JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 12 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v12i3.4997

PENENTUAN PUSAT KLASTER SECARA OTOMATIS PADA ALGORITMA DENSITY PEAKS CLUSTERING BERBASIS METODE INTER QUARTILE RANGE

Ridwan Efendi[1], Achmad Junaidi[2], Agung Mustika Rizki[3]

[1,2,3]Universitas Pembangunan Nasional "Veteran" Jawa Timur, Jl. Rungkut Madya No.1, Gn. Anyar, Surabaya

Received: 14 Juli 2024
Accepted: 31 Juli 2024
Published: 7 Agustus 2024

Keywords:
Clustering;
Density Peaks Clustering;
Deteksi Pusat Otomatis;

Corespondent Email:
20081010071@student.upnjatim.ac.id

Abstrak. *Clustering* adalah sebuah metode untuk mengelompokkan data yang sejenis ke dalam satu bagian yang sama. Proses ini mampu membantu manusia untuk mendapatkan informasi secara lebih cepat. Dalam konteks media sosial misalnya, metode *clustering* dapat memberikan informasi terkait konten yang cenderung disukai dan kurang disukai. Algoritma Density Peaks Clustering (DPC) adalah salah satu algoritma yang cukup populer digunakan untuk mengelompokkan sebuah data. Sudah banyak penelitian yang menggunakan algoritma ini. Namun, algoritma DPC memiliki kekurangan dalam hal penentuan pusat klaster. Pusat Klaster dalam algoritma DPC masih dipilih secara manual melalui grafik keputusan. Pemilihan pusat klaster secara otomatis menambah subjektivitas dan ketidakstabilan dalam algoritma. Untuk mengatasi masalah tersebut, diusulkan sebuah algoritma 'Penentuan Pusat Otomatis' yang berbasis pada metode Inter Quartile Range (IQR). Algoritma ini diuji menggunakan dataset iris, aggregation, flame, dan spiral. Hasil pengujian menunjukkan bahwa algoritma yang diusulkan dapat memperoleh hasil *clustering* yang lebih baik dan lebih akurat.

Abstract. *Clustering is a method for grouping similar data into the same segment. This process helps humans to obtain information more quickly. In the context of social media, for example, the clustering method can provide information about content that tends to be liked or disliked. The Density Peaks Clustering (DPC) algorithm is one of the more popular algorithms used for clustering data. Many studies have utilized this algorithm. However, the DPC algorithm has a drawback in determining cluster centers. Cluster centers in the DPC algorithm are still manually chosen through a decision graph. The manual selection of cluster centers introduces subjectivity and instability into the algorithm. To address this issue, an 'Automatic Center Determination' algorithm based on the Inter Quartile Range (IQR) method is proposed. This algorithm is tested using the iris, aggregation, flame, and spiral datasets. The test results show that the proposed algorithm can achieve better and more accurate clustering results.*

1. PENDAHULUAN

Seiring dengan pesatnya perkembangan teknologi, internet, dan banyaknya perangkat elektronik, data pun akan bertambah semakin cepat. Pada tahun 2017 saja terdapat 2.5 triliun bytes data yang dihasilkan per hari [1]. Menurut perkiraan terbaru, saat ini data dihasilkan mencapai 328.77 juta *terabytes* per hari [2]. Dengan banyaknya data yang ada, kita membutuhkan metode canggih yang dapat secara otomatis melakukan analisis dan mengenali pola-pola sehingga dapat menghasilkan sebuah informasi yang berguna. Contohnya dalam konteks media sosial,

pengguna akan cenderung memposting opini mengenai produk maupun orang (*content creator*), data opini ini dapat dikumpulkan dan diproses untuk mengetahui aspek apa yang banyak disukai dan tidak disukai [3]. Dalam hal ini, metode *clustering* merupakan salah satu metode dapat digunakan untuk menganalisis sebuah data secara otomatis.

Clustering merupakan salah satu algoritma *unsupervised learnin*g yang paling penting [4][5]. Untuk mencari informasi pada data yang kompleks, kita dapat melakukan *clustering* data terlebih dahulu, yaitu dengan mengelompokkan data yang sejenis dalam satu kelas sesuai dengan karakteristik data. Metode *clustering* telah banyak mengalami perkembangan dan telah diterapkan secara luas pada data *mining* [6], *image processing* [7], *bioinformatics* [8], *recommendation* [9], dan masih banyak lagi. Sampai saat ini, banyak algoritma *clustering* yang telah dikembangkan, diantaranya adalah *K-means clustering, fuzzy C-means clustering (FCM), Density-based Spatial Clustering of Application (DBSCAN),* dan lainnya.

