Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 11, November 2019, hlm. 10678-10683 http://j-ptiik.ub.ac.id

Optimasi Penentuan *Centroid* pada Algoritme *K-Means* Menggunakan Algoritme *Pillar* (Studi Kasus: Penyandang Masalah Kesejahteraan Sosial di Provinsi Jawa Timur)

Alan Primandana[1], Sigit Adinugroho[2], Candra Dewi[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]alanprimandana@gmail.com, [2]sigit.adinu@ub.ac.id, [3]dewi_candra@ub.ac.id

Abstrak

Metode *k-means clustering* merupakan metode pengelompokan *non hierarchy* yang mengelompokkan data ke beberapa pusat klaster (*centroid*). Kesederhanaan metode *k*-means banyak digunakan diberbagai bidang karena memiliki beberapa keunggulan yaitu mudah diimplementasikan dan memiliki tingkat ketelitian yang cukup tinggi terhadap ukuran objek sehingga metode ini relatif lebih terukur dan efisien. Akan tetapi algoritme *k-means* awal perhitungan menggunakan nilai C (centroid) yang secara acak akan menyebabkan hasil yang acak pula. Ketergantungan pada nilai C (centroid) membuat akurasi pada algoritme *k-means* kurang maksimal. Hasil yang dari perhitungan *k-means* seringkali didapatkan dengan melakukan percobaan beberapa kali dan cenderung menghasilkan klaster yang berbeda. Tapi dalam mendapatkan hasil yang lebih baik, sulitnya menentukan batasan eksperimen. Penentuan titik pusat klaster secara acak menyebabkan metode *k-means* belum mampu mendapatkan hasil pengelompokan terbaik. Pada penelitian ini menjabarkan algoritma yang juga digunakan untuk mengoptimalkan pemilihan titik pusat awal pada metode *k*-means yaitu algoritma *pillar*. Algoritma ini merupakan penentuan posisi centroid awal dengan menghitung jarak akumulasi metrik antara setiap data dengan semua centroid sebelumnya. Pemilihan titik ditentukan oleh titik data yang memiliki jarak maksimum. Adapun penelitian ini melakukan penentuan *centroid* menggunakan algoritme *Pillar* kemudian hasil dari algoritme tersebut digunakan untuk titik pusat klaster pada algoritme *k-means*. Pada setiap klaster algoritme *pillar* mampu mendapatkan nilai *Sum of Squeared Error* (SSE) lebih baik dibandingkan dengan *centroid* acak dibuktikan dengan menurunnya nilai SSE.

Kata kunci: *pengelompokan, k-means, pillar*

Abstract

The k-means clustering method is a non-hierarchical grouping method that groups data into several centroid centers. The simplicity of the k-means method is widely used in various fields because it has several advantages, namely it is easy to implement and has a high level of accuracy of the size of the object so that this method is relatively more measurable and efficient. However, the initial k-means algorithm calculates using a C (centroid) value that randomly causes random results. Dependence on C (centroid) values makes the accuracy of the k-means algorithm less than optimal. The results of k-means calculations are often obtained by experimenting several times and tend to produce different clusters. But in getting better results, it is difficult to determine the limits of an experiment. The random determination of cluster centers causes the k-means method has not been able to get the best grouping results. In this study, we describe an algorithm that is also used to optimize the selection of the initial center point in the k-means method, the pillar algorithm. This algorithm is an initial centroid determination by calculating the distance of metric accumulation between each data and all previous centroids. The choice of points is determined by data points that have a maximum distance. This research determines centroid using the Pillar algorithm and the results of the algorithm are used for the cluster's focal point on the k-means algorithm. In each cluster pillar algorithm is able to get the value of Sum of Squeared Error (SSE) better than random centroids as evidenced by the decreasing value of SSE.

