OPTIMASI PERFORMA *CLUSTER K-MEANS* MENGGUNAKAN *SUM OF SQUARED ERROR* (SSE)

[1]Rena Nainggolan, [2]Gortap Lumbantoruan
[1,2]Universitas Methodist Indonesia
[1]renanain99olan@gmail.com, [2]lumbantoruan.gortap@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol2No2.pp103-108

ABSTRAK

K-Means merupakan suatu algoritma pengklasteran yang cukup sederhana yang mempunyai kemampuan mengelempokkan data dalam jumlah yang cukup besar, mempartisi dataset kedalam beberapa klaster k. Algoritmanya cukup mudah untuk diimplementasikan dan dijalankan, relative cepat, dan efisien. Disi lain K-Means masih memiliki beberapa kelemahan, yaitu dalam menentukan jumlah cluster. Hasil *cluster* yang terbentuk dari metode *K-means* ini sangatlah tergantung pada inisiasi nilai pusat awal *cluster* yang diberikan. Hal ini menyebabkan hasil *cluster*nya berupa solusi yang sifatnya *local optimal*. Pada penelitian ini dilakukan untuk mengatasi kelemahan yang ada pada algoritma K-Means yaitu: perbaikan pada algoritma K-Means menghasilkan cluster yang lebih baik yaitu penerapan *Sum Of Squared* Error *(SSE) untuk membatu K-Means Clustering dalam* menetukan jumlah cluster yang paling optimum, dari proses modifikasi ini, diharapkan pusat cluster yang diperoleh nantinya akan menghasilkan *cluster-cluster*, dimana antar anggota *cluster* memiliki tingkat kemiripan yang tinggi. Perbaikan performa cluster K-Means akan diterapkan pada penentuan pusat cluster.

Kata Kunci: Optimasi, K-Means, Sum of Squared Error

PENDAHULUAN

Latar Belakang

Teknologi clustering data merupakan suatu teknik yang menunjukan persamaan karakteristik dalam suatu kelompok sehingga akan menghasilkan informasi yang bermaanfaat. Algoritma clustering data sudah banyak dipergunakan diberbagai bidang misalnya untuk proses pengolahan citra, data mining proses pengambilan keputusan, pengenalan pola.

Salah satu metode yang terdapat pada *Data Mining* untuk menentukan pola-pola penggalian informasi dengan menggunakan metode *Clustering* yang dimana metode tersebut mengelompokan objek yang memiliki kesamaan karakteristik hingga menemukan pola-pola yang diinginkan. Proses menentukan pola-pola pengelompokan atau *clustering* diantaranya menggunakan algoritma *K-Means*. Algoritma *K-Means clustering* merupakan salah satu algoritma pengelompokan data dengan sistem partisi *K-Means* (Debalty, 2014).

Algoritma *K-Means* memiliki tingkat ketelitian yang tinggi, efektif serta membutuhkan waktu eksekusi yang relatif cepat karena bersifat linear. Tahap awal dalam algoritma *K-Means* adalah menentukan jumlah k atau jumlah *cluster* terlebih dahulu. Untuk menentukan jumlah *cluster* terbaik digunakan metode *elbow.* Pada metode *elbow* nilai *cluster* terbaik yang akan diambil dari nilai *Sum of Square Error* (SSE) yang mengalami

penurunan yang signifikan dan berbentuk siku (Kaur etAl, 2013).

Beberapa penelitian menggunakan *K-Means* pernah dilakukan sebelumnya, *data mining* menggunakan algoritma *k-means clustering* untuk menentukan strategis promosi Universitas Dian Nuswantoro (Ramadhani, 2013). Analisa Penentuan Jumlah Cluster Terbaik Pada Metode *K-Means Clustering* (Merliana dkk., 2015). Algoritma Modifield K-Means Clustering pada Penentuan Cluster Center Berbasis Sum of Squared Error (SSE) (Nainggolan, 2014). Penerapan algoritma *k-means* pada data mining untuk memilih produk dan pelanggan potensial (Prasetyo, 2013).

Dengan bantuan teknik *data mining*, seperti algoritma clustering, yang memungkinkan untuk menemukan karakteristik-karakteristik dari pasien dan menggunakan karakteristik mereka sebagai bahan acuan atau pedoman pihak manajemen dan pihak fungsional untuk memprediksi akan keberhasilan pelayanan dari layanan kesehatan atau rumah sakit di masa yang akan mendatang.

