PRISMA 5 (2022): 721-725
PRISMA, PROSIDING SEMINAR NASIONAL MATEMATIKA
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

Algoritme *Clustering Large Application (CLARA)* untuk MenanganiData *Outlier*

Pravasta Rama Fitrayana[a,*], Dewi Retno Sari Saputro[b]

[a,b] *Program Studi Matematika FMIPA Universitas Sebelas Maret, Jl. Ir. Sutami No.36 A, Surakarta 57126, Indonesia*

**Alamat Surel: pravasta11@student.uns.ac.id, dewiretnoss@staff.uns.ac.id*

Abstrak

Clustering merupakan salah satu metode pengelompokkan dalam data *mining*. *Clustering* mengelompokkan objek yang mempunyai kesamaan kedalam satu *cluster*. *Outlier* merupakan objek yang memiliki nilai menyimpang jauh dengan objek-objek lainnya sehingga terlihat tidak mengikuti pola dari sebagian besar objek. Pada*clustering* terdapat dua metode pendekatan yang dapat digunakan yaitu pendekatan *non-hierarki* (partisi) dan pendekatan *hierarki*. *K-Medoids* atau algoritme *Partitioning Around Medoid (PAM)* merupakan algoritme dari pendekatan partisi yang digunakan untuk pengelompokkan data yang mengandung *outlier*.Algoritme *PAM* menggunakan median *(medoid)* sebagai pusat*cluster*, sehingga tidak terpengaruh oleh adanya *outlier*. Algoritme *PAM* hanya dapat digunakan untuk pengelompokkan data berskala kecil dan perhitungan jarak antara objek terhadap pusat *cluster* menggunakan jarak *euclidean*, namun hasil perhitungan jaraknya dapat terpengaruh adanya *outlier*.Algoritme *Clustering Large Application (CLARA)* merupakan algoritme untuk mengelompokkan data berskala besar yang mengandung *outlier* menggunakan teknik pengambilan sampelkemudian menerapkan algoritme *PAM*dengan perhitungan jarak antara objek terhadap pusat *cluster*menggunakan jarak *manhattan*. Pada penelitian ini dilakukan kajian ulang terhadap algoritme CLARA dari aspek kompleksitas dan perhitungan jarak yang digunakan. Hasil menunjukkan bahwa algoritme CLARA dengan jarak manhattan lebih efisien dan akurat dalam mengelompokkan data berskala besar yang mengandung outlier.

Kata kunci:

Clustering, Outlier, K-Medoid, CLARA

1. Pendahuluan

Data *mining* adalah proses dalam *machine learning* yang digunakan untuk menentukan pola dan informasi terkait suatu data. MenurutTaruna R & Hiranwal(2013),data *mining* merupakan proses pegumpulan informasi dari sejumlah data menggunakan teknik-teknik dalam bidang *statistika*. Dalam statistika terdapat beberapa jenis analisis data berdasarkan jumlah variabel, seperti *univariate*, *bivariate* dan *multivariate*. *Clustering* merupakan suatu analisis statistika yang menggunakan lebih dari satu variable atau*multivariate* dalam pengelompokkan objek yang diamati menjadi beberapa *cluster* berdasarkan kemiripan objek (Utami & Saputro, 2018). *Clustering* adalah proses pembentukan *cluster* dimana penempatan objek yang sama ditempatkan dalam kelompok yang sama, namun akan menempati kelompok yang berbeda untuk objek yang tidak sama. Tujuan *clustering*adalah mengelompokkan sejumlah data menjadi beberapa kelompok(Lailiyah & Hafiyusholeh, 2016).

Menurut Johson & Wichern (2014), terdapat dua metode pendekatan dalam *clustering* yaitu pendekatan *non-hierarki* (partisi) dan pendekatan *hierarki*. Pendekatan *hirarki* mengelompokkan objek pengamatan secara terstruktur berdasarkan sifat kemiripan objek, sedangkan pendekatan partisi mengelompokkan objek ke dalam *cluster-cluster* yang sudah ditentukan(Hair *et al.*, 2014).Pada *clustering*, keberadaan *outlier* dapat menyebabkan penyimpangan hasil, sehingga *cluster*yang diperoleh tidak merepresentasikan objek dengan tepat. *Outlier* adalah suatu objek pengamatan yang terletak jauh dari pusat objek dan memiliki karakter unik sehingga terlihat berbeda dengan objek-objek lainnya. *Outlier* dapat muncul dalam bentuk nilai yang sangat rendah atau sangat tinggi(Barnett & Lewis, 1994).

