PROSIDING SEMINAR NASIONAL MULTI DISIPLIN ILMU & *CALL FOR PAPERS* UNISBANK (SENDI_U)
Kajian Multi Disiplin Ilmu untuk Mewujudkan Poros Maritim dalam Pembangunan Ekonomi Berbasis Kesejahteraan Rakyat
ISBN: 978-979-3649-81-8

ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING

Ni Putu Eka Merliana, Ernawati, Alb. Joko Santoso
Program Studi Magister Teknik Informatika, Fakultas Teknik Industri, Universitas Atma Jaya
Jl. Babarsari 43 Yogyakarta 55281, Indonesia
Email: putuekamerliana@gmail.com, ernawati@mail.uajy.ac.id, albjoko@mail.uajy.ac.id

Abstract

Clustering is a technique used to analyze data either in machine learning, data mining, pattern recognition, image analysis and bioinformatics. So as to produce useful information need for an analysis of data using clustering process because data has a lot of variety and quantity. In this case the researchers will use the K-Means method in which these methods into an efficient and effective algorithms to process data with the variety and number of lots. K-means algorithm has a problem in determining the best number of clusters. So in this paper the researchers will conduct research to search for the best number of clusters in K-means method. There are many ways to determine this, one of them with methods Elebow. The determination of these methods seen from the graph SSE (Sum Square Error) of some number of clusters. Results from this study will be the basis for determining the number clusters in the process clustering with K-Means method in a case study, and this case study will be conducted at the institute STAHN (Sekolah Tinggi Agama Hindu Negeri) Tampung Penyang Palangkaraya.
Keywords: Clustering, K-Means, Method Elbow, SSE (Sum Square Error)

Abstrak

Clustering merupakan suatu teknik yang digunakan untuk menganalisa data baik itu dalam mesin learning, data mining, pengenalan pola, analisa gambar maupun bioinformatika. Untuk menghasilkan informasi yang bermanfaat diperlukannya suatu analisa data dengan menggunakan proses clustering, ini disebabkan karena data memiliki variasi dan jumlah yang banyak. Peneliti akan menggunakan metode K-Means dimana metode ini menjadi algoritma yang efisien dan efektif dalam mengolah data dalam jumlah yang banyak. Algoritma K-means memiliki permasalahan dalam penentuan jumlah cluster yang terbaik. Sehingga dalam paper ini peneliti akan melakukan penelitian dengan mencari jumlah cluster terbaik pada metode K-means. Terdapat banyak cara dalam menentukan hal tersebut, salah satunya dengan metode Elebow. Penentuan metode ini dilihat dari grafik SSE (Sum Square Error) dari beberapa jumlah cluster. Hasil dari penelitian ini akan dijadikan dasar untuk penentuan jumlah cluster dalam melakukan proses clustering dengan metode K-Means pada suatu studi kasus, dalam hal ini penelitian akan dilakukan di lembaga STAHN (Sekolah Tinggi Agama Hindu Negeri) Tampung Penyang Palangka Raya.
Kata kunci : Clustering, K-Means, Metode Elbow, SSE (Sum Square Error)

I. PENDAHULUAN

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, maupun dalam bidang bioinformatika[1]. Clustering data akan mengelompokan objek yang paling dekat dimana terdapat kesamaan dengan objek lain, serta data yang akan dicluster diambil secara acak atau random. Data yang dikelompokkan dengan memiliki karakteristik sama mempergunakan metode clustering.

Ada beberapa algoritma yang diusulkan untuk dapat melakukan proses clustering pada suatu dataset dalam jumlah yang banyak. Pada penelitian ini, peneliti akan menggunakan metode algoritma K-Means dalam menentukan jumlah cluster terbaik. K-Means merupakan algoritma yang sangat banyak dipergunakan karena effektif dan effisien. Ini dikarenakan K-means sangat mudah dipelajari dan dari segi waktu proses komputasinya relatif singkat[2]. Penentuan nilai hasil cluster dilihat dari jarak terdekat antar objek data. Sebelum menghitung jarak terdekatnya, harus ditentukan terlebih dahulu

ISBN: 978-979-3649-81-8

jumlah cluster centroid[3]. Selain itu K-Means juga memiliki ketelitian yang cukup tinggi terhadap ukuran objek, sehingga algoritma ini relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah yang besar[4]. Yang menjadi kelemahan dalam algoritma K-Means adalah menganalisa dan menentukan jumlah k terbaik dalam mengcluster data pada suatu dataset. Identifikasi jumlah cluster k merupakan cara yang paling penting dan utama pada proses clustering dengan menggunakan algoritma K-Means dimana hasil cluster akan bergantung pada jumlah cluster awal. Sehingga jika jumlah cluster yang ditentukan tidak baik maka hasil cluster juga tidak akan sesuai dengan yang diharapkan yaitu tidak akan menghasilkan informasi yang dibutuhkan oleh pengguna.