Clustering

Clustering termasuk dalam klasifikasi tanpa *pengawasan* (*unsupervised classification*). Pengertian *Clustering* adalah proses mengelompokkan atau penggolongan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/cluster. *Clustering* dalam data mining berguna untuk menemukan pola distribusi di dalam sebuah dataset yang berguna untuk proses analisa data. Kesamaan objek biasanya diperoleh dari kedekatan nilai-nilai atribut yang menjelaskan objek-objek data, sedangkan objek-objek data biasanya direpresentasikan sebagai sebuah titik dalam ruang multi dimensi (Jain etAl, 1988).

K-Means Clustering

K-Means merupakan salah satu metode pengelompokan data nonhierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan kedalam kelompok yang lain. Adapun tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang diatur dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antar kelompok (Prasetyo, 2013).

Algoritma *K-Means clustering* merupakan salah satu algoritma pengelompokan data dengan sistem partisi [1]. Untuk itu digunakan aturan dalam Algoritma *K-Means* sebagai berikut :

- Jumlah *cluster* atau k harus diinisialisasikan terlebih dahulu
- Atribut bersifat numerik
- Keterbatasan atribut
- Kompleksitas algoritma linear (n)

Algoritma *K-Means* termasuk dalam metode *non-hierarchical* yang mempartisi data ke dalam bentuk satu atau lebih *cluster*, sehingga data yang mempunyai karakteristik yang sama dikelompokkan dalam satu cluster yang sama dan data yang memiliki karakteristik berbeda dikelompokkan ke dalam *cluster* lain. Algoritma ini merupakan algoritma yang paling umum digunakan karena mudah untuk diimplementasikan. Adapun kelemahannya adalah algoritma ini sangat sensitif terhadap inisialisasi *cluster*.

Berikut ini urutan Algoritma *K-Means*:

1. Menentukan jumlah k-cluster yang akan dibentuk.
2. Membangkitkan k-centroid (titik pusat cluster) secara acak.

Menghitung jarak setiap data terhadap masing-masing centroid. Rumus yang digunakan adalah rumus jarak Euclidean (*Euclidean Distance*) dengan persamaan (1) sebagai berikut :

$$\sqrt{(X1 - X2)^2 + (y1 - y2)^2}$$

Ukuran jarak atau ketidaksamaan antar objek ke-i denganobjek ke-j, disimbolkan dengan dij dan k=1,…..,p. Nilai dij diperoleh melalui perhitungan jarak kuadrat Euclidean sebagai berikut:

$$dij = \sqrt{\sum_{k=1}^{p} (xik - xjk)^2}$$

Keterangan :

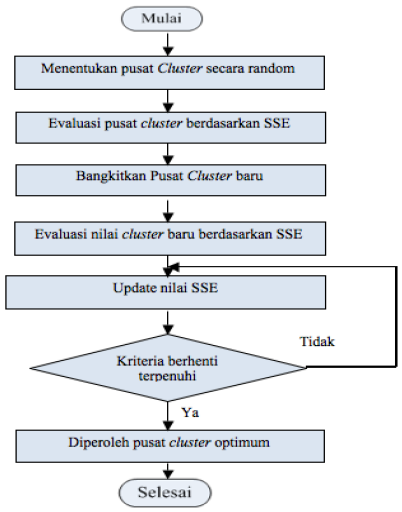
dij = Jarak Kuadrat Euclidean antar objek ke-i dengan obejk ke-j

P = Jumlah Variabel cluster

Xik = Nilai atau data dari objek ke-i pada variable ke-k

Xjk = Nilai atau data dari objek ke-j pada variable ke-k

Berikut tahapan algoritma *K-Means* menggunakan *flowchart*



Gambar 1. Flowchart *Modified K-Mean Clustering* berbasis SSE

METODE PENELITIAN

Metode penelitian pada penelitian ini dapat dijelaskan dalam bentuk langkah-langkah sebagai berikut:

1. Menentukan parameter *MK-Means Clustering* (jumlah iterasi)

Jumlah iterasi yang akan diuji Adalah 100 pusat *cluster*. Jumlah iterasi (i) yang akan diuji adalah masing-masing 5 pusat *cluster* pada setiap iterasi, yaitu sebanyak 20 iterasi.

2. Jumlah *cluster* adalah sebanyak 3 *cluster*.
3. Membangkitkan solusi awal yaitu pusat *cluster* secara random . Menghitung nilai *Sum of Squared Error (SSE)* dengan menggunakan persamaan SSE.
4. Evaluasi pusat *cluster* berdasarkan nilai SSE.
5. Membangkitkan pusat *cluster* baru bedasarkan nilai SSE yaitu nilai SSE yang paling minimum pada pusat *cluster* sebelumnya.
6. Update nilai SSE.
7. Kriteria berhenti apabila iterasi telah melakukan pencarian SSE sebanyak 20 kali iterasi atau 100 pusat *cluster* yang berbeda beda yang dibangkitkan secara random.
8. Output dari sebuah system adalah nilai SSE yang paling minimum yang merupakan pusat *cluster* yang paling optimum.