To cite this article:
Fitrayana, P. R. & Saputro, D. R. S. (2022). Algoritme *Clustering Large Application (CLARA)* untuk Menangani Data *Outlier*.*PRISMA, Prosiding Seminar Nasional Matematika* 5, 721-725

Salah satu algoritme dari pendekatan partisi adalah algoritme*k-means*yang mudah diimplementasikan. Pada algoritme *k-means*banyaknya*cluster* yang akan dibentuk ditentukan sebelum dilakukan *clustering*. Menurut Han & Kamber(2012), algoritme*k-means* sensitif terhadap *oulier*, karena algoritme ini menggunakan nilai *rata-rata* sebagai pusat *cluster* sehingga ketika *outlier* dikelompokkan kedalam suatu *cluster* maka *outlier* tersebut dapat mempengaruhi nilai rata-rata dari suatu *cluster*. Dalam mengatasi kelemahan algoritme*k-means*, dikembangkanalgoritme*k-medoid* yang menggunakan median *(medoids)* sebagai pusat *cluster*(Kaufmaan & Rousseeuw, 2009). Dua algoritme ini memiliki cara kerja yang sama dimana hasil *cluster* k dibentuk dengan mengukur jarak objek dengan titik pusat, kemudian objek dikelompokkan dalam *cluster* dengan melihat titik pusat terdekat.

Chu *et al*., (2002)menyatakan algoritme*k-medoid* atau biasa disebut *PAM* dikatakan lebih *robust* terhadap *outlier* jika dibandingkan dengan algoritma pendekatan partisi lainnya. Algoritme*PAM* bekerja optimal terhadap data berskala kecil, namun kurang efektif untuk data berskala besar. Algoritme*PAM* menghasilkan keakuratan sebesar 83%(Nyoman & Smrti, 2015), namun untuk data berskala besar dapat mempengaruhi keakuratan algoritme*PAM*. Keakuratan tersebut dipengaruhi oleh pengukuran jarak objek ke pusat cluster yang menggunakan jarak *Euclidien*. Sehingga dikembangkan algoritme*clustering large application (CLARA)*dengan jarak *manhattan*yang efektif untuk data berskala besar dan *robust* terhadap adanya *outlier*. Oleh karena itu, pada artikel ini dilakukan kajian mengenai algoritme *CLARA* untuk menangani data *outlier* dari aspek kompleksitas dan perhitungan jarak.

2. Pembahasan

Clustering merupakan proses pengelompokkan objek ke dalam suatu*cluster* dimana objekyang berada dalam satu *cluster* mempunyai tingkat kesamaan yang tinggi dan antar *cluster* mempunyaitingkat kesamaan yang rendah(Tan *et al*., 2016). Menurut Halkidi et al.(2001), *clustering* dibagi ke dalam dua pendekatan yaitu pendekatan hirarki dan pendekatan partisi. Pendekatan hirarki (hierarki), yaitu sebuah proses pengelompokkan objek berdasarkan kemiripan karakter yang dilakukan secara terstruktur dengan jumlah *cluster* belum ditentukan. Metode ini dimulai dengan mengelompokkan objek yang memiliki kesamaan ke dalam satu *cluster*. Proses tersebut dilakukan hingga *cluster-cluster* tersebut membentuk sebuah hirarki (tingkatan). Sedangkan pendekatan partisi (non-hierarki), yaitu sebuah proses pengelompokkan objek yang jumlah *cluster* ditentukan diawal sesuai dengan kehendak. Metode ini dimulai dengan menentukan pusat *cluster* di masing-masing *cluster* dan menentukan jarak antara semua objek dan pusat *cluster* tersebut. Objek dengan jarak tertentu ditempatkan pada *cluster* yang ditentukan. Proses tersebut dilakukan hingga setiap *cluster* memiliki anggota objek yang tetap. *Clustering* dapar dilakukan untuk data yang mengandung *outlier.*