Keywords: *clustering, k-means, pillar*

1. PENDAHULUAN

Penyandang Masalah Kesejahteraan Sosial atau disingkat PMKS adalah seseorang, keluarga, bahkan sekelompok masyarakat yang tidak dapat melaksanakan kegiatan sosialnya yang disebabkan oleh suatu hambatan, kesulitan, atau juga gangguan. Hal Ini cukup berpengaruh dalam kebutuhan hidupnya secara jasmani, rohani, maupun sosial secara memadai tidak dapat tercukupi (Ariyanto & Utami, 2016). Kesejateraan sosial harus di evaluasi karena sangat berpengaruh untuk membangun perekonomian dan stabilitas suatu pemerintahan (Hafiludien & Istiawan, 2018). Maka dari itu perlu adanya pengelompokan daerah berdasarkan penyandang masalah kesejahteraan sosial (PMKS) dengan tujuan untuk mendapatkan gambaran daerah berdasarkan PMKS sehingga pemerintah dapat menentukan kebijakan dan dapat menentukan sasaran kebijakan tersebut (Hafiludien & Istiawan, 2018).

Metode *k-means clustering* merupakan metode klasterisasi *non hieracrchy* yang mengelompokkan data ke beberapa pusat klaster (centroid) terdekat dengan data. Nilai akhir dari metode *k-means* sendiri adalah memaksimalkan kemiripan data di dalam satu klaster dan meminimalkan kemiripan data antar klaster. Ukuran kemiripan dalam klaster dengan fungsi jarak yang berarti jarak yang terdekat merupakan data yang mirip dengan titik klaster (centroid) (Asroni & Adrian, 2015). Kesederhanaan metode *k*-means banyak digunakan diberbagai bidang karena memiliki beberapa keunggulan yaitu mudah diimplementasikan dan memiliki tingkat ketelitian yang cukup tinggi terhadap ukuran objek sehingga metode ini relatif lebih terukur dan efisien. Selain itu metode ini juga mudah dijalankan, relatif cepat dan mudah beradaptasi (Purnamaningsih, et al., 2014). Akan tetapi algoritme *k-means* awal perhitungan menggunakan nilai C (centroid) yang secara acak akan menyebabkan hasil yang acak pula. Ketergantungan pada nilai C (centroid) membuat akurasi pada algoritme *k-means* kurang maksimal. Bahkan dapat dilakukan banyak iterasi jika nilai C yang ditentukan tidak tepat (Fikri, et al., 2017). Menurut Pratama dan Harjoko (2015) dalam beberapa literatur didapatkan bahwa penentuan titik pusat klaster sangat mempengaruhi terhadap hasil yang diperoleh. Hasil yang dari perhitungan *k-means* seringkali didapatkan dengan melakukan percobaan beberapa kali dan cenderung menghasilkan klaster yang berbeda. Tapi dalam mendapatkan hasil yang lebih baik, sulitnya menentukan batasan eksperimen. Penentuan titik pusat klaster secara acak menyebabkan metode *k-means* belum mampu mendapatkan hasil pengelompokan terbaik. Alasan lain yang membuat metode *k-means* tidak mampu mencapai solusi global karena dalam penentuan titik pusat klaster yang baru pada setiap iterasi diperoleh dengan menghitung nilai tengah atau nilai *mean* data setiap klaster. Hal ini berdampak pada setiap iterasi penelusuran calon titik pusat klaster yang baru berada dalam wilayah sekitar titik pusat klaster awal. Keadaan ini membuat *k-means* terjebak dalam solusi lokal optima bukan global optima (Pratama & Harjoko, 2015).

Ketergantungan penentuan titik pusat awal dapat diatasi pada penelitian sebelumnya yang dilakukan oleh Pratama dan Harjoko (2015) melakukan optimasi terhadap *k-means* menggunakan algoritma *Invasive Weed Optimization* (IWO). Algoritma ini adalah algoritma yang terinspirasi pada proses kolonisasi rumput liar. Rumput liar memiliki sifat adaptif dalam penyebarannya terhadap perubahan lingkungan. Sifat acak dan adaptif inilah yang ditiru pada algoritma IWO dalam membangun koloni. Ide yang mendasari algoritma IWO dengan menyebarkan rumput sekaligus dua karakteristik berbeda yaitu area yang luas dan sempit (Pratama & Harjoko, 2015).

