PENENTUAN PUSAT AWAL KLASTER ALGORITMA K-MEANS UNTUK PEMETAAN TINGKAT KESEJAHTERAAN DI PROVINSI JAWA TENGAH

DETERMINING OF INITIAL CLUSTER CENTER ON K-MEANS ALGORITHM FOR MAPPING IN POSPERITY DEGREE OF CENTRAL JAVA

Deden Istiawan
Akademi Statistika Muhammadiyah Semarang
Email: dedenistiawan@gmail.com

ABSTRAK

Kesejahteran mempunyai arti yang relatif, dinamis dan kuantitatif. Sampai saat ini rumusanya tidak pernah selesai karena akan terus berkembang seiring dengan perkembangan zaman. Kesejahteraan secara umum adalah suatu keadaan dimana segenap warga negara selalu berada dalam kondisi yang serba kecukupan dalam segala kebutuhannya. Kemiskinan di Provinsi Jawa Tengah masih berada di atas kemiskinan nasional. Pengelompokkan kemiskinan merupakan salah satu cara untuk mengidentifikasi karakteristik tingkat kesejahteraan rakyat pada tiap daerah agar dalam mengambil kebijakan dan strategi pembangunan tepat sasaran dan tepat guna. Algoritma K-means merupakan salah satu algoritma klastering yang paling sering digunakan untuk pengelompokan objek dikarena kemudahan dalam pengaplikasianya dan sangat efisien untuk mengelompokan data yang besar, namun algoritma K-means memiliki kelemahan pada pemilihan pusat awal klaster secara acak, sehingga menyebabkan kinerja algoritma K-means menurun. Pada penelitian ini diusulkan metode penentuan pusat awal klaster pada algoritma K-means. Hasil penelitian menunjukan bahwa metode yang diusulkan memiliki kinerja yang lebih baik daripada algoritma K-means Standar.

Kata Kunci: Algoritma K-means, Centroid, Kesejahteraan, Klaster

PENDAHULUAN

Kesejahteran mempunyai arti yang relatif, dinamis dan kuantitatif. Sampai saat ini rumusanya tidak pernah selesai karena akan terus berkembang seiring dengan perkembangan zaman. Kesejahteraan secara umum adalah suatu keadaan di mana segenap warga negara selalu berada dalam kondisi yang serba kecukupan dalam segala kebutuhannya (Roestam, 1993). Kemiskinan merupakan salah satu masalah dalam kesejahteraan yang harus dituntaskan. Potret kemiskinan di Jawa Tengah mencapai 13.58 persen ini tercatat lebih tinggi dari angka kemiskinan secara nasional. Pengelompokkan kemiskinan merupakan salah satu cara untuk mengidentifikasi karakteristik tingkat kesejahteraan rakyat pada tiap daerah agar dalam mengambil kebijakan dan strategi pembangunan tepat sasaran dan tepat guna.

Dalam beberapa tahun terkahir, terdapat beberapa penelitian yang mengusulkan algoritma klastering untuk pengelompokkan kabupaten/kota berdasarkan indikator kesejahteraan di Provinsi Jawa Tengah. Yulianto dan Hidayatullah (2014) mengusulkan metode hirarki menghasilkan tiga kelompok kabupaten/kota di Jawa Tengah berdasarkan indikator kesejahteraaan rakyat dengan variabel yang digunakan meliputi PDRB perkapita, kepadatan penduduk, penduduk miskin, jumlah angkatan kerja, pengeluaran riil perkapita yang disesuaikan, angka harapan hidup dan rata-rata lama sekolah (Yulianto & Hidayatulloh, 2014). Kelebihan metode hirarki adalah tidak membutuhkan informasi jumlah klaster (Basu & Murthy, 2015) dan juga penentuan nilai-nilai awal sehingga hasil klastering selalu sama (Hsu, Lu, & Lin, 2012), namun membutuhkan waktu komputasi yang tinggi pada *dataset* yang besar daripada metode partisi (Pimentel & de Souza, 2016).

Penelitian lain tentang pengelompokkan kesejahteraan dilakukan oleh Putriana (2015) di Provinsi Jawa Tangah. Pada penelitian ini dilakukan perbandingan algoritma K-means dan Metode Hirarki. Hasil penelitian menunjukkan bahwa algoritma K-means lebih unggul daripada Metode Hirarki dengan menghasilkan tiga klaster (Putrina, 2015). Hidayat *et al* (2017) juga melakukan perbandingan

antara algoritma K-means dan Fuzzy C-Means untuk pengelompokkan kemisikinan di Provinsi Jawa Tengah. Hasil penelitian juga menyebutkan bahwa algoritma K-means lebih baik daripada algoritma Fuzzy C-Means (Hidayat, Wasono, & Darsyah, 2017). Algoritma K-means sangat efisien untuk mengelompokan *dataset* yang besar (Fahad et al., 2014), kemudahan dalam pengaplikasiannya (Jain, 2010) dan metode yang efisien dalam hal komputasi (Cura, 2012) menjadi alasan utama popularitas K-means, meskipun telah diusulkan lebih dari 50 tahun yang lalu (Jain, 2010).