Pengumpulan Data

Berikut ini adalah data pasien sebanyak 200 data untuk dilakukan uji coba terhadap algoritma *Modified K-Mean Clustering* dalam penentuan pusat *cluster* berbasis *Sum of Squared Error (SSE).*

Tabel 1. Data Awal

No	Nama	Wilayah	Pekerjaan	Umur
0	Daniel	Medan Perjuangan	TKI	33
1	Hendrik	Medan Selayang	Peg. Swasta	25
2	Rinaldi	Belawan	Pedagang	44
3	Marusaha	Medan Perjuangan	TKI	31
4	Ana	Medan Perjuangan	TKI	37
5	Linda	Medan Selayang	Peg. Swasta	40
199	Hendry	Medan Kota	TKI	65

Transformasi Data

Agar data diatas dapat diolah dengan menggunakan metode *K-Means Clustering*, maka data yang berjenis data nominal seperti wilayah dan pekerjaan harus diinisialisasikan terlebih dahulu dalam bentuk angka.

Untuk melakukan inisialisasi wilayah dilakukan dengan cara mengurutkan dari yang terbesar berdasarkan frekuensi pasien yang berasal dari wilayah tersebut. Setelah itu wilayah yang memiliki frekuensi terbesar diberi inisial dengan angka 1 dan wilayah yang memiliki frekuensi terbesar kedua diberi inisial dengan angka 2. Hasil dari inisialisasi wilayah dapat dilihat pada tabel 2.

Tabel 2. Inisialisasi Data Wilayah Kota Asal

No	Wilayah	Frekuensi	Inisialisasi
1	Belawan	135	1
2	Medan Perjuangan	35	2
3	Medan Selayang	25	3
4	Medan Kota	5	4

Selain wilayah, pekerjaan juga termasuk ke dalam jenis data nominal, sehingga perlu diinisialisasikan ke dalam bentuk angka. Seperti pada wilayah, pada pekerjaan juga diberikan inisialisasi berdasarkan frekuensi pekerjaan pasien tersebut. Hasil dari inisialisasi pekerjaan tersebut dapat dilihat pada tabel 3.

Tabel 3. Inisialisasi Data Pekerjaan

No	Pekerjaan	Frekuensi	Inisialisasi
1	TKI	98	1
2	Peg. Swasta	75	2
3	Pedagang	10	3
4	PNS	15	4
5	Therapys	2	5

Tabel 4. Keseluruhan Data yg diinisialisasikan

No	Nama	Wilayah	Pekerjaan	Umur
0	Daniel	2	1	33
1	Hendrik	3	2	25
2	Rinaldi	1	3	44
3	Marusaha	2	1	31
4	Ana	2	1	37
5	Linda	3	2	40
199	Hendry	5	1	24

Setelah semua data pasien ditransformasi ke dalam bentuk angka, maka data tersebut telah dapat dikelompokkan dengan menggunakan metode *K-Mean Clustering*.

HASIL PENELITIAN

Algoritma *Modified K-Mean Clustering* berbasis *Sum of Squared Error (SSE)* diawali dengan pembangkitan pusat *cluster* secara random dengan batas maksimum yang sudah ditentukan sebelumnya yaitu 20 iterasi. Tahap pertama adalah pembangkitan kembali pusat *cluster* yang memiliki nilai SSE yang paling minimum. Tahap kedua adalah *sorting* atau pengurutan pusat *cluster* berdasarkan nilai SSE. Pada pengujian ini, iterasi awal terdiri dari 4 pusat *cluster*, masing – masing pusat *cluster* berbeda-beda.

Tabel 5 adalah pusat *cluster* pertama yang dibangkitkan secara random yaitu pada data ke [0,1,2] yaitu dengan pasien data ke-0 berada di wilayah Medan Perjuangan dan pekerjaan pasien adalah Tenaga Kerja Indonesia (TKI) dengan umur 33 Tahun, untuk data pasien ke-1 adalah berada di wilayah Medan Selayang dan pekerjaan pasien adalah pegawai swasta dengan umur 25 tahun dan untuk data pasien ke- 2 adalah berada di wilayah Belawan dan pekerjaan pasien adalah pedagang dengan umur 44. Dan hasil inisialisasi pusat *cluster* dengan data pasien ke 0, 1, 2 adalah [2, 1, 33], [3, 2, 25], [1, 3, 44].

