$d_{16} = \sqrt{(5-8)^2 + (5-8)^2 + (6-9)^2 + (7-9)^2 + (6-9)^2} = 6.32$$

$$d_{17} = \sqrt{(6-8)^2 + (5-8)^2 + (7-9)^2 + (7-9)^2 + (6-9)^2} = 4.69$$

$$d_{18} = \sqrt{(7-8)^2 + (7-8)^2 + (8-9)^2 + (7-9)^2 + (8-9)^2} = 2.83$$

$$d_{19} = \sqrt{(8-8)^2 + (7-8)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2} = 1$$

Maka, diperoleh data centroid baru untuk iterasi 1 yang dilanjutkan ke pencarian data iterasi 2.

Tabel 3. Hasil Centroid Iterasi 1

Data ke-i	Cluster				
	Administrasi	Kesehatan Awal	Psikologi	Wawancara	Kesehatan Akhir
4	7.273	7.364	8.273	7.455	8
10	5.579	5.842	6.526	7.105	6.316

Sehingga nilai posisi *centroid* yang baru dengan menghitung nilai rata-rata

dari suatu data pada *centroid* yang sama, adalah sebagai berikut:

$$C_{11} = \frac{(7+8+7+8+7+6+7+7+6+7+7+8+7+8)}{11} = 7.273$$

$$C_{12} = \frac{(7+8+7+8+7+8+7+7+7+7+7+8+7+8)}{11} = 7.364$$

$$C_{13} = \frac{(7+8+7+8+7+7+7+7+6+7+7+8+7+8)}{11} = 8.273$$

$$C_{14} = \frac{(7+8+7+8+7+6+7+7+6+7+7+8+7+8)}{11} = 7.455$$

$$C_{15} = \frac{(7+8+7+8+7+6+7+7+6+7+7+8+7+8)}{11} = 8.182$$

V. SIMPULAN

Di dalam kasus ini, hasil akurasi *clustering* dengan optimasi *centroid* awal dengan menggunakan metode K-Means mengunakan *cosine similarity* memiliki signifikan 86.67%, sedangkan menggunakan K-Means dengan cosine similarity tidak jauh berbeda yaitu sebesar 89.7%, maka dari itu hasilnya cukup baik.

DAFTAR PUSTAKA

Agusta, Y. 2007. K-Means Penerapan, Permasalahan dan Metode Terkait. Denpasar, Bali: Jurnal Sistem dan Informatika Vol.3, pp : 47-60.

Budiman, I. 2012. Data Clustering Menggunakan Metodologi CRISP-DM Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma. Tesis. Universitas Diponegoro.

Huliman. 2013. Analisis Akurasi Algoritma Pohon Keputusan dan K-Nearest Neighbor (KNN). Tesis.Universitas Sumatera Utara.

Larose Daniel,T .2005. Discovering knowledge in data : an introduction to data mining, John Wiley & Sons, Inc.

Mirza, M. 2008. Mengenal Diabetes Melitus. Kata Hati. Yogyakarta.

Nugraheni, Y. 2011. Data Mining degan Metode Fuzzy Untuk Customer Relationship Management (CRM) pada Perusahaan Retail. Universitas Udayana.

Nurjayanti B. 2011. Identifikasi shorea menggunakan K-Nearest Neighbor berdasarkan karakteristik morfologi daun. Skripsi. Institut Pertanian Bogor.

Ong, J. O. 2013. Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University(12):10-20.

Oscar Ong, J .2013. Implementasi Algoritma k-means Clustering untuk no. 1, pp. Menentukan Strategi Marketing President University. Jurnal Ilmiah Teknik Industri vol. 12,10-13.

Paulanda, Z. 2012. Model Profil Mahasiswa Yang Potensisal Drop Out Menggunakan Teknik Kernel-K-Mean Clustering Dan Decision Tree. Tesis. Universitas Sumatera Utara. 2013.

Rismawan, T & Kusumadewi, S. 2008. Aplikasi K-Means Untuk Pengelompokkan Mahasiswa Berdasarkan Nilai Body Mass Index (BMI) & Ukuran Kerangka, SNATI. Yogyakarta.

Santosa, B. 2007. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, Teori dan Aplikasi. Graha Ilmu. Yogyakarta.

Soegondo, S. 2004. Penatalaksanaan Diabetes Mellitus Terpadu. FKUI. Jakarta.

Soraya, Y. 2011. Perbandingan Kinerja Metode Single Linkage, Metode Complete Linkage dan Metode K-Means dalam Analisis Cluster. Universitas Negeri Semarang.

Usino, W., Prabuwono. A. A., Allehaibi. K. H. S., Barmantoro. A., Hasniaty. A., & Amaldi. Wahyu. 2019. Document Similarity Detection using K-Means and Cosine Distance. International Journal of Advanced Computer Science and Applications. Vol. 10, No. 2, pp. 165-170.

Xiong, C. 2017. Using K-Means Clustering and Similarity Measur to Deal with Missing Rating in Collaborative Filtering Recommencation Systems. Graduate Programin Information System and Technology. Ontario: York University Toronto.