2.1. Data Outlier

Outlier atau biasa disebut pencilan merupakan objek pengamatan yang memiliki nilai-nilai ekstrim, baik terlalu rendah atau terlalu tinggi sehingga menyebabkan selisih yang besar dengan objek-objek lainnya. Dalam analisis, *outlier* biasanya mendapatkan perlakuan khusus supaya tidak menyebabkan masalah dalam penentuan hasil akhir analisis. Tidak semua *outlier* dapat dihilangkan karena terkadang *outlier* memiliki nilai yang penting dan berpengaruh dalam kumpulan objek. Terdapat beberapa metode untuk mendeteksi adanya *outlier*, seperti metode *boxplot*, grafik, standarisasi, dan jarak kuadrat *Mahalanobis*. Pada analisis *multivariate*, jarak kuadrat *mahalanobis* digunakan dalam mendeteksi adanya *outlier*(Utami & Saputro, 2018).

2.2. K-Medoid (PAM)

K-medoid merupakan salah satu algoritme dalam pendekatan partisi. Algoritme *k-medoid* dikembangkan pada tahun 1987 oleh Kaufman & Rousseeuw. Algoritme *PAM* merupakan sebuah algoritme yang merepresentasikan *cluster*yang dibentuk menggunakan*medoid* sebagai pusat *cluster*. Algoritme ini mengambil satu objek pada kumpulan objek sebagai perwakilan dari sebuah *cluster* dimana objek tersebut disebut *medoid*(Utami & Saputro, 2018). Secara umum teknik dari *clustering* dengan algoritme *k-medoid* adalah menghitung kedekatan antara *medoid* dengan objek menggunakan perhitungan jarak. Algoritme *k-medoid*ditulis sebagaiberikut.

(1) Menentukan *medoid* awal secara acak dan menentukan jumlah*k cluster.*

(2) Menghitung jarak setiap objek terhadap *medoid* awal yang telah ditentukan pada *cluster* terdekat menggunakan persamaan jarak *Euclidean* yang ditulis sebagai,

$$d(j,k) = ||j-k|| = \sqrt{\sum_{i=1}^{n}(j_i - k_i)\,; n = 1,2,3,\ldots}$$

dengan
d(j,k) : jarak data ke-*j*ke pusat *cluster k*
j_i : data ke-*j*atribut ke-*i*
k_i : pusat data ke-*k* atribut ke-*i*

(3) Mengalokasikan setiap objek ke suatu *cluster* terhadap *medoid* terdekat.
(4) Menentukan calon *medoid* baru pada setiap *cluster* secara acak.
(5) Menghitung jarak setiap objek pada setiap *cluster* dengan calon *medoid* baru.
(6) Mengalokasikan setiap objek ke suatu *cluster* terhadap calon *medoid* baru terdekat.
(7) Menghitung simpangan (S) dengan menentukan selisih antara nilai total jarak objek ke calon *medoid* baru dan total jarak objek ke *medoid* awal. Jika diperoleh S < 0, maka ganti *medoid* awal dengan calon *medoid* baru sebagai *medoid* baru.
(8) Mengulangi langkah 4 hingga 7, dengan sedemikian sehingga diperoleh S > 0 atau *medoid* tidak mengalami perubahan, sehingga diperoleh *cluster* beserta anggota *cluster* masing-masing.

