Penerapan Seleksi Atribut Berdasarkan Koefisien Variansi dan Korelasi untuk Inisialisasi Pusat Awal Klaster pada Algoritma K-Means dalam Pemetaan *E-Government* Tahun 2016

Ivon Dewi Apriliyaningsih[1*], Deden Istiawan[2]
[1,2]Statistika, Akademi Statistika (AIS) Muhammadiyah Semarang
*Email: iphon.dewi19@gmail.com

Keywords:
Klastering; K-means; metode a new algorithm erisoglu; calis; & sakallioglu

Abstrak

Salah satu metode yang terkenal dalam data mining adalah klastering. Klastering adalah metode pengelompokan data sehingga setiap klaster berisi data yang semirip mungkin dan berbeda antar klaster. K-means merupakan salah satu algoritma yang sering digunakan dalam klastering, karena algoritmanya mudah, bisa digunakan untuk menangani jumlah data yang besar dan waktu komputasinya relatif singkat. Namun K-means sangat sensitif terhadap inisialisasi pusat awal klaster, hasil klaster berupa solusi yang bersifat lokal optimal. Metode a new algorithm erisoglu, calis, & sakallioglu didasarkan pada pemilihan dua dari n atribut yaitu atribut pertama dan atribut kedua. Atribut pertama dihitung menggunakan koefisien variansi, sedangkan atribut kedua dihitung menggunakan koefisien korelasi. Kemudian menghitung rata-rata kedua atribut untuk menjadi pusat klaster. Hasil dari implementasi algoritma K-means dengan metode a new algorithm erisoglu, calis, & sakallioglu pada survei E-Government Perserikatan Bangsa-Bangsa (PBB) 2016. Setelah di evaluasi dengan Davies Bouldin Index (DBI) dapat dibuktikan nilai DBI metode a new algorithm erisoglu, calis, & sakallioglu lebih kecil dan waktu komputasi lebih efisien. Hal ini menunjukan bahwa metode a new algorithm erisoglu, calis, & sakallioglu bekerja lebih baik untuk mengatasi inisialisasi pusat awal klaster pada algoritma K-means.

1. PENDAHULUAN

Survei *E-Government* merupakan layanan publik teknologi informasi yang berisi database tingkat pembangunan 192 negara anggota PBB. Data mining belakangan ini sangat terkenal dalam dunia informasi. Data mining adalah integrasi dari berbagai disiplin ilmu, database, kecerdasan buatan, statistiak dll. Data mining terbagi beberapa kelompok berdasarkan tugas yang dilakukan yaitu model prediksi, analisis klaster, analisis asosiasi, dan deteksi anomali [1]. Namun yang paling terkenal yaitu klastering.

Klastering merupakan metode pengelompokan, dalam satu klaster berisi data semirip mungkin dan berbeda antar klaster. Klastering bersifat tanpa arah (*unsepervised learning)* yaitu tanpa variabel y. Analisis klaster berbasis partisi terdiri dari K-means, K-Harmonic Means, K-Modes, Fuzzy C-means, tetapi yang sangat terkenal adalah algoritma K-means [2].

Algoritma K-means sangat luas penggunaannya karena K-means merupakan

algoritma sederhana, mudah diimplementasikan, bisa digunakan untuk menangani jumlah data yang besar dan relatif singkat waktunya [3]. Namun algoritma K-means mempunyai kelemahan terhadap inisialisasi pusat awal klaster. Hasil klaster yang terbentuk sangatlah tergantung pada inisialisasi pusat awal klaster yang diberikan. Hal ini menyebabkan hasil klasternya berupa solusi yang sifatnya lokal optimal [4].

Beberapa penelitian pernah dilakukan untuk mengatasi inisialisasi pusat awal klaster dan metode yang digunakan. Penelitian yang dilakukan oleh (Gonzalez, 1985) yaitu mengusulkan metode *maximin initialization.* Pilih pusat awal klaster pertama C_1 secara acak. Pusat awal klaster selanjutnya atau C_i dihitung menggunkan jarak *Euclidean*. Hitung sampai dengan jumlah klaster terpenuhi. Metode tersebut memiliki waktu komputasi terbaik dan menghasilkan solusi optimal dengan nilai fungsi objektif dua kali dari solusi lokal optimal [5].