Untuk mengatasi ini, penulis melakukan penelitian dengan mencari nilai k terbaik menggunakan metode *elbow*. Metode ini sudah cukup lama dipergunakan, dimana metode ini akan melihat fungsi dari nilai cluster pada suatu data [5]. Metode *elbow* juga sangat mudah diimplementasikan dengan cara melihat grafik dari nilai k yang akan diinputkan. Nilai fungsi k yang akan dibandingkan pada metode *elbow* adalah dengan melihat nilai SSE (*Sum of Square Error*) pada nilai cluster yang ditentukan. Hasil jumlah cluster k terbaik akan dijadikan dasar untuk melakukan proses clustering dengan menggunakan metode K-Means dalam suatu studi kasus yaitu pada Sekolah Tinggi Agama Hindu Negeri Tampung Penyang Palangka Raya. Diharapkan hasil cluster dengan metode K-Means menghasilkan hasil cluster yang optimal dan dapat membantu menghasilkan informasi yang dibutuhkan.

II. CLUSTERING

Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah data set kedalam subset, sehingga data dalam setiap subset memiliki arti yang bermanfaat. Sebuah cluster terdiri dari kumpulan benda-benda yang mirip antara satu dengan yang lainnya dan berbeda dengan benda yang terdapat pada cluster lainnya. Algoritma clustering terdiri dari dua bagian yaitu secara hirarkis dan secara partitional. Algoritma hirarkis menemukan cluster secara berurutan dimana cluster ditetapkan sebelumnya, sedangkan algoritma partitional menentukan semua kelompok pada waktu tertentu[6]. Clustering juga bisa dikatakan suatu proses dimana mengelompokan dan membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokkan pada cluster yang sama dan memisahkan diri dengan membentuk pola yang berbeda di cluster yang berbeda[7]. Clustering dapat ditemukan dibeberapa aplikasi yang ada di berbagai bidang. Sebagai contoh pengelompokan data yang digunakan untuk menganalisa data statistik seperti pengelompokan untuk pembelajaran mesin, data mining, pengenalan pola, analisis citra dan bioinformatika[1].

III.K-Means

Algoritma K-Means merupakan salah satu algoritma dengan partitional, karena K-Means didasarkan pada penentuan jumlah awal kelompok dengan mendefinisikan nilai centroid awalnya[6]. Algoritma K-Means menggunakan proses secara berulang-ulang untuk mendapatkan basis data cluster. Dibutuhkan jumlah cluster awal yang diinginkan sebagai masukan dan menghasilkan jumlah cluster akhir sebagai output. Jika algoritma diperlukan untuk menghasilkan cluster K maka akan ada K awal dan K akhir. Metode K-Means akan memilih pola k sebagai titik awal centroid secara acak. Jumlah iterasi untuk mencapai cluster centroid akan dipengaruhi oleh calon cluster centroid awal secara random dimana jika posisi centroid baru tidak berubah. Nilai K yang dipilih menjadi pusat awal, akan dihitung dengan menggunakan rumus Euclidean Distance yaitu mencari jarak terdekat antara titik centroid dengan data/objek. Data yang memiliki jarak pendek atau terdekat dengan centroid akan membentuk sebuah cluster[8].

Algoritma K-Means

1. Tentukan k sebagai jumlah cluster yang akan dibentuk
2. Tentukan k Centroid (titik pusat cluster) awal secara random/acak.

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ;i = 1,2,3,...n \qquad \text{.....(1)}$$

Dimana; v : centroid pada cluster
x_i : objek ke-i
n : banyaknya objek/jumlah objek yang menjadi anggota cluster

3. Hitung jarak setiap objek ke masing-masing centroid dari masing-masing cluster. Untuk menghitung jarak antara objek dengan centroid dapat menggunakan Euclidian Distance

$$d(x,y) = \|x-y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad ;i = 1,2,3,...,n \qquad \text{.....(2)}$$

Dimana; x_i : objek x ke-i
y_i: daya y ke-i

ISBN: 978-979-3649-81-8

n : banyaknya objek

4. Alokasikan masing-masing objek ke dalam centroid yang paling dekat
5. Lakukan iterasi, kemudian tentukan posisi centroid baru dengan menggunakan persamaan (1)
6. Ulangi langkah 3 jika posisi centroid baru tidak sama [9].