Pada penelitian ini menjabarkan algoritma yang juga digunakan untuk mengoptimalkan pemilihan titik pusat awal pada metode *k*-means yaitu algoritma *pillar*. Algoritma *pillar* terinspirasi oleh proses penentuan pilar pada rumah atau bangunan agar mendapatkan kestabilan. Penentuan pilar ditempatkan sejauh mungkin antara pilar satu dengan yang lainnya agar penyebaran tekanan atap dapat ditopang dengan baik. Algoritme ini merupakan penentuan posisi centroid awal dengan menghitung jarak akumulasi metrik antara setiap data dengan semua centroid sebelumnya. Pemilihan titik ditentukan oleh titik data yang memiliki jarak maksimum. Pendekatan ini dapat menemukan semua centroid terpisah sejauh mungkin antara centroid awal pada data yang tersebar (Barakbah & Kiyoki, 2009).

2. METODE PENELITIAN

Tujuan dalam penelitian ini adalah untuk mendapatkan informasi mengenai hasil dari optimasi penentuan *centroid* pada algoritme *pillar* dengan hasil penentuan *centroid* secara acak dan juga parameter optimal. Penelitian ini menggunakan data sekunder diperoleh dari BPS Provinsi Jawa Timur tahun 2016 pada kasus penyandang masalah kesejahteraan sosial di provinsi Jawa Timur. Proses awal pada penelitian ini dengan melakukan proses perhitungan menggunakan algoritme *pillar* yang bertujuan untuk mendapatkan titik pusat klaster atau *centroid*. Kemudian proses algoritme *k-means* dilakukan untuk mengelompokan data terhadap setiap titik pusat klaster yang diperoleh dari hasil algoritme *pillar*.

3. ALGORITME *K-MEANS*

Metode *K-Means* merupakan salah satu metode pengelompokan data dengan sistem *partitioned clustering* yang dipartisi ke dalam bentuk satu atau lebih *cluster* atau kelompok yang memiliki kemiripan data yang sama dan kemiripan data yang berbeda dikelompokan dalam *cluster* yang berbeda. Dengan kata lain, dalam satu *cluster* memiliki variasi yang minimal dan antar *cluster* memiliki variasi yang maksimal menurut Daniati dan Nugroho (2016) dalam penelitian Saksono (2018).

3.1 Langkah-langkah algoritme *pillar*

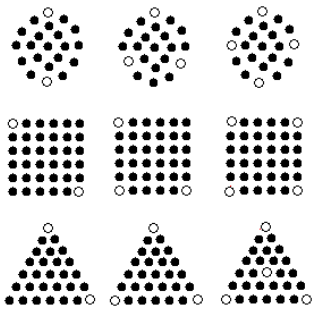
Terdapat beberapa langkah yang harus dilakukan dalam menyelasaikan maslah dengan metode *K-Means Clustering* (Saksono, 2018). Tahapan langkah tersebut dijelaskan sebagai berikut :

1. Menentukan jumlah dari k-*cluster* yang ingin dibangun.
2. Menentukan nilai *centroid* awal secara random sebanyak k-*cluster*.
3. Data input terhadap masing-masing *centroid* dihitung jaraknya menggunakan persamaan jarak *Eucledian*. Persamaan jarak Eucledian sebagai berikut :
   $$d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \quad (1)$$
4. Mengklasisfikasikan setiap data input yang memiliki jarak terdekat pada *centroid* .
5. Memperbarui nilai *centroid* berdasarkan dari nilai rata-rata *cluster* yang bersangkutan dengan rumus sebagai berikut :
   $$\mu_i(t+1) = \frac{1}{N_{sj}} \sum_{j \in S_j} x_j \quad (2)$$
   Keterangan :
   $\mu_i(t+1)$ = *centroid* pada iterasi ke $(t+1)$
   N_{sj} = Jumlah data pada *cluster* S_j
6. Melakukan iterasi dari langkah 3 ke langkah 5 hingga anggota disetiap *cluster* tidak terdapat perubahan.

