CLUSTERING DATA NUMERIK MENGGUNAKAN ALGORITME *X-MEANS*

(*Clustering Numeric Data Using X-Means Algorithm*)

Ayya Agustina Riza[1*)], Dewi Retno Sari Saputro[2)]

[1,2)] Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sebelas Maret, Jl. Ir Sutami No. 36A, Surakarta 57126, Indonesia
e-mail: rizaagustina480@student.uns.ac.id [1*)], dewiretnoss@staff.uns.ac.id [2)]
*)penulis korespondensi

Abstract. Data mining is the extraction of new and useful information from large data sets that helps in the decision-making process. Clustering is a technique of grouping data that has similar characteristics into the same cluster. Generally, the Clustering process is used for numeric or categorical data. The K-Means algorithm is one of the algorithms that can be used for numeric type data. The stage carried out in the K-Means algorithm is to divide n observations into k clusters so that each observation is included in the cluster with the closest average (centroid), but K-Means still has a weakness in determining the number of clusters. This must be determined specifically by the user. To overcome the weakness of K-Means, the X-Means algorithm was developed by Dan Pelleg and Andre Moore. In X-Means, the value of k is estimated by inputting a range of clusters based on the dataset itself, so that no specific determination of the number of clusters is needed. The purpose of this study is to examine the X-Means algorithm. The results showed that the division of clusters in the X-Means algorithm used the Bayesian Information Criterion (BIC) value. In the X-Means algorithm, inputting a range of clusters for the number of clusters can make the clustering process more efficient.

Keywords: Clustering, K-Means, numeric data, X-Means.

1. Pendahuluan

Data *mining* merupakan proses pengekstrakan atau penggalian informasi bermanfaat dari sekumpulan data *base* yang membantu proses pengambilan suatu keputusan atau *Knowledge Discovery* [13]. Data *mining* sering digunakan untuk menggali informasi dari pola *cluster* suatu data. Terdapat beberapa teknik dalam data *mining* diantaranya deskriptif, prediksi, estimasi, asosiasi, klasifikasi, dan *clustering* (pengelompokkan).

Clustering merupakan pengelompokkan objek yang dipartisi ke dalam suatu *cluster* yang memiliki kesamaan karakteristik. Kumpulan objek yang serupa di antara suatu kelompok dan berbeda dengan objek kelompok lain disebut *cluster* [7]. *Clustering* termasuk salah satu teknik analisis multivariat. Menurut Sitepu dkk. [11] analisis multivariat merupakan analisis yang menggunakan tiga variabel atau lebih dan bersifat multidimensial. Pengembangan metode *clustering* umumnya digunakan untuk data yang bertipe numerik.

Pengelompokkan data bertipe numerik dapat dilakukan menggunakan metode *partitional clustering*. Dalam metode *partitional clustering* salah satu algoritmenya adalah algoritme

K-Means. Dalam penelitian Arai & Barakbah [2], algoritme *K-Means* memiliki kemampuan untuk mengelompokkan objek berjumlah besar dengan waktu komputasi relatif efisien dan cepat. Salah satu kelemahan dari algoritme *K-Means* yaitu jumlah cluster harus ditentukan secara manual dan spesifik oleh user. Kemudian Pelleg & Moore [10] melakukan penelitian tentang algoritme *X-Means* yaitu mengekstensi algoritme *K-Means*, pada algoritme ini tidak diperlukan kembali memasukkan jumlah *cluster* secara spesifik namun dapat berupa *range cluster* yang akan dibentuk. Dengan demikian pada penelitian ini dilakukan kajian ulang terkait algoritme *X-Means*.

2. Metodologi

Penelitian ini merupakan penelitian berbasis teori yakni melakukan kajian terhadap teori algoritme *X-Means.* Dalam penelitian ini langkah-langkah yang dilakukan adalah mempelajari dan menganalisis kajian pustaka dari beberapa referensi buku, jurnal, serta tulisan lainnya terkait teori algoritme *X-Means*.

3. Hasil dan Pembahasan

Data merupakan nilai representasi dari suatu objek atau kejadian [9]. Data juga dapat diartikan sebagai fakta-fakta yang diolah sedemikian sehingga menghasilkan informasi tertentu. Data dibagi menjadi dua menurut jenis variabelnya, yaitu data numerik serta data kategorik [1]. Data numerik merupakan data yang berisi sekumpulan angka yang dapat diukur. Data kategorik merupakan data yang berisi sekumpulan kategori. Sekumpulan data besar yang di proses membentuk sebuah pola disebut data *mining*. Data *mining* dapat dikatakan juga suatu proses untuk mencari serta menambah informasi yang sebelumnya tidak diketahui dari basis data. Menurut Syahril dkk. [12] tujuan data *mining* yaitu untuk mendapatkan pola atau hubungan yang dapat memberikan petunjuk yang bermanfaat. Salah satu teknik data *mining* diantaranya adalah *Clustering*.