Tabel 5. Titik Pusat *Cluster* pada solusi pertama

Cluster Baru	Wilayah	Pekerjaan	Umur
C1	2	1	33
C2	3	2	25
C3	1	3	44

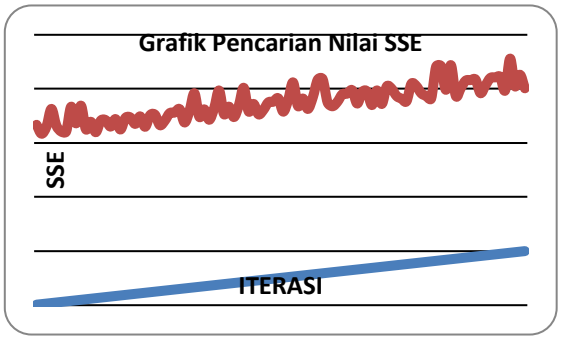
Pada tabel 6 ditampilkan hasil pencapain nilai SSE yang paling minimum pada setiap iterasi dan pusat *cluster* yang paling optimum pada setiap iterasi.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Tabel 6. Pencapaian Nilai SSE Minimum pada Setiap Iterasi

Iterasi	Pusat *Cluster*	Nilai SSE Minimum
1	[2, 1, 33][3, 1, 22][1, 3, 44]	315.0354
2	[2, 5, 30][2, 1, 33][1, 3, 41]	314.2185
3	[3, 2, 25][3, 2, 40][2, 5, 30]	306.8306
4	[3, 2, 25][3, 2, 40][2, 5, 30]	306.8306
5	[3, 2, 25][3, 2, 40][2, 5, 30]	306.8306
6	[3, 2, 25][3, 2, 40][2, 5, 30]	306.8306
7	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
8	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
9	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
10	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
11	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
12	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
13	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
14	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
15	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
16	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
17	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
18	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
19	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
20	[2, 1, 37][2, 1, 33][3, 1, 27]	301.0533
	SSE Minimum	**301.0533**

Pada Gambar 2 ditampilkan grafik hasil pencarian nilai SSE, dimana untuk setiap iterasi menghasilkan nilai yang berbeda beda, untuk mencari pusat *cluster* yang paling optimum maka akan dicari nilai SSE yang paling minimum.



Gambar 2. Grafik Pencarian SSE Terbaik dengan Metode *Modified K-Mean Clustering* berbasis *Sum of Squared Error (SSE)*

Dari grafik diatas dapat kita simpulkan bahwa nilai SSE yang terbaik adalah 301.0533 dengan pusat *cluster* [2, 1, 37][2, 1, 33][3, 1, 27].

Tabel 7. Pusat *Cluster* Optimum

Pusat *Cluster*	Wilayah	Pekerjaan	Umur
C1	2	1	37
C2	2	1	33
C3	3	1	27

Tabel 8. Jarak Setiap Data Pasien ke Titik Centroid pada Iterasi ke – 1

No	Nama	W	P	U	Jarak Ke			Min *Cluster*			
					C1	C2	C3	C1	C2	C3	
0	Daniel	2	1	33	4.0	0.0	6.083		*		C2
1	Hendrik	3	2	25	12.083	8.124	2.236			*	C3
2	Rinaldi	1	3	44	7.348	11.225	17.234	*			C1
3	Marusaha	2	1	31	6.0	2.0	4.123		*		C2
4	Ana	2	1	37	4.0	10.05	10.05	*			C1
5	Linda	3	2	40	3.317	7.141	13.038	*			C1
199	Hendry	3	2	26	11.091	7.141	1.414			*	C3

Langkah selanjutnya menghitung pusat *cluster* baru. Pusat *cluster* baru ditentukan berdasarkan pengelompokan anggota masing-masing *cluster* berdasarkan tabel diatas, *cluster* pertama untuk parameter wilayah pasien memiliki 7 anggota (tujuh) yaitu pasien ke-2, ke-4, ke-5, ke-7, ke-8, ke-14, ke-15. Pusat *cluster* baru untuk *cluster* pertama dihitung berdasarkan rata-rata koordinat ketujuh anggota tersebut adalah:

Tabel 9 Pusat *Cluster* pada iterasi ke-2

Cluster Baru	Wilayah	Pekerjaan	Umur
C1	2.286	2.714	40.857
C2	2.0	2.333	21.333
C3	2.5	2.1	25.2