Hasil pengolompokan algoritma K-means bergantung pada pemilihan acak pusat awal klaster (Han, Kamber, & Pei, 2012). Untuk mendapatkan hasil klastering yang baik penentuan pusat awal klaster menjadi sangat penting untuk algoritma K-means. Menurut beberapa peneliti penentuan pusat awal klaster secara acak menyebabkan satu atau beberapa klaster kosong, hasil klastering merupakan solusi sub-optimal dan cepat konvergen ke dalam minimum lokal (Celebi, Kingravi, & Vela, 2013), M. C. Naldi (Naldi & Campello, 2014), (Xu, Ding, Liu, & Luo, 2015).

Pada penelitian ini akan diusulkan metode penentuan pusat awal klaster pada algoritma K-means dengan cara memilih atribut yang memiliki nilai standar deviasi tertinggi. Selanjutnya dari atribut yang terpilih diurutkan dari nilai terkecil ke nilai terbesar. Tahap selanjutnya partisi data menjadi n *subset* sesuai dengan jumlah klaster yang ditentukan, masing-masing *subset* dihitung nilai rata-ratanya. Kemudian nilai rata-rata dari masing-masing *subset* akan digunakan untuk penentuan pusat awal klaster. Tujuannya adalah untuk meningkatkan kinerja algoritma K-means yang mudah terjebak dalam kondisi minimun lokal yang disebabkan pemilihan pusat awal klaster secara acak. Hasil evaluasi dalam penelitian ini adalah Davis-Bouldin Index.

METODE

Algoritma K-MEANS

Algoritma K-means mengelompokan objek ke dalam beberapa kelompok atau klaster sehingga objek dalam satu klaster memiliki kemiripan yang tinggi, sedangkan antar klaster memiliki kemiripan yang sangat rendah. Algoritma K-means dimulai dengan menentukan jumlah klaster sebanyak k, kemudian membangkitkan k pusat klaster secara acak. Selanjutnya setiap objek akan dikelompokan berdasarkan jarak terdekat dengan pusat klaster, pusat klaster diperbaharui berdasarkan titik data dalam setiap klaster. Proses ini diulangi sampai kriteria konvergen terpenuhi. Berikut ini adalah tahapan dari algoritma K-means:

1. Menentukan nilai k sebagai jumlah klaster yang dibentuk.
2. Memilih k pusat klaster secara acak untuk menjadi pusat klaster awal.
3. Alokasikan semua data ke pusat klaster terdekat dengan matrik jarak.
4. Hitung kembali pusat klaster baru berdasarkan data yang mengikuti klaster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai atau tidak ada data yang berpindah dari satu klaster ke klaster yang lainnya.

Kemiripan antar data dapat diketahui dengan menghitung jarak antar tiap data dengan pusat klaster. Untuk kemiripan yang digunakan adalah jarak euclidean yang diformulasikan oleh persamaan berikut:

$$d(x, c) = \sqrt{\sum_{i=1}^{n}(x_i - C_k)^2}$$

Dimana $x = x_1, x_2, \dots, x_n$ dan $c = c_1, c_2, \dots, c_k$ pada tahap keempat, setiap representasi klaster direlokasi ke pusat klaster dengan rata-rata aritmatika dari setiap klaster. Hal ini jugalah yang menyebabkan algoritma ini sering disebut dengan *cluster mean* atau *cluster centroid* seperti nama yang dimiliki.