Penelitian yang dilakukan oleh (Deelers & Auwatanamongkol, 2007) mengusulkan sebuah algoritma partisi data untuk menghitung inisialisai pusat awal klaster. Partisi data mencoba membagi ruang data kedalam sel kecil atau kelompok, yang mana jarak interklaster sebesar mungkin dan jarak intraklaster sekecil mungkin. Sel dipartisi satu persatu sampai jumlah sel samadengan jumlah klaster k yang telah ditetapkan, dan pusat-pusat sel k menjadi awal pusat klaster untuk K-means. Hasil percobaan menunjukan bahwa algoritma partisi data bekerja lebih baik dibandingkan dengan inisialisasi pusat klaster secara acak dari sebagian kasus eksperimental dan dapat mengurangi waktu *running* algoritma K-means untuk dataset yang besar [6].

Penelitian yang dilakukan oleh (Naiggolan, 2014) mengusulkan sebuah metode inisialisasi pusat awal klaster yaitu *Modifield K-means clustering* berbasis *Sum of Squared Error* (SSE). Pertama menentukan jumlah iterasi, kemudian bangkitkan pusat awal klaster secara. Menghitung nilai SSE, nilai SSE terendah yang akan dipilih untuk dibangkitkan pada iterasi selanjutnya. Kriteria berhenti apabila iterasi telah melakukan pencarian SSE dengan jumlah iterasi yang sudah ditentukan. Nilai SSE yang paling minimum merupakan pusat awal klaster yang paling optimum. Hasil percobaan menunjukan *Modifield K-means clustering* berbasis *Sum of Squared Error* (SSE) lebih efektif untuk mencari pusat awal klaster dibanding dengan algoritma K-means acak [7].

Pada Penelitian ini, mengusulkan metode a new algorithm erisoglu, calis, & sakallioglu. Metode tersebut didasarkan pada pemilihan dua dari *n* atribut yaitu atribut pertama dan atribut kedua. Atribut pertama dihitung menggunakan koefisien variansi. Atribut kedua dihitung menggunakan koefisien korelasi antar atribut. Menghitung rata-rata dari kedua atribut untuk menjadi pusat klaster [8]. Metode a new algorithm erisoglu, calis, & sakallioglu yang diharapkan dapat meningkatkan kinerja algoritma K-means .

2. METODE

2.1. Data

Data yang digunakan dalam penelitian ini adalah data *United Nations on E-Government Survey* 2016 yaitu data *E-Government Development Index* (EGDI). Data tersebut merupakan data kuantitatif dengan jumlah 192 data dan terdiri dari 3 atribut. Adapun atribut-atribut yang digunakan dalam penelitian ini terdapat pada Tabel 3.1 sebagai berikut:Atribut Dari Data *E-Government Development Index* (EGDI).

Tabel 1. Atribut Dari Data *E-Government Development Index* (EGDI)

Atribut	Keterangan	Skala Data
X1	*Online Service Component*	Interval

ISSN 2407-9189

X2	*Telecom Infrastructure Component*	Interval
X3	*Human Capital Component*	Interval

2.2. Metode a new algorithm erisoglu, calis, & sakallioglu

Metode pusat awal klaster yang diusulkan berdasarkan pada penelitian-penelitian terdahulu oleh (Erisoglu et al., 2011). Metode ini di dasarkan pada pemilihan dua dari n atribut yaitu atribut pertama dan atribut. Metode tersebut dimulai dengan menghitung nilai absolut dari koefisien variasi, yang memiliki persamaan sebagai berikut:

$$cv_i = \left|\frac{s(x_i)}{\bar{x}_i}\right| \quad i = 1,2,3,\dots n$$

dengan:

cv_i = koefisien variansi
$s(x_i)$ = standar deviasi
$\bar{x}_i$ = rata-rata

Atribut pertama dipilih berdasarkan atribut yang mempunyai koefisien variasi tertinggi. Kemudian untuk mencari atribut kedua menggunakan koefisien korelasi, dengan cara menggunakan atribut yang mempunyai nilai koefisien variansi tertinggi, kemudian menghitung koefisien korelasi antar atribut menggunakan persamaan sebagai berikut:

$$r = \frac{n\sum_{i=1}^{n} xy - (\sum_{i=1}^{n} x)(\sum_{i=1}^{n} y)}{\sqrt{(n\sum_{i=1}^{n}(x)^2 - (\sum_{i=1}^{n} x)^2)(n\sum_{i=1}^{n}(y)^2 - (\sum_{i=1}^{n} y)^2}}$$

dengan:

n = jumlah titik pasang (x,y)
x = nilai atribut x
y = nilai atribut y

Atribut kedua dipilih berdasarkan nilai koefisien korelasi terkecil. Menghitung nilai rata-rata atribut pertama dan atribut kedua untuk menjadi pusat klaster.