4. ALGORITME *PILLAR*

Algoritme ini bertujuan untuk memaksimalkan penempatan *centroid* awal yang ditentukan di dalam ruang fitur dengan cara penempatan inisialisasi masing-masing *centroid* awal memiliki akumulasi jarak terjauh satu sama lain. Algoritme ini terinsipirasi oleh proses berpikir dalam menentukan serangkaian lokasi pilar dalam membuat rumah atau bangunan yang stabil. Yang mana penempatan dua, tiga, dan empat pilar dapat menahan distribusi tekanan atap bangunan (Barakbah & Kiyoki, 2009).

4.1 Konsep Dasar



Gambar 1 Ilustrasi penempatan lokasi pilar (titik putih)

Algoritma yang diusulkan dalam penelitian ini terinspirasi oleh proses berpikir untuk menentukan serangkaian lokasi pilar membangun sebuah rumah atau bangunan yang stabil. Pada gambar 1 mengilustrasikan penempatan dua, tiga, dan empat pilar, untuk menahan distribusi tekanan dari beberapa atap yang memiliki struktur berbeda terdiri dari titik-titik diskrit. Hal tersebut menjadi inspirasi untuk mendistribusikan pilar sejauh mungkin satu sama lain dalam penyebaran tekanan atap, pilar bisa tahan tekanan atap dan menstabilkan rumah atau bangunan.

4.2 Langkah-langkah algoritme *pillar*

Misalkan $X = \{xi \mid i = 1, \dots, n\}$ adalah data, k adalah jumlah *cluster*, $C = \{ci|i1, \dots, k\}$ adalah centroid awal, $SX \subseteq X$ menjadi identifikasi untuk X yang sudah di pilih dalam urutan proses, $DM = \{xi|i1, \dots, n\}$ adalah akumulasi metric jarak, $D = \{i|i1, \dots, n\}$ adalah metric jarak untuk setiap iterasi, dan m menjadi grand mean X.

1. Hitung m. m adalah titik tengah data atau rata-rata data.
2. Hitung jarak setiap data(x) terhadap m.
3. Menghitung nilai nmin.
   $nmin = \alpha \frac{n}{k}$ (3)
4. dmax = data dengan jarak terjauh dari m.
5. Tentukan nbdis yaitu batas lingkungan
   $nbdis = \beta . dmax$ (4)
6. Tentukan $i = 1$ sebagai penghitung untuk menentukan centroid awal ke-i
7. Menghitung nilai DM.
   $DM = DM + D$ (5)
8. Pilih calon centroid(ж) = data dengan nilai DM tertinggi
9. Menghitung nilai SX.
   $SX = SX \cup ж$ (6)
10. D sebagai jarak setiap data(x) ke ж.
11. Tentukan no = banyak data yang memenuhi $D \leq nbdis$
12. Menetapkan nilai DM(ж)
    $DM(ж) = 0$ (7)
13. Jika $no < nmin$, kembali ke langkah 8
14. Menetapkan nilai D(SX).
    $D(SX) = 0$ (8)
15. Menentukan nilai C.
    $C = C \cup ж$ (9)
16. Menambahkan iterasi $i = i + 1$
17. Jika $i \leq k$, kembali ke langkah 7
18. Selesai, dimana C adalah solusi sebagai centroid awal yang dioptimalkan.

5. HASIL

Pada penelitian ini hasil yang didapatkan dari pengujian parameter alpha dan beta terhadap beberapa klaster. Hasil penelitian juga didapatkan dari pengujian pengelompokan secara acak. Kedua hasil pengujian tersebut kemudian dibandingkan.

5.1 Banyak Tetangga (alpha)

Pengujian alpha dilakukan dengan cara merubah parameter alpha dan parameter beta tetap. Parameter alpha dirubah dari rentang 0,1 hingga 0,9. Pertama pengujian dilakukan dengan pengelompokan sebanyak 3 klaster. Parameter beta adalah 0,8 dapat dilihat pada tabel 1 pengujian alpha pertama.