Gb. 1. Flowchart algoritma K-Means

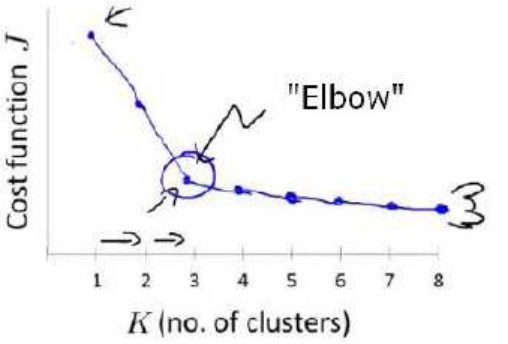
IV. Metode *Elbow*

Metode *Elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik[6]. Metode ini memberikan ide/gagasan dengan cara memilih nilai cluster dan kemudian menambah nilai cluster tersebut untuk dijadikan model data dalam penentuan cluster terbaik. Dan selain itu persentase perhitungan yang dihasilkan menjadi pembanding antara jumlah cluster yang ditambah [12]. Hasil persentase yang berbeda dari setiap nilai cluster dapat ditunjukan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai cluster pertama dengan nilai cluster kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai cluster tersebut yang terbaik [5].

Untuk mendapatkan perbandingannya adalah dengan menghitung SSE (*Sum of Square Error*) dari masing-masing nilai cluster. Karena semakin besar jumlah cluster K maka nilai SSE akan semakin kecil. Rumus SSE pada K-Means [10]

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} \left\| X_i - C_k \right\|_2^2$$

Setelah dilihat akan ada beberapa nilai K yang mengalami penurunan paling besar dan selanjutnya hasil dari nilai K akan turun secara perlahan-lahan sampai hasil dari nilai K tersebut stabil. Misalnya nilai cluster K=2 ke K=3, kemudian dari K=3 ke K=4, terlihat penurunan drastis membentuk siku pada titik K=3 maka nilai cluster k yang ideal adalah K=3[11].



Gb 2. Grafik metode *Elbow* [11]

Algoritma Metode *Elbow* dalam menentukan nilai K pada K-Means

1. Mulai
2. Inisialisasi awal nilai K
3. Naikkan nilai K
4. Hitung hasil *sum of square error* dari tiap nilai K
5. Melihat hasil *sum of square error* dari nilai K yang turun secara drastis
6. Tetapkan nilai K yang beberbentuk siku
7. Selesai [5]

V. Hasil dan Pembahasan

Penelitian ini dilakukan dengan mengambil data selama 3 tahun terakhir untuk mahasiswa yang masih aktif yaitu sebanyak 855 data pada Sekolah Tinggi Agama Hindu Negeri Tampung Penyang Palangka Raya. Data yang akan diuji coba adalah data jumlah kunjungan mahasiswa ke perpustakaan, IPK yang dimiliki dan jumlah buku yang dipinjam dengan attribute pendukung jenis kelamin, tahun akademik, program studi dan jumlah sks yang ditempuh oleh mahasiswa. Jumlah cluster yang akan diuji adalah dari K=2 sampai dengan K=8. Untuk melakukan uji

ISBN: 978-979-3649-81-8

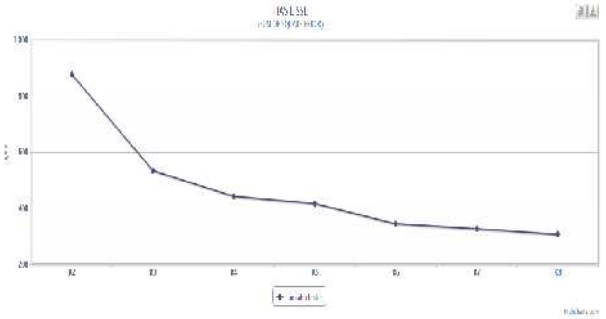
coba dalam menentukan jumlah k terbaik maka uji coba akan dilakukan dalam 4(empat) kali uji coba dengan jumlah data yang berbeda-beda