$$m = [\bar{x}_I \quad \bar{x}_{II}]$$

Dengan $\bar{x}_i$ adalah rata-rata dari atribut pertama, $\bar{x}_{ii}$ adalah *defined similarly.*

Jarak *euclidean* digunakan untuk mencari jarak data dengan pusat awal klaster, persamaanya adalah sebagai berikut:

$$d_{im} = \sqrt{(x_{iI} - \bar{x}_I)^2 + (x_{iII} - \bar{x}_{II})^2} \quad i = 1,2,\dots,n$$

dengan:

d_{im} = jarak *euclidean*
x_{iI} = data ke- i
$\bar{x}_I$ = rata-rata atribut pertama
$\bar{x}_{II}$ = rata-rata atribut kedua

Dari hasil persamaan di atas, hasil perhitungan yang mempunyai nilai tertinggi dipilih sebagai kandidat pertama pusat awal klaster yaitu C_1. Selanjutnya menghitung data dengan pusat awal klaster untuk C_2, persamaanya sebagai berikut:

$$d_{ic1} = \sqrt{(x_{iI} - \bar{x}_{c1I})^2 + (x_{iII} - \bar{x}_{c1II})^2} \quad i = 1,2,\dots,n$$

dengan:

d_{ic1} = jarak *euclidean* untuk kandidat c1
x_{iI} = data ke- i
$\bar{x}_{c1I}$ = kandidat c1 untuk atribut pertama
$\bar{x}_{c1II}$ = kandidat c1 untuk atribut kedua

3.6

Dari persamaan di atas, hasil perhitungan yang mempunyai nilai tertinggi dipilih sebagai kandidat pusat awal klaster yang kedua yaitu C_2.

Untuk memilih kandidat pusat awal klaster selanjutnya yaitu C_r, menggunakan d_{icr} (dimana r adalah langkah-langkah iterasi) yaitu menghitung

jarak data dan C_{r-1}. Sd_{ir} merupakan penjumlahan dari jarak d_{icr}. Misal sd_{i3} persamaanya sebagai berikut:

$$Sd_{i3} = d_{ic_1} + d_{ic_2} \quad i = 1,2,\dots,n$$

dengan:

Sd_{i3} = jarak *euclidean*
d_{ic1} = jarak *euclidean* untuk kandidat c1
d_{ic2} = jarak *euclidean* untuk kandidat c2

Akumulasi diatas dapat menghindari jarak paling dekat antara kandidat pusat awal klaster dengan kandidat pusat awal klaster yang lainya. Selanjutnya hasil perhitungan yang mempunyai nilai terbesar akan dijadikan sebagai kandidat pusat awal klaster ketiga yaitu C_3. Sehingga didapatkan nilai pusat awal klaster yaitu C_1, C_2 dan C_3. Kemudian hasil dari eksperimen di jadikan sebagai pusat awal klaster algoritma K-means.

2.3. Algoritma K-means

Algoritma K-means merupakan algoritma pengelompokan yang melakukan partisi data ke dalam sejumlah k klaster yang sudah ditetapkan di awal [9]. Algoritma K-means sederhana untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaanya dalam praktek. Algoritma K-means menjadi salah satu algoritma yang paling penting dalam bidang data mining.

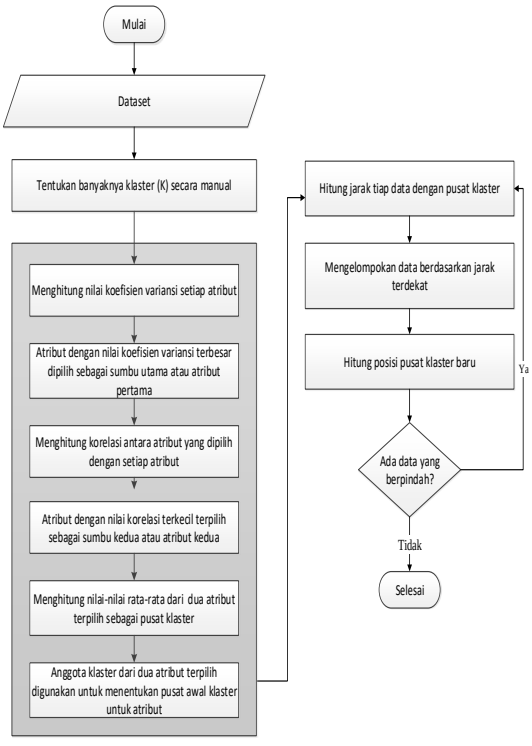
Langkah-langkah Algoritma K-means:

2.3.1. Tentukan k sebagai jumlah klaster yang diinginkan.
2.3.2. Bangkitkan inisialisasi pusat awal klaster secara acak.
2.3.3. Menghitung jarak antara data dengan pusat awal klaster dan kelompokkan data berdasarkan klaster yang diikuti.
2.3.4. Hitung posisi pusat awal klaster baru.
2.3.5. Ulangi langkah 3 sampai 5 hingga kondisi konvergen terpenuhi, yaitu tidak ada data yang berpindah klaster.