Tabel 1 Pengujian banyak tetangga (alpha) pertama

Alpha	Nilai SSE
0.1	11604126.46
0.2	11604126.46
0.3	11604126.46
0.4	11604126.46
0.5	11604126.46
0.6	11604126.46
0.7	11604126.46
0.8	11604126.46
0.9	11604126.46

Hasil yang didapatkan dengan merubah parameter alpha dari 0,1 hingga 0,9 dan parameter beta 0,8 menunjukkan nilai SSE tetap yaitu sebesar 11604126. kemudian pengujian dilakukan dengan pengelompokan sebanyak 4 klaster dengan parameter beta adalah 0,1. Hasil pengujian pada jumlah klaster 4 dapat dilihat pada tabel 2 pengujian alpha kedua.

Tabel 2 pengujian banyak tetangga (alpha) kedua

Alpha	Nilai SSE
0.1	5125735.25
0.2	5125735.25
0.3	5125735.25
0.4	5125735.25
0.5	5125735.25
0.6	12353309.88
0.7	12353309.88
0.8	12353309.88
0.9	12353309.88

Nilai SSE yang didapatkan dari pengelompokan dengan jumlah klaster 4 mengalami perubahan. Namun variasi nilai SSE yang didapat hanya 2 dan perubahan nilai SSE yang didapat juga semakin tinggi.

5.2 Batas Lingkungan (Beta)

Pengujian beta dilakukan dengan cara merubah parameter beta dan parameter alpha tetap. Parameter beta dirubah dari rentang 0,1

hingga 0,9. Pertama pengujian beta dilakukan dengan pengelompokan sebanyak 3 klaster. Parameter alpha adalah 0,1. Hasil pengujian beta dengan alpha 0,1 dapat dilihat pada tabel 3.

Tabel 3 Pengujian batas lingkungan (beta) pertama

Beta	Nilai SSE
0.1	12959494
0.2	12013669
0.3	11604126
0.4	11604126
0.5	11604126
0.6	11604126
0.7	11604126
0.8	11604126
0.9	7896065

Pengujian beta pada alpha 0,1 mengalami perubahan nilai SSE. Perubahan nilai beta semakin tinggi maka nilai SSE semakin kecil. Pengujian dilakukan dengan pengelompokan sebanyak 3 klaster. Parameter alpha adalah 0,4. Hasil pengujian beta dengan alpha 0,4 dapat dilihat pada tabel 4.

Tabel 4 Pengujian batas lingkungan (beta) kedua

Beta	Nilai SSE
0.1	12959494.52
0.2	12959494.52
0.3	12013669.86
0.4	12013669.86
0.5	11604126.46
0.6	11604126.46
0.7	11604126.46
0.8	11604126.46
0.9	11604126.46

Hasil pengujian yang didapatkan dengan alpha tetap 0,4 yaitu sebesar 11604126,46. Pengujian beta pada alpha 0,4 mengalami perubahan nilai SSE dan beta semakin tinggi nilai SSE yang didapat semakin rendah.

5.3 Pengelompokan Optimasi *Pillar*

Pengujian terhadap setiap klaster dilakukan dengan nilai alpha dan beta terbaik. Pada pengujian sebelumnya nilai beta diperoleh hasil yang terbaik pada 0,9 dengan nilai alpha 0,1. Pengujian jumlah klaster diuji menggunakan nilai alpha 0,1 dan beta 0,9 pada beberapa klaster. Berikut pengujian terhadap beberapa jumlah klaster dapat dilihat pada tabel 5.

Tabel 5 Pengujian terhadap setiap klaster

Klaster	Nilai SSE
3	7896065
5	2814718
7	2046913
9	1427727
11	1177569
13	943280
14	644343
15	604712
16	496919
17	548196

Terlihat bahwa nilai SSE semakin kecil ketika jumlah klaster semakin banyak. Dapat dilihat nilai dari klaster 3 hingga klaster 16 mengalami penurunan dan mengalami kenaikan pada klaster 17 sehingga harus dihentikan.