Metode Yang Diusulkan

Pada penelitian ini akan diusulkan sebuah metode dengan cara memilih atribut yang memiliki nilai standar deviasi terbesar. Standar deviasi merupakan ukuran penyebaran statistik yang digunakan untuk melihat sebaran data dari rata-rata hitungnya. Semakin besar nilai standar deviasi maka semakin

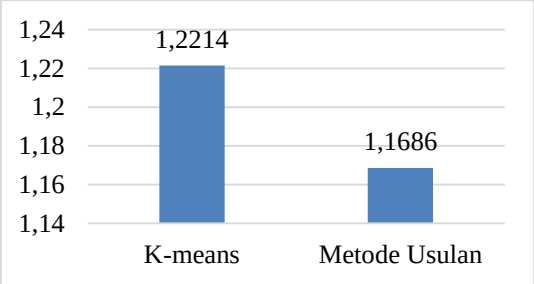
heterogen data tersebut. Berikut merupakan tahapan metode yang diusulkan pada penelitian ini dengan alur sebagai berikut:

1. Mengumpulkan dan menghapus data informasi pendukung terhadap dataset yang digunakan.
2. Menentukan banyaknya klaster (K) secara manual
3. Menghitung nilai standar deviasi setiap atribut.
4. Pilih atribut dengan nilai standar deviasi terbesar.
5. Paritisi data menjadi *n subset* sesuai jumlah klaster yang ditentukan
6. Nilai rata-rata dari masing-masing subset digunkan sebagai pusat awal klaster.

HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini merupakan data pemukhiran dari pengumpulan data PPLS yang digunakan oleh Tim Nasional Percepatan Penanggulangan Kemisnikan (TNP2K) berdasarkan status kesejahteraan 40% terendah di Provinsi Jawa Tengah dengan atribut jumlah penduduk, kepala rumah tangga (KRT) perempuan, anak tidak sekolah, individu yang cacat, individu yang memiliki penyakit kronis, pengangguran, sumber air minum tidak terlindungi, sumber penerangan tidak listrik, bahan bakar memasak menggunakan kayu bakar/arang/minyak tanah, fasilitas tempat buang air besar (BAB) tidak tersedia.

Pada penelitian ini kinerja algoritma K-means dievaluasi dengan menggunakan Davies-Bouldin Index. Hasil penelitian menunjukkan bahwa metode yang diusulkan dapat meningkatkan kinerja algoritma K-means daripada K-means standar. Gambar 1 menunjukan bahwa nilai *Davies-Bouldin Index* metode yang diusulkan lebih kecil. *Davies-Bouldin Index* mengukur rata-rata kesamaan antara setiap klaster dan satu klaster lagi yang paling mirip. Rata-rata yang paling rendah pada *Davies-Bouldin Index* menunjukan bahwa klaster yang terbentuk adalah kompak dan dapat terpisah pada klaster yang tepat.



Gambar 2 Diagram perbandingan Davies-Bouldin Index

Tabel 2 Hasil Pengelompokan Algoritma K-means

Klaster	Anggota Klaster
1	Kab. Wonosobo, Kab. Sukoharjo, Kab. Karanganyar, Kab. Sragen, Kab. Rembang, Kab. Kudus, Kab. Semarang, kab. Temanggung, Kab. Batang, Kota Magelang, Kota Surakarta, Kota Salatiga, Kota Semarang, Kota Pekalongan, Kota tegal
2	Kab. Banjarnegara, Kab. Purworejo, kab. Magelang. Kab. Boyolali, Kab. Wonogiri, Kab. Grobogan, Kab. Blora, Kab. Pati,

	Kab. Jepara, kab. Demak, Kab. Kendal, Kab. Pekalongan
3	Kab. Cilacap, Kab. Banyumas, Kab. Purbalingga, Kab. Kebumen, kab. Klaten, Kab. Pemalang, Kab. Tegal, Kab. Brebes

Hasil pengelompokan algoritma K-means dengan metode yang diusulan dapat dilihat pada Tabel 1 dimana kabupaten/kota dikelompokan menjadi tiga klaster. Ketiga klaster yang terbentuk dapat diuraikan karakteristik masing-masing kelompok bersadarkan atribut status kesejahteraan. Tabel 2 merupakan karakteristik dari masing-masing kelompok yang terbentuk.

Tabel 3 Karakteristik Klaster

Atribut	Klaster		
	1	2	3
At1	1.6158	3.0268	4.9302
At2	1.5537	3.7963	3.8924
At3	1.3425	2.9940	5.4918
At4	1.4841	3.4463	4.5479
At5	1.4651	3.4924	4.5145
At6	1.3975	2.9824	5.4060
At7	1.1749	2.6464	6.3274
At8	0.7737	2.6632	7.0545
At9	1.4634	3.7378	4.1495
At10	0.9409	2.9986	6.2379

Klaster pertama memiliki karakteristik jumlah penduduk terkecil, jumlah rumah tangga yang berkepala rumah tangga perempuan terkecil, jumlah anak yang tidak bersekolah terkecil, jumlah individu yang cacat terkecil, jumlah individu yang memilki penyakit kronis terkecil, jumlah pengangguran terkecil, jumlah rumah tangga yang memiliki sumber air minum tidak terlindungi terkecil, jumlah rumah tangga yang tidak memiliki sumber penerangan listrik terkecil, jumlah rumah tangga yang menggunakan kayu bakar/arang/minyak tanah terkecil dan jumlah rumah tangga yang tidak memiliki fasilitas jamban untuk BAB terkecil. Sehingga klaster ini merupakan klaster kabupaten/kota yang berkategori hampir miskin.