Density-Based Clustering merupakan salah satu algoritma yang paling penting dan banyak digunakan [10]. *Density-Based Clustering* bergantung pada gagasan menemukan kepadatan pada suatu wilayah dalam kumpulan data [11]. Tujuannya adalah untuk menemukan klaster pada tingkat yang berbeda-beda. Dalam *Density-Based Clustering*, daerah dengan kepadatan lebih tinggi akan dipisahkan dan diidentifikasi sebagai pusat klaster, sedangkan daerah dengan kepadatan lebih rendah digunakan sebagai partisi [1]. Algoritma *Density-Based Clustering* tidak membutuhkan jumlah klaster sebagai parameter masukan [12]. Algoritma-algoritma yang termasuk ke dalam kategori *Density-Based Clustering*, yaitu *Density-based Spatial Clustering of Application (DBSCAN), Mean Shift, Spectral Method, Subtractive Method*, dan lain-lain.

Pada tahun 2014 Rodriguez dan Laio mengusulkan sebuah algoritma canggih yang bernama *Density Peaks Clustering (DPC)* dan termasuk dalam kategori *Density-Based Clustering*. Sejak pertama kali diusulkan, algoritma DPC telah banyak diterapkan pada berbagai penelitian [13]. Algoritma DPC memiliki prinsip yang sederhana dan efisiensi yang tinggi [14]. Seperti algoritma *Density-Based* lainnya, algoritma DPC juga menghitung kepadatan untuk setiap titik data. Proses *clustering* dari algoritma DPC dimulai dengan menentukan puncak atau pusat klaster, kemudian data lainnya akan digabungkan dengan data yang terdekat dan memiliki nilai kepadatan lebih tinggi.

Namun algoritma DPC masih memiliki masalah dalam penentuan pusat klaster. Penentuan pusat klaster masih dilakukan secara manual melalui grafik keputusan dengan melihat titik-titik data yang mempunyai nilai kepadatan (ρ) dan jarak lokal (δ) yang relatif tinggi, atau dapat diidentifikasi dengan memilih titik-titik data yang berada pada kanan atas grafik keputusan. Namun, seberapa besar nilai yang dianggap tinggi? Tidak ada standar yang jelas terkait batasan sebuah nilai yang akan dianggap tinggi. Proses ini akan menjadi sangat subjektif tergantung dari batasan yang ditetapkan oleh masing-masing pengguna. Selain itu, ketika grafik keputusan tidak dapat menunjukkan pusat klaster secara akurat, maka semakin sulit untuk memilihnya secara manual melalui grafik keputusan.

Penelitian terdahulu yang berkaitan dengan algoritma DPC juga masih mengabaikan masalah pemilihan pusat klaster yang manual. Algoritma Density Peaks Clustering Based on K-Nearest Neighbors and Principal Component Analysis (DPC-KNN dan DPC-KNN-PCA) telah diusulkan untuk memperbaiki kekurangan pada rumus ρ dan meningkatkan akurasi pada data berdimensi tinggi [15]. Namun, sulit untuk menentukan nilai parameter d_c atau p yang tepat dan pusat klaster tidak dipilih secara otomatis. Selanjutnya, terdapat algoritma Density Peaks Clustering Based on Density Backbone and Fuzzy Neighborhood (DPC-DBFN) yang dapat mengatasi masalah seperti sensitivitas terhadap nilai d_c, kurangnya efektivitas dalam perhitungan nilai ρ, dan reaksi berantai pada algoritma DPC [16]. Namun, algoritma DPC-DBFN masih memerlukan kebutuhan untuk menentukan pusat klaster secara manual melalui nilai parameter c. Algoritma Density Peaks Clustering and Gravitational Search Method (GSA-DPC) telah diusulkan untuk mengatasi masalah yang timbul akibat nilai d_c dalam menentukan pusat klaster [17]. Namun, algoritma GSA-DPC tidak bisa memilih pusat klaster secara otomatis, karena masih membutuhkan bantuan manusia

untuk memilih pusat klaster dengan menentukan nilai parameter *k*.

Oleh karena itu, untuk mengatiasi masalah tersebut, dalam penelitian ini diusulkan sebuah algoritma "Penentuan Pusat Otomatis" yang berbasis pada metode *Inter Quatile Range* (IQR). Gabungan antara algoritma DPC dan "Penentuan Pusat Otomatis" disebut dengan algoritma DPC-IQR.