5.4 Pengelompokan *K-Means*

Pengujian ini dilakukan dengan variasi klaster sebanyak 9 klaster dan 10 kali iterasi pada setiap klaster. Setiap iterasi pemilihan *centroid* secara acak dilakukan oleh sistem komputer. Berikut hasil dari pengujian *centroid* acak ditampilkan pada tabel 6.

Tabel 6 Pengujian dengan *centroid* acak

Percobaan	Nilai SSE			
	3	…	16	17
1	12959495	…	733284	886581
2	11640568	…	821535	710799
3	11658090	…	830325	466573
4	12119617	…	791668	710603
5	7806033	…	949337	543643
6	11545727	…	883133	738198
7	12119617	…	848940	651027
8	12959495	…	919721	874064
9	12959495	…	844072	603504
10	12959495	…	909322	582076

Percobaan	Nilai SSE			
	3	…	16	17
Rata-Rata	11872763	…	853134	676707

5.5 Perbandingan Hasil

Hasil pengujian yang diperoleh dari pengujian *pillar* dan *centroid* acak kemudian dibandingkan. Nilai perbandingan SSE *centroid* acak yang dibandingkan merupakan nilai SSE rata-rata dari 10 kali percobaan. Perbandingan *centroid acak* dan optimasi *pillar* dapat dilihat pada tabel 7.

Tabel 7 Perbandingan hasil

Klaster	Nilai SSE	
	pillar	acak
3	7896065	11872763
5	2814718	7500001
7	2046913	2894402
9	1427727	2291586
11	1177569	1236891
13	943280	1182520
14	644343	995163
15	604712	885031
16	496919	853134
17	548196	676707

6. KESIMPULAN

1. Pengelompokan daerah berdasarkan penyandang masalah kesejahteraan sosial dilakukan dengan menentukan centroid terlebih dahulu dengan algoritme pillar dan dilanjutkan dengan pembentukan kelompok dan cluster menggunakan k-means.
2. Berdasarkan hasil evaluasi, penggunaan algoritme pillar dapat meningkatkan kinerja dari algorime clustering k-means yang dibuktikan dengan menurunnya nilai SSE.

DAFTAR PUSTAKA

Ariyanto, J. T. & Utami, A. W., 2016. Sistem Pendukung Keputusan Penerima Bantuan Penyandandang Masalah Kesejahteraan Sosial Menggunakan Metode Weighted Product Studi Kasus di IPSM Kelurahan Kertajaya Kota Surabaya. *Jurnal Manajemen Informatika*, Volume 5, pp. 107-116.

Asroni & Adrian, R., 2015. Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik. *Jurnal Ilmiah Semesta Teknika*, Volume 18.

Barakbah, A. R. & Kiyoki, Y., 2009. A Pillar Algorithm for K-Means Optimization by Distance. *IEEE.*

Fikri, C. M., Agustin, F. E. M. & Mintarsih, F., 2017. Pengelompokan Kualitas Kerja Pegawai Menggunakan Algoritma K-MEANS++ dan COP-KMEANS Untuk Merencanakan Program Pemeliharaan Kesehatan Pegawai di PT. PLN P2B JB Depok. *Jurnal Pseudocode*, Volume IV.

Hafiludien, A. & Istiawan, D., 2018. Penerapan Algoritma Self Organizing Maps Untuk Pemetaan Penyandang Kesejahteraan Sosial (PMKS) di Provinsi Jawa Tengah Tahun 2016. *University Research Colloqium.*

Pratama, I. P. A. & Harjoko, A., 2015. Penerapan Algoritma Invasive Weed Optimnization untuk Penentuan Titik Pusat Klaster pada K-Means. *IJCCS*, Volume 9, pp. 65-76.

Purnamaningsih, C., Saptono, R. & Aziz, A., 2014. Pemanfaatan Metode K-Means Clustering dalam Penentuan Penjurusan Siswa SMA. *ITSMART*, Volume 3.

Saksono, 2018. *Rekomendasi Lokasi Wisata Kuliner Menggunakan Metode K-Means Clustering dan Simple Additive Term Weighting.* Malang: Universitas Brawijaya.