Tabel 10. Jarak Setiap Data Pasien ke Titik Centroid pada Iterasi ke – 2

No	Nama	W	P	U	Jarak Ke			Min *Cluster*			
					C1	C2	C3	C1	C2	C3	
0	Daniel	2	1	33	8.047	2.134	7.893		*		C2
1	Hendrik	3	2	25	15.889	6.42	0.548			*	C3
2	Rinaldi	1	3	44	3.408	12.724	18.881	*			C1
3	Marusaha	2	1	31	10.009	1.374	5.925		*		C2
4	Ana	2	1	37	4.23	5.822	11.862	*			C1
5	Linda	3	2	40	1.324	8.731	14.809	*			C1
199	Hendry	3	2	26	14.891	5.436	0.949			*	C3

Langkah selanjutnya menghitung pusat *cluster* baru. Pusat *cluster* baru ditentukan berdasarkan pengelompokan anggota masing-

masing *cluster* berdasarkan tabel diatas, *cluster* pertama untuk parameter wilayah pasien memiliki 7 anggota (tujuh) yaitu pasien ke-2, ke-4, ke-5, ke-7, ke-8, ke-14, ke-15. Pusat *cluster* baru untuk *cluster* pertama dihitung berdasarkan rata-rata koordinat ketujuh anggota tersebut adalah:

Tabel 11 Pusat *Cluster* pada iterasi ke-3

Cluster Baru	Wilayah	Pekerjaan	Umur
C1	2.286	2.714	40.857
C2	2.0	2.333	31.333
C3	2.5	2.1	25.2

Pengulangan dihentikan karena hasil perhitungan menunjukkan adanya pusat *cluster* yang sama pada iterasi ke-2 dan iterai ke-3.

KESIMPULAN

Berdasarkan hasil pengelompokan data menggunakan metode *Modified K-Means Clustering* berbasis *Sum Of Squares Error (SSE),* di dapatkan hasil *cluster*ing hingga iterasi ke-3, dimana titik pusat tidak lagi berubah dan tidak ada data yang berpindah antar *cluster*. Dari hasil *cluster* 1, terlihat bahwa karakteristik pasien pada *cluster* 1 dominasi oleh pasien yang berada di wilayah Medan Perjuangan dan sebagian kecil berada di wilayah Medan Selayang, dan pekerjaan pasien di dominasi oleh pedagang dan sebagian kecil pekerjaan pasien adalah pegawai swasta dengan umur pasien sekitar 40 tahun.

Kemudian, hasil *cluster* 2 diatas dapat dilihat bahwa karakteristik pasien pada *cluster* 2 didominasi oleh pasien yang berasal dari wilayah Medan Perjuangan dan pekerjaan pasien adalah pegawai swasta dan sebagian kecil adalah pedagang dengan umur pasien sekitar 31 Tahun.

Sedangkan, dari hasil *cluster* di atas dapat dilihat bahwa karakteristik pasien pada *cluster* 3 didominasi oleh pasien yang berada di wilayah Medan Perjuangan dan Medan Selayang dan pekerjaan pasien adalah pegawai swasta dengan umur pasien sekitar 25 Tahun.

DAFTAR PUSTAKA

Debatty, T., et.Al, (2014). Determining the k in k-means with Map Reduce. *Proceedings of the EDBT/ICDT 2014 Joint Conference* (ISSN 1613-0073), 19-28

Jain, A.K., and Dubes, R.C., (1988). *Algorithms for Clustering Data*, New Jersey: Prentice-Hall, Inc.

Kaur, K., Dhaliwal, D.S. & Vohra, K.R., (2013). Statistically Refining the Initial Points for KMeans Clustering Algorithm. *International Journal of Advanced Research in Computer Engineering & Technology, II* (11), pp.2972-2977

Merliana, P.E., Ernawati & Santoso, A.J., (2015). Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means. *Skripsi.* UNISBANK.

Nainggolan, R., (2014). Algoritma Modifield K-Means Clustering pada Penentuan Cluster Center Berbasis Sum of Squared Error (SSE), *Tesis.* Universitas Sumatera Utara.

Prasetyo, E., (2012). *Data Mining: Konsep dan Aplikasi menggunakan MATHLAB*. Yogyakarta: Andi Offset.

Ramadhani, D, (2014). Data Mining Menggunakan Algoritma K-Means Clustering untuk Menentukan Strategi Promosi Universitas Dian Nuswantoro. *Skripsi.* Jurusan Sistem Informasi Universitas Dian Nuswantoro.