Clustering merupakan alat bantu data *mining* yang memiliki tujuan untuk mengelompokkan objek ke dalam suatu *cluster* [5]. Menurut Han *et al*., [4] *clustering* merupakan pengelompokkan objek ke setiap *cluster* yang memiliki karakteristik tinggi antar objek dalam suatu *cluster* namun tidak dengan objek dari yang *cluster* lain. Metode dalam *clustering* dapat dibagi menjadi dua, yaitu *hierarchical clustering* dan *partitional clustering*. Pada *hierarchical clustering* objek dikelompokkan melalui bagan berupa hirarki, dimana terdapat penggabungan dua grup terdekat di setiap iterasinya ataupun pembagian seluruh dataset ke dalam *cluster*. *Partitional clustering* adalah pengelompokkan data ke dalam sejumlah *cluster* yang mempunyai *centroid* tanpa terdapatnya struktur hirarki, dengan maksud untuk meminimumkan jarak seluruh data ke

centroid [14]. Terdapat salah satu data yang kelas *cluster*nya dapat ditentukan dengan algoritme *clustering* yaitu data numerik.

3.1 Data Numerik

Data numerik merupakan suatu data riil hasil perhitungan atau pengukuran [15]. Data numerik disebut juga data kuantitatif yang diperoleh dari suatu variabel yang nilainya mempunyai arti ukuran/besaran. Data numerik dapat berupa ordinal, interval, atau rasio. Data bertipe numerik dapat digunakan dalam *clustering* dengan metode *partitional clustering*. Salah satu algoritme dalam *partitional clustering* yang dapat digunakan untuk data numerik adalah algoritme *K-Means* dan pengembangannya *X-Means*.

3.2 Algoritme *K-Means*

Algoritme *K-Means* merupakan metode pengelompokkan data yang dilakukan dengan mempartisi data ke dalam suatu *cluster* yang jumlah *cluster (K)* telah ditentukan di awal. Secara umum menurut Everitt *et al.*,[3] langkah-langkah algoritme *K-Means* dituliskan sebagai berikut.

1. Menentukan jumlah K, dengan K merupakan banyak *cluster.*
2. Menentukan nilai awal *centroid* dari *cluster.*
3. Menghitung jarak data terhadap *centroid* menggunakan jarak *Euclidean*

$$d(x,y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \quad (1)$$

dengan

$d(x,y)$: jarak antara data ke-x ke pusat *cluster* y

x_i : data ke-x variabel ke-i

y_i : *centroid cluster* ke-y untuk variabel ke-i

n : banyaknya variabel.

4. Mengalokasikan data ke dalam *cluster* berdasarkan jarak minimum dengan *centroid.*
5. Menentukan *centroid* baru dengan persamaan

$$C_{ij} = \frac{1}{N}\sum_{i=0}^{N} x_{ij} \quad (2)$$

dengan

C_{ij} : *centroid* baru pada iterasi k

N : jumlah data dari anggota *cluster* ke-k

x_{ij} : data ke-k dalam *cluster.*

6. Mengulangi langkah 3 hingga tidak ada anggota *cluster* yang mengalami perubahan letak *cluster.*

3.3 Algoritme *X-Means*

Algoritme *X-Means* merupakan pengembangan dari algoritme *K-Means* dimana dalam *X-Means* terdiri dari beberapa operasi berulang hingga eksekusi berakhir serta pada algoritme ini mengoptimalkan nilai *Bayesian Information Criterion* (BIC). Pelleg & Moore [10] menguraikan algoritme *X-Means* yang ditulis sebagai berikut.

1. Menentukan *range cluster* $K(K_{min}, K_{max})$.
2. Menginisiasi nilai $K = K_{min}$.
3. Menjalankan *K-Means* hingga konvergen (anggota *cluster* tidak mengalami perubahan letak *cluster*).
4. Memperbaiki struktur, langkah ini dimulai dengan memecah setiap *centroid* hasil langkah 3 menjadi dua *children* dalam arah yang berlawanan di sepanjang vektor yang dipilih secara acak. Setelah itu menjalankan *K-Means* secara lokal di dalam setiap *cluster* untuk dua *cluster*. Keputusan masing-masing pusat *cluster* sendiri dengan membandingkan nilai-nilai BIC.
5. Memperbaharui nilai K, jika $K>K_{max}$ maka proses berhenti dan melaporkan struktur terbaik yang ditemukan selama pencarian, jika tidak kembali ke langkah 3.