Inisialisasi pusat awal klaster juga mempengaruhi hasil klastering. Sifat ini menjadi karakteristik alami algoritma K-means yang dapat mengakibatkan hasil klaster yang didapatkan pada percobaan berbeda mendapatkan hasil berbeda. Kondisi seperti ini dikenal dengan solusi yang lokal optimal, yang artinya algoritma K-means sangat sensitif terhadap inisialisasi pusat awal klaster. Dengan kata lain, inisialisasi pusat awal klaster yang berbeda dapat mengakibatkan hasil klaster yang berbeda, meskipun dataset yang digunakan adalah sama.

2.4. Metode Evaluasi

Metode evaluasi klaster yang digunakan dalam penelitian ini adalah *Davies Bouldin Index* (DBI). *Davies Bouldin Index* merupakan salah satu metode evaluasi internal yang mengukur evaluasi klaster pada suatu metode pengelompokkan. Semakin kecil nilai DBI yang diperoleh (non-negatif ≥ 0), maka semakin baik klaster yang diperoleh dari pengelompokan algoritma K-means yang digunakan [10].



ISSN 2407-9189

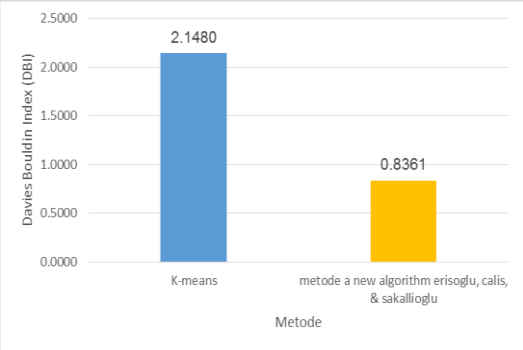
Gambar 1. Tahapan Penelitian

3. HASIL DAN PEMBAHASAN

Eksperimen dalam penelitian ini menggunakan program *Microsoft Excel* dan SPSS dengan data *E-Government Development Index* (EGDI).

Pada bagian ini, pembahasan tentang kinerja metode a new algorithm erisoglu, calis, & sakallioglu, dengan membandingan hasil evaluasi klastering menggunakan *Davies Bouldin Index* (DBI) dan jumlah iterasi algoritma K-means dengan inisialisasi pusat awal klaster metode a new algorithm erisoglu, calis, & sakallioglu dan algoritma K-means. Hasil terbaik dari perbandingan inisialisasi pusat awal klaster algoritma K-means dalam pemetaan *E-Government* tahun 2016.

3.1. Perbandingan hasil evaluasi klastering menggunakan *Davies Bouldin Index* (DBI) ditunjukan pada Gambar 2.

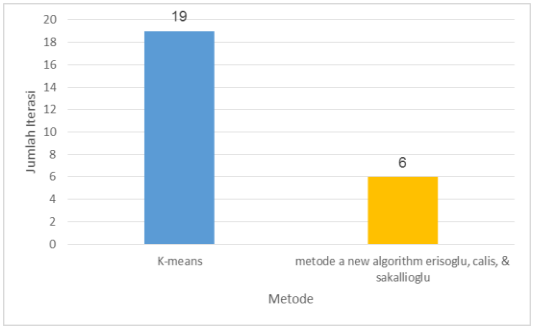


Gambar 2. Perbandingan hasil evaluasi klastering menggunakan *Davies Bouldin Index* (DBI)

Hasil perhitungan *Davies Bouldin Index* (DBI) diperoleh, yaitu 2.1480 untuk inisialisasi pusat awal klaster algoritma K-means dan 0.8361 metode a new algorithm erisoglu, calis, & sakallioglu. Semakin kecil nilai DBI yang diperoleh (non-negatif ≥ 0), maka semakin baik klaster yang diperoleh dari pengelompokan algoritma K-means yang digunakan. Sehingga dapat disimpulkan metode a new algorithm erisoglu, calis, & sakallioglu lebih baik, dibanding dengan inisialisasi pusat awal klaster algoritma K-means.