2. TINJAUAN PUSTAKA

2.1. *Density Peaks Clustering*

Algoritma DPC diusulkan oleh Rodriguez dan Laio pada jurnal Science AS tahun 2014. Ide dibalik algoritma DPC adalah pusat-pusat klaster ditandai oleh kepadatan yang lebih tinggi daripada tetangga-tetangga mereka dan jarak yang relatif besar dari titik-titik dengan kepadatan yang lebih tinggiAlgoritma DPC memiliki 2 variabel penting, yaitu kepadatan lokal *(ρ)* dan jarak lokal *(δ)*.

Algoritma DPC perlu menghitung matrik jarak Euclidean dari kumpulan data terlebih dahulu. Jarak Euclidean didefinisikan dengan persamaan berikut:

$$d(x_i, x_j) = \|x_i \times x_j\|_2 \quad (1)$$

Kepadatan lokal *(ρ)* dari titik x_i, dilambangkan dengan ρ_i dan didefinisikan sebagai berikut:

$$\rho_i = \sum_j \exp\left(-\frac{d(x_i,x_j)^2}{{d_c}^2}\right) \quad (2)$$

d_c merupakan satu-satunya variabel pada Persamaan 2. d_c diberi nilai tertentu untuk membuat jumlah rata-rata tetangga sekitar 2% dari total jumlah objek dalam kumpulan data [18]. Untuk mendapatkannya d_c yang optimal, dapat digunanakan nilai parameter (*p*) $1 \leq p \leq 2$ [19]. Persamaan d_c didefinisikan sebagai berikut:

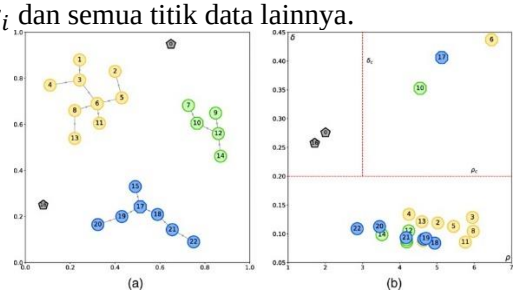
$$d_c = d_{\left\lceil N_d \times \frac{p}{100} \right\rceil} \quad (3)$$

Dimana $N_d = \binom{N}{2}$ dan dan $d_{\left\lceil N_d \times \frac{p}{100} \right\rceil} = \epsilon D = [d_1, d_2, \dots, d_{N_d}]$. *N* merupakan banyak data dalam kumpulan data. *D* adalah himpunan semua jarak antara dua titik dalam kumpulan data yang telah diurutkan secara menaik (*ascending*). $\left\lceil N_d \times \frac{p}{100} \right\rceil$ merupakan *subscript* dari $d_{\left\lceil N_d \times \frac{p}{100} \right\rceil}$, dimana $\lceil * \rceil$ adalah *ceiling function.*

Nilai jarak lokal *(δ)* dihitung dengan mencari jarak $d(x_i, x_j)$ minimum antara titik x_i dan titik lain yang memiliki kepadatan yang lebih tinggi, didefinisikan sebagai berikut:

$$\delta_i = \begin{cases} min_{j:\rho_i<\rho_j}\left(d(x_i,x_j)\right) \\ max_j\left(d(x_i,x_j)\right) \end{cases} \quad (4)$$

Perhatikan jika sebuah titik memiliki nilai kepadatan tertinggi, maka nilai δ ditetapkan sebagai jarak $d(x_i, x_j)$ maksimum antara titik x_i dan semua titik data lainnya.



Gambar 1. DPC dalam Grafik dua Dimensi. (A) Distribusi data. (B) Grafik Keputusan

Nilai ρ_i dan δ_i digunakan untuk membuat sebuah grafik keputusan dengan ρ_i sebagai sumbu x dan δ_i sebagai sumbu y, yang digunakan agar pengguna dapat memilih pusat klaster dari suatu dataset. Titik data dengan nilai ρ dan δ yang relatif tinggi dapat dipilih sebagai pusat klaster (Gambar 1B). Setelah pusat klaster ditemukan, algoritma DPC menugaskan titik-titik data yang tersisa ke klaster yang sama dengan tetangga terdekatnya yang memiliki nilai kepadatan lebih tinggi (Gambar 1A). Proses ini didefinisikan dengan persamaan berikut:

$$Cluster(i) = Cluster(min\_num(i)) \quad (5)$$

2.2. *Interquartile Range*

Metode *Inter Quartile Range (IQR)* adalah sebuah metode statistika deskriptif yang digunakan untuk mengukur sebaran atau keragaman data dalam sebuah himpunan data. IQR digunakan untuk mengetahui seberapa jauh nilai-nilai data tersebar dari nilai tengahnya. IQR memberikan informasi tentang sebaran data di sekitar nilai tengah (median) tanpa terpengaruh oleh nilai-nilai ekstrim (*outlier*) yang ada pada dataset. Semakin besar nilai IQR, semakin besar keragaman data di

dalam himpunan data tersebut. IQR juga dapat digunakan untuk mendeteksi *outlier* dalam sebuah dataset. Nilai-nilai di luar rentang batas bawah dan batas atas dapat dianggap sebagai *outlier* [20] [21].

3. METODE PENELITIAN

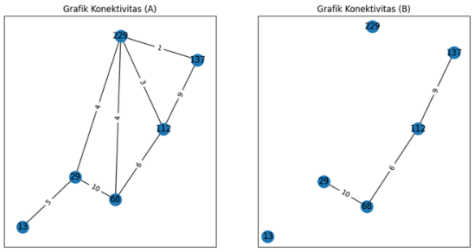
Untuk menyelesaikan permasalah pada algoritma DPC, maka perlu dilakukan perbaikan pada algoritma DPC. Pada bagian ini, akan diberikan penjelasan maupun cara kerja terkait algoritma 'Penentuan Pusat Otomatis' untuk mengatasi masalah pada algoritma DPC, yang telah dijelaskan sebelumnya. Selain itu, juga terdapat informasi terkait metode evaluasi yang digunakan untuk mengukur kualitas hasil *clustering* dari algoritma DPC-IQR.

3.1. Algoritma Penentuan Pusat Otomatis

$$\gamma_i = \rho_i \times \delta_i \quad (6)$$

Algoritma ini dimulai dengan merubah grafik keputusan dari dua dimensi data (ρ_i dan δ_i) menjadi satu dimensi data gamma (γ). Nilai gamma diperoleh dengan mengalikan ρ_i dan δ_i (Persamaan 6). Titik dengan nilai *gamma* yang lebih dari batas atas dianggap sebagai pusat klaster. Batas atas dihitung dengan rumus:

$$batas\ atas = (Q_3 \times 2) + (1.5 \times IQR) \quad (7)$$



Gambar 1. Konektivitas

Dari langkah awal tersebut akan dihasilkan jumlah klaster awal. Banyak klasternya sesuai dengan banyak pusat klaster yang diperoleh. Dari sini akan dihitung nilai konektivitas antar dua klaster (Gambar 1A). Nilai konektivitas didefinisikan sebagai berikut:

$$C(A,B) = \sum_{i \in A} Cross_k(i,B) + \sum_{i \in B} Cross_k(i,A) \quad (8)$$

Dimana $Cross_k(i,B)$ meruapakan jumlah tetangga dari titik ke-i pada klaster A yang berada pada klaster B dengan mempertimbangkan *k* tetangga terdekat, $Cross_k(i,A)$ berarti sebaliknya. Jumlah *k* adalah dua persen dari total data.

Proses selanjutnya adalah memfilter seluruh nilai konektivitas yang telah diperoleh. Nilai konektivitas yang kurang dari rata-rata semua nilai konektivitas akan dihapus (Gambar 1B). Klaster-klaster yang masih memiliki nilai konektivitas berarti mereka seharusnya berada dalam satu klaster yang sama. Sehingga akhirnya didapatkan pusat klaster atau jumlah klaster yang optimal.

3.2. Metode Evaluasi

Kualitas dan keakuratan algoritma DPC-IQR diverifikasi melalui pengujian dan evaluasi terhadap beberapa data. Data yang digunakan dalah dataset (kumpulan data) iris, aggregation, flame dan spiral. Informasi detail mengenai kumpulan data ini disajikan dalam Tabel 1, termasuk jumlah fitur, jumlah data, dan jumlah klaster untuk setiap kumpulan data.

Tabel 1. Detail Kumpulan Data

Dataset	Fitur	Data	Klaster
Iris	3	150	3
Aggregation	2	778	7
Flame	2	240	2
Spiral	2	312	3

Hasil *clustering* dari algoritma DPC-IQR dibandingkan dengan beberapa algoritma klastering lainnya, yaitu DPC [22], GB-DPC [3], dan K-Means [23]. Dalam penelitian ini digunukan dua metrik evaluasi yang populer untuk membandingkan hasil *clustering* dari algoritma yang digunakan. Kedua metrik tersebut adalah