3.4 *Bayesian Information Criterion* (*BIC*)

BIC merupakan salah satu alat yang dapat digunakan dalam pemilihan model statistik dimana memiliki kesederhanaan komputasi dan kinerja yang efektif [8]. Dalam *clustering, BIC* digunakan untuk mengetahui jumlah *cluster* terbaik yang dibentuk. *BIC* menggunakan probabilitas posteriors untuk memberi nilai dari model [6]. Persamaannya sebagai berikut

$$BIC(M_j) = \hat{l}_j(D) - \frac{P_j}{2}.log\ log\ R \qquad (3)$$

dengan

$\hat{l}_j(D)$: fungsi *log-likehood* data berdasarkan model ke-j

P_j : jumlah parameter pada M_j

Cluster akan dibagi menjadi dua *cluster*, jika *BIC* pada pembagian setelah *cluster* lokal nilainya lebih besar dari nilai *BIC cluster* awal. Selain itu, jika *BIC cluster* awal nilainya lebih besar dari *BIC* setelah dibagi menjadi *cluster* lokal maka *cluster* tersebut akan tetap.

4. Kesimpulan

Berdasarkan pembahasan diperoleh simpulan bahwa algoritme *X-Means* merupakan algoritme *clustering* untuk data bertipe numerik. Algoritme *X-Means* menggunakan *BIC* untuk mengontrol proses pemisahan *cluster* sehingga hasil *cluster* yang diperoleh dapat optimal. Pada algoritme *X-Means inputa*n suatu *range cluster* untuk jumlah *cluster* dapat membuat proses *clustering* lebih efisien.

Daftar Pustaka

[1] Anderson, T. W., & Sclove, S. L., (1974), *Introductory Statistical Analysis*, Houghton Mifflin.

[2] Arai, K. & Barakbah, A. R., (2007), Hierarchical *K-Means*: An Algorithm for Centroids Initialization for *K-Means*, *Reports of the Faculty of Science and Engineering*, **36(1),** 25-31.

[3] Everitt, B. S., Landau S., Leese M., & Stahl D., (2011), *Cluster Analysis 5th Edition*, Wiley, UK.

[4] Han, J., Kamber, M., & Pei, J., (2011), *Data Mining Concepts and Techniques* 3^{rd} *Edition*, Morgan Kaufmann, USA.

[5] Hariyanto, M. & Shita, R. T., (2018), *Clustering* Pada Data *Mining* untuk Mengetahui Potensi Penyebaran Penyakit DBD Menggunakan Metode Algoritma *K-Means* dan Metode Perhitungan Jarak Euclidean Distance, *SKANIKA*, **1(1)**, 117-122.

[6] Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the american statistical association*, **90(431)**, 928-934.

[7] Madhulatha, T. S., (2012) An Overview On *Clustering* Methods, *IOSR Journal of Engineering*, **2(4)**, 719-725, https://doi.org/10.48550/arXiv.1205.1117.

[8] Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian Information Criterion: Background, Derivation, and Applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, **4(2)**, 199-203, https://doi.org/10.1002/wics.199.

[9] Pamungkas, C. A., (2017), *Pengantar dan Implementasi Basis Data*, Deepublish, Yogyakarta.

[10] Pelleg, D., & Moore, A. W., (2000) *X-Means*: Extending *K-Means* with Efficient Estimation of the Number of Clusters, *In Icml* (Vol. 1, pp. 727-734).

[11] Sitepu, R., Irmeilyana, I., & Gultom, B. (2011). Analisis Cluster Terhadap tingkat pencemaran udara pada sektor industri di Sumatera Selatan. *Jurnal Penelitian Sains*, **14(3)**, https://doi.org/10.56064/jps.v14i3.208.

[12] Syahril, M., Erwansyah, K., & Yetri, M., (2020), Penerapan Data *Mining* untuk Menentukan Pola Penjualan Peralatan Sekolah pada Brand Wigglo dengan Menggunakan Algoritma Apriori. *Jurnal Teknologi Sistem Informasi Dan Sistem Komputer TGD*, **3(1)**, 118-136, https://doi.org/10.53513/jsk.v3i1.202.

[13] Tan, P., Steinbach, M., & Kumar, V., (2006), *Introduction to Data Mining*, Pearson Education Inc, New Delhi.

[14] Wanto, A., Siregar, M. N. H., Windarto, A. P., Hartama, D., Ginantra, N. L. W. S. R., Napitupulu, D., Negara, E. S., Dewi, M. R. L. S. V., & Prianto, C., (2020), *Data Mining: Algoritma dan Implementasi*. Yayasan Kita Menulis, Medan.

[15] Widaningsih, S., & Suheri, A., (2018), Pengelompokkan Data Profil Dosen Kopertis Wilayah IV Menggunakan Metode Two Step Cluster Analysis, *Konferensi Nasional Sistem Informasi (KNSI) 2018*.