Klaster kedua memiliki karakteristik jumlah penduduk sedang, jumlah rumah tangga yang berkepala rumah tangga perempuan terbesar, jumlah anak yang tidak bersekolah sedang, jumlah individu yang cacat sedang, jumlah individu yang memilki penyakit kronis sedang, jumlah pengangguran sedang, jumlah rumah tangga yang memiliki sumber air minum tidak terlindungi sedang, jumlah rumah tangga yang tidak memiliki sumber penerangan listrik sedang, jumlah rumah tangga yang menggunakan kayu bakar/arang/minyak tanah sedang dan jumlah rumah tangga yang tidak memiliki fasilitas jamban untuk BAB sedang. Sehingga klaster ini merupakan klaster kabupaten/kota berkategori miskin.

Klaster Ketiga memiliki karakteristik jumlah penduduk terbesar, jumlah rumah tangga yang berkepala rumah tangga perempuan sedang, jumlah anak yang tidak bersekolah terbesar, jumlah individu yang cacat terbesar, jumlah individu yang memilki penyakit kronis terbesar, jumlah pengangguran terbesar, jumlah rumah tangga yang memiliki sumber air minum tidak terlindungi terbesar, jumlah rumah tangga yang tidak memiliki sumber penerangan listrik terbesar, jumlah rumah tangga yang menggunakan kayu bakar/arang/minyak tanah terbesar dan jumlah rumah tangga yang tidak memiliki.

KESIMPULAN

Hasil pengelompokan algoritma K-means dengan metode yang diusulkan menghasilkan tiga klaster. Klaster pertama berganggotakan 15 kabupaten/kota, klater kedua beranggotakan 12 Kabupaten dan klaster ketiga beranggotakan 8 kabupaten. Setalah dilakukan perbandingan kinerja algoritma, metode yang diusulkan mempunyai kinerja yang lebih baik daripada algoritma K-means standar.

DAFTAR PUSTAKA

Basu, T., & Murthy, C. A. (2015). A similarity assessment technique for effective grouping of documents. *Information Sciences*, *311*, 149–162. https://doi.org/10.1016/j.ins.2015.03.038

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

Cura, T. (2012). A particle swarm optimization approach to clustering. *Expert Systems with Applications*, *39*(1), 1582–1588. https://doi.org/10.1016/j.eswa.2011.07.123

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., … Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279. https://doi.org/10.1109/TETC.2014.2330519

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Third). Waltham: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Hidayat, R., Wasono, R., & Darsyah, M. Y. (2017). Pengelompokan Kabupaten Kota Di Jawa Tengah Menggunakan Metode K-Means dan Fuzzy C-Means. *Prosiding Seminar Nasional Pendidikan, Sains Dan Teknologi*, 240–250.

Hsu, F. M., Lu, L. P., & Lin, C. M. (2012). Segmenting customers by transaction data with concept hierarchy. *Expert Systems with Applications*, *39*(6), 6221–6228. https://doi.org/10.1016/j.eswa.2011.12.005

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Naldi, M. C., & Campello, R. J. G. B. (2014). Evolutionary k-means for distributed data sets. *Neurocomputing*, *127*, 30–42. https://doi.org/10.1016/j.neucom.2013.05.046

Pimentel, B. A., & de Souza, R. M. C. R. (2016). Multivariate Fuzzy C-Means algorithms with weighting. *Neurocomputing*, *174*, 946–965. https://doi.org/10.1016/j.neucom.2015.10.011

Putrina, U. (2015). *Metode Cluster Analysis untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Variabel yang Mempegaruhi Kemiskinan pada Tahun 2013*. Institut Sains & Teknologi AKPRIND.

Roestam, S. (1993). *Pembangunan nasional untuk kesejahteraan rakyat*. Jakarta: Kantor Kementrian Koordinator Bidang Kesejahteraan Rakyat Republik Indonesia.

Xu, Q., Ding, C., Liu, J., & Luo, B. (2015). PCA-guided search for K-means. *Pattern Recognition Letters*, *54*, 50–55. https://doi.org/10.1016/j.patrec.2014.11.017

Yulianto, S., & Hidayatulloh, K. H. (2014). Analisis Klaster untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal Statistika*, *2*(1), 56–63.