3.2. Jumlah Iterasi Pada Algoritma K-means

Jumlah iterasi algoritma K-means dengan inisilisasi pusat klaster algoritma K-means biasa dan metode a new algorithm erisoglu, calis, & sakallioglu dapat ditunjukan pada Gambar 3.



Gambar 3. Perbandingan Jumlah Iterasi pada algoritma K-means

Iterasi dalam algoritma K-means berhenti jika anggota klaster yang diikuti tidak berpidah klaster atau data sudah konvergen. Dapat dilihat jumlah iterasi untuk inisalisasi pusat awal klaster algoritma K-means, yaitu iterasi ke-19 dan ke-6 untuk inisialisasi pusat awal klaster metode a new algorithm erisoglu, calis, & sakallioglu. Semakin sedikit jumlah iterasi, maka semakin cepat waktu komputasi algoritma K-means dalam proses klastering. Sehingga dapat disimpulkan bahwa metode a new algorithm erisoglu, calis, & sakallioglu lebih baik dibanding dengan algoritma K-means.

3.3. Hasil klastering Klastering algoritma K-means Pada pemetaan *E-Government Survey* 2016

Tabel 2. Pusat awal klasteryang Terbentuk

Klaster	X1	X2	X3
1	0.7909	0.6850	0.8474
2	0.2057	0.1356	0.3931
3	0.4986	0.3931	0.7036

- Klaster 1 yaitu klaster negara maju sebanyak 49 negara anggota PBB
- Klaster 2 yaitu klaster negara kurang berkembang sebanyak 72 negara anggota PBB

- Klaster 3 yaitu klaster negara berkembang sebanyak 71 negara anggota PBB.

4. KESIMPULAN

Kesimpulan yang diperoleh berdasarkan pengujian adalah sebagai berikut:

1. Berdasarkan Hasil evaluasi *Davies Bouldin Index* (DBI) metode a new algorithm erisoglu, calis, & sakallioglu lebih baik yaitu memiliki nilai DBI lebih kecil 0.8361, sedangkan DBI algoritma K-means biasa 2.1480.
2. Jumlah iterasi metode a new algorithm erisoglu, calis, & sakallioglu berhenti pada iterasi ke-6, sedangkan algoritma K-means biasa berhenti pada iterasi ke-19. Hal ini menunjukan waktu komputasi metode a new algorithm erisoglu, calis, & sakallioglu lebih efisien.
3. Secara keseluruhan metode a new algorithm erisoglu, calis, & sakallioglu meningkatkan kinerja algoritma K-means dan mudah diimplementasikan untuk mengatasi masalah inisialisasi pusat awal klaster pada algoritma K-means.

REFERENSI

[1] Prasetyo, E. *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlap*. Yogjakarta: Penerbit Andi. 2014.

[2] Andriane, Guilherme, S., Felipe, C., Veronica, G., & Carvalho. (2015). Combining K-Means and K-Harmonic with Fish School Search Algorithm for Data Clustering Task on Graphics Processing Units. *University Estadual Paulista, Brazil*. 2015. Available from: https://doi.org/10.1016/j.asoc.2015.12.032

[3] Widiartha, I. M., Arifin, A. Z., & Yuniarti, A. Adaptive Data Clustering Method Based on Artifical Bee Colony and K-Harmonic Means. *Universitas Udayana, Bali*. 2012. *6*:129–137.

[4] Alfina, T., Santosa, B., & Barakbah, R. Analisa Perbandingan Metode Hierarchical Clustering , K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS), *1*. 2012.

[5] Gonzalez, F. Clustering to Minimize the Maximum Intercluster Distance. 1985; *38*; 293–306

[6] Deelers, S., & Auwatanamongkol, S. (2007). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance. 2007; *1*(11): 518–523.

[7] Naiggolan, R. *Algoritma Modifiend K-Means Clustering Pada Penentuan Cluster Centre Berbasis Sum Of Squared Error (SSE)*. Universitas Sumatera Utara. 2014.

[8] Erisoglu, M., Calis, N., & Sakallioglu, S. A new algorithm for initial cluster centers in k-means algorithm. *Pattern Recognition Letters*. 2011; *32*(14): 1701–1705. Available from: https://doi.org/10.1016/j.patrec.2011.07.011

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

[10] Alith Fajar Muhammad. Klasterisasi Proses Seleksi Pemain Menggunakan Algoritma K-Means, 1–5. 2015.

ISSN 2407-9189