Uji Coba I dengan 54 data mahasiswa

Dari hasil proses perhitungan *Sum of Square Error* terhadap 54 data mahasiswa maka hasil yang mengalami penurunan yang paling besar adalah pada K=3. Ini dapat dilihat pada Tabel 1 dan Grafik 1

Tabel 1. Hasil *Sum of Square Error* dari tiap-tiap cluster untuk 54 Data

Cluster	Hasil Sum Square Error	Selisih
K2	876,427	876,427
K3	533,305	343,122
K4	441,387	91,918
K5	415,409	25,978
K6	343,584	71,825
K7	326,248	17,336
K8	305,867	20,381



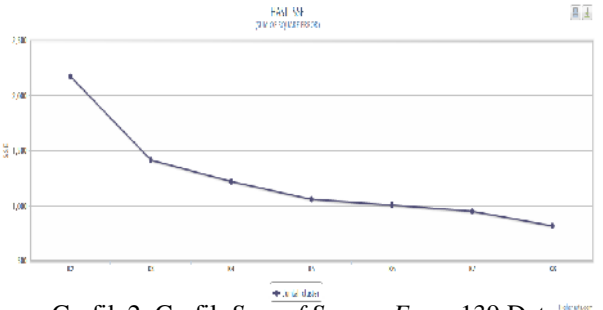
Grafik 1. Grafik *Sum of Square Error* 54 Data

Uji Coba II dengan 139 data mahasiswa

Sama seperti uji coba I pada 54 data mahasiswa, pada uji coba 139 data mahasiswa terdapat penurunan hasil *Sum of Square Error* paling besar adalah pada K=3 ini dapat dilihat pada Tabel 2 dan Grafik 2 dibawah ini.

Tabel 2. Hasil *Sum of Square Error* dari tiap-tiap cluster untuk 139 Data

Cluster	Hasil Sum Square Error	Selisih
K2	2171,64	2171,64
K3	1414,61	757,03
K4	1218,69	195,92
K5	1056,53	162,16
K6	1004,82	51,71
K7	948,043	56,777
K8	815,712	132,331



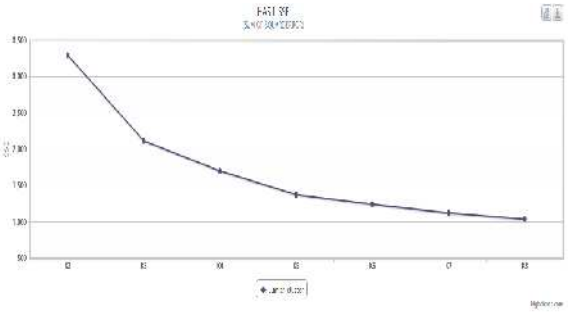
Grafik 2. Grafik *Sum of Square Error* 139 Data

Uji Coba III dengan 261 data mahasiswa

Sama pada uji coba I dan II, uji coba dengan 261 data mahasiswa dimana data yang diambil adalah data mahasiswa selama satu tahun, hasil dari perhitungan *Sum of Square Error* dari masing-masing cluster yang mengalami penurunan paling besar adalah pada K=3 dapat dilihat pada Tabel 3 dan Grafik 3 dibawah ini.

Tabel 3. Hasil *Sum of Square Error* dari tiap-tiap cluster untuk 261 Data Mahasiswa

Cluster	Hasil Sum Square Error	Selisih
K2	3291,32	3291,32
K3	2110,59	1180,73
K4	1697,28	413,31
K5	1371,04	326,24
K6	1238,2	132,84
K7	1119,83	118,37
K8	1034,02	85,81



Grafik 3. Grafik *Sum of Square Error* 261 Data

Uji Coba IV dengan 885 data mahasiswa

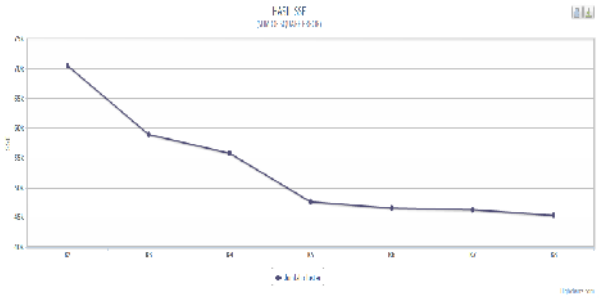
Uji coba IV yang dilakukan pada 885 data mahasiswa, hasil yang didapat pada perhitungan *Sum of Square Error* sama seperti uji coba I, II dan III, dimana penurunan yang paling besar adalah terdapat

ISBN: 978-979-3649-81-8

pada K=3 ini dapat dilihat pada Tabel 4 dan Grafik 4 dibawah ini.