2.3. CLARA

Clustering Large Application (CLARA) merupakan pengembangan dari algoritme *k-medoid* yang *robust* terhadap *outlier* dan dapat digunakan dalam data berskala besar(Rifa *et al*., 2020). Menurut Kaufmaan& Rousseeuw(2009), *CLARA* menerapkan pengambilan sampel data kemudian menerapkan algoritme *PAM* terhadap sampel untuk mendapatkan *medoid* yang optimal.. Algoritme *CLARA* ditulis sebagai

(1) Mengambil sampel dengan ukuran *40+2k* secara acak dari data.
(2) Menentukan*medoid* awal secara acak dan menentukan jumlah*k cluster*.
(3) Menghitung jarak setiap objek terhadap *medoid* awal yang telah ditentukan pada *cluster* terdekat menggunakan persamaan jarak Manhattan.
(4) Mengalokasikan setiap objek ke suatu *cluster* terhadap *medoid* terdekat.
(5) Menentukan calon *medoid* baru pada setiap *cluster* secara acak.
(6) Menghitung jarak setiap objek pada setiap *cluster* dengan calon *medoid* baru.
(7) Mengalokasikan setiap objek ke suatu *cluster* terhadap calon *medoid* baru terdekat.
(8) Menghitung simpangan (S) dengan menentukan selisih antara nilai total jarak objek ke calon *medoid* baru dan total jarak objek ke *medoid* awal. Jika diperoleh S < 0, maka ganti *medoid* awal dengan calon *medoid* baru sebagai *medoid* baru.
(9) Mengulangi langkah 5 hingga 8, dengan sedemikian sehingga diperoleh S > 0 atau *medoid* tidak mengalami perubahan, sehingga diperoleh *cluster* beserta anggota *cluster* masing-masing.

Ukuran sampel yang digunakan pada algoritme *CLARA* bergantung pada jumlah *cluster* yang akan dibuat. Berdasarkan penelitian yang dilakukan oleh Kaufman & Rousseeuw, yaitu*clustering*terhadap 1000 objek menjadi 2 *cluster* dengan 8 variabel menggunakan algoritme CLARA diperoleh bahwa ukuran sampel terbaik untuk *clustering* menggunakan algoritme *CLARA* yaitu sebesar 40 + 2*k* dengan *k* adalah jumlah *cluster* yang akan dibuat. Jumlah *cluster* yang dapat dibuat dari algoritme CLARA harus diantara 1 sampai 30 *cluster* sehingga banyaknya sampel yang digunakan sebanyak 42 sampai 100 sampel. Fungsi ukuran sampel 40+2*k*memiliki probabilitas yang tinggi dan masuk akal untuk menentukan anggota objek dari semua *cluster* yang ada. Pehitungan jarak dalam algoritme *CLARA* menggunakan jarak Manhattan. Persamaan jarak Manhattan ditulis sebagai

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \quad (1)$$

dengan
d_{ij} : jarak objek *i* ke objek *j*
x_{ik} : objek *i* atribut ke-*k*
x_{ik} : objekke-*j* atribut ke-*k*

Jarak Manhattandigunakan untuk menentukan jarak antara dua objek menggunakan perbedaan absolut antar objek, sedangkan jarak *euclidean* digunakan untuk menentukan jarak antar dua objek menggunakan perbedaan kuadrat disetiap objek. Berdasarkan hal tersebut diketahui bahwa jarak *eculidean* akan memperbesar jarak antara dua objek dekat dengan sebagian besar objek yang berbeda dengan salah satu objek, sedangkan jarak Manhattan akan mengabaikannya karena lebih dipengaruhi oleh kedekatan antar objeknya. Oleh karena itu jarak *manhattan* akan memberikan hasil yang lebih kuat terhadap *clustering* data yang mengandung *outlier*, sedangkan jarak *euclidean* cenderung dipengaruhi oleh data *outlier*(Mohibullah *et al*., 2015).

Algoritme *CLARA* bekerja efisien untuk clustering data berskala besar seperti 2000 sampai 3000 objek. Hal tersebut dibuktikan dengan hasil analisis kompleksitas algoritme *CLARA*dan algoritme *PAM*. Komplektitas algortime *PAM* dimana *k(n-k)* sebagai pasangan *medoid* dan *(n-k)* sebagai banyaknya pengujian objek *non-medoid*, diperoleh kompleksitasnya yaitu $O(k(n-k)^2$dengan *n* adalah total objek dan *k* adalah jumlah *cluster*. Hasil komplekstias algoritme *PAM* merupakan kompleksitas dari hanya satu iterasi, sehingga algoritme *PAM* menjadi kurang efisien untuk nilai *n* dan *k* yang besar(Sagvekar *et al*., 2013).

Kompleksitas algoritme *CLARA* yang menerapkan pengambilan sampel dan algortime *PAM* dimana *40+2k* sebagai jumlah sampel yang digunakan dan $O(k(n-k)^2)$ sebagai kompleksitas dari algoritme *PAM*, diperoleh kompleksitas algoritme *CLARA* yaitu $O(k(40+2k)^2 + k(n-k))$ untuk setiap iterasi. Berdasarkan kompleksitasnya, algoritme *CLARA* lebih efisien untuk nilai *n* dan *k* yang besar(Sagvekar *et al*., 2013).

Algoritme *CLARA*menggunakan *medoid* sebagai pusat *cluster* sehingga hasil *clustering* dengan algoritme *CLARA* tidak dipengaruhi adanya *outlier*. Dengan demikian,hasil *clustering* algoritme *CLARA* dapat teralokasi dengan tepat, sehingga dapat dikatakan bahwa *clustering* dengan algoritme *CLARA* memenuhi sifat *robust* terhadap *outlier.*

3. Simpulan

Berdasarkan pembahasan diperoleh simpulan bahwa algoritme *CLARA* memiliki sifat robust sehingga dapat mengatasi adanya data *outlier.* Penggunaan jarak Manhattandalam perhitungan jarak antara objek dan pusat *cluster* memberikan hasil keakuratan yang baik dalam *clustering* data berskala besar karena kompleksitas algoritme *CLARA* yang digunakan untuk setiap iterasi.

Daftar Pustaka

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons.

Chu, S.-C., Roddick, J. F., & Pan, J. S. (2002). Efficient K-Medoids Algorithms Using Multi-Centroids With Multi-Runs Sampling Scheme. In *Workshop on Mining Data for CRM* (pp. 14–25).

Hair, J., Black, W., Babin, B., & Anderson, R. (2014). *Pearson New International Edition. Multivariate Data Analysis*. Pearson Education Limited Harlow.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, *17*(2), 107–145. https://doi.org/10.1007/978-3-642-48618-0_7

Han, J., & Kamber, M. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher.

Johson, R., & Wichern, D. (2014). *Applied Multivariate Statistical Analysis*. Pearson.

Kaufmaan, L., & Rousseeuw, P. (2009). *Finding Groups in Data : an Introduction to Cluster Analysis* (Vol. 148). John Wiley & Sons.

Lailiyah, S., & Hafiyusholeh, M. (2016). Perbandingan Antara Metode K-Means Clustering Dengan Gath-Geva Clustering. *Jurnal Matematika "MANTIK,"1*(2), 26–37. https://doi.org/10.15642/mantik.2016.1.2.26-37

Mohibullah, M., Hossain, M. Z., & Hasan, M. (2015). Comparison of Euclidean Distance Function and Manhattan Distance Function Using K-Mediods. *International Journal of Computer Science and Information Security (IJCSIS)*, *13*(10), 61–71.

Nyoman, N., & Smrti, E. (2015). Otomatisasi Klasifikasi Buku Perpustakaan Dengan Menggabungkan Metode K-Nn Dengan K-Medoids. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *4*(1), 201–214.

Sagvekar, V., Sagvekar, V., & Deorukhkar, K. (2013). Performance assessment of CLARANS: A Method for Clustering Objects for Spatial Data Mining. *Global Journal of Engineering, Design & Technology/Global Institute for Reserach & Education*, *2*(6), 1–8. http://gifre.org/library/upload/volume/1-8-vol-2-6-13-gjedt.pdf

Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to Data Mining*. Pearson Education.

Taruna R, S., & Hiranwal, S. (2013). Enhanced Naïve Bayes Algorithm for Intrusion Detection in Data Mining. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *4*(6), 960–962.

Utami, D. S., & Saputro, D. R. S. (2018). Pengelompokkan Data yang Memuat Pencilan dengan Kriteria Elbow dan Koefisien Silhouette (Algoritme K-Medoid). *Konferensi Nasional Penelitian Matematika Dan Pembelajaran (KNPMP) III*, 448–456.