Tabel 4. Hasil *Sum of Square Error* dari tiap-tiap cluster untuk 885 Data Mahasiswa

Cluster	Hasil Sum Square Error	Selisih
K2	70430,5	70430,5
K3	58879,6	11550,9
K4	55743,5	3136,1
K5	47578,9	8164,6
K6	46526,2	1052,7
K7	46250,3	275,9
K8	45311,1	939,2



Grafik 4. Grafik *Sum of Square Error* 885 Data

Dari keempat uji coba diatas berdasarkan studi kasus data yang diambil pada Sekolah Tinggi Agama Hindu Negeri Tampung Penyang Palangka Raya, hasil *Sum of Square Error* yang mengalami penurunan paling besar adalah pada K=3 dengan jumlah data yang berbeda. Sehingga untuk kasus ini jumlah cluster yang ideal adalah K=3 dan dijadikan default cluster untuk menentukan karakteristik dari data-data tersebut.

VI. Kesimpulan dan Saran

Berdasarkan hasil yang didapat dari proses dalam menentukan jumlah cluster terbaik dengan metode K-Means maka dapat disimpulkan sebagai berikut :

a. Penentuan jumlah cluster terbaik dengan metode *elbow* dapat menghasilkan jumlah cluster K yang sama pada jumlah data yang berbeda-beda
b. Hasil penentuan jumlah cluster terbaik dengan metode *elbow* akan dijadikan default untuk proses karakteristik berdasarkan studi kasus yang dilakukan.

Saran-saran terhadap penelitian ini adalah sebagai berikut :

a. Jika data sebagai centroid diinputkan secara random atau acak, maka hasil *Sum of Square Error* dari masing-masing cluster terkadang tidak membentuk grafik *elbow*, sehingga akan sulit dalam menentukan nilai cluster terbaik. Diharapkan kedepannya dapat menemukan metode yang bisa mengatasi kelemahan tersebut.
b. Dalam penentuan jumlah cluster terbaik, metode *elbow* harus dilakukan proses uji coba berkali-kali sehingga jika ingin menggunakan metode ini harus menentukan data centroid terlebih dahulu secara berurutan sesuai dengan jumlah cluster yang akan diuji.

REFERENSI

[1] Debatty, Thibault., et.Al, (2014). *Determining the k in k-means with MapReduce*. Proceedings of the EDBT/ICDT 2014 Joint Conference (ISSN 1613-0073), 19-28

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

[3] Vora, P. & Oza, B., 2013. *A Survey on K-mean Clustering and Particle Swarm Optimization*. International Journal of Science and Modern Engineering, I(3), pp.24-26

[4] Ediyanto, et.al, 2013. *Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis*. Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster), II(2), pp. 133-136.

[5] Bholowalia, Purnima & Kumar, Arvind, 2014. *EBK-Means: A Clustering Techiniques based on Elbow Method and K-Means in WSN. International Journal of Computer Application* (0975-8887), IX(105), pp. 17-24

[6] Madhulatha, T.S., 2012. *An Overview On Clustering Methods*. IOSR Journal of Engineering, II(4), pp.719-725

[7] HUNG, C.M., WU, J., CHANG, J.H. & YANG, D.L., 2005. *An Efficient k-Means Clustering Algorithm Using Simple Partitioning*. JOURNAL OF INFORMATION SCIENCE AND ENGINEERING, XXI(1), pp.1157-77

[8] Agrawal, A. & Gupta, H., 2013. *Global K-Means (GKM) Clustering Algorithm: A Survey*.

ISBN: 978-979-3649-81-8

International Journal of Computer Applications, LIX(2), pp.20-24

[9] Ediyanto, Mara, M.N. & Satyahadewi, N., 2013. *Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis*. Buletin Ilmiah Mat. Stat. dan Terapannya , II(2), pp.133-36.

[10]Irwanto, et. al (2012). *Optimasi Kinerja Algoritma Klasterisasi K-Means untuk kuantisasi Warna Citra*. Jurnal Teknik ITS, I(1), pp.197-202.

[11]Kodinariya, Trupti M. & Makwana, Prashant R., (2013). *Review on determining number of cluster in K-Means Clustering*. International Journal of Advance Research in Computer Science and Management Studies, I(6), pp. 90-95

ISBN: 978-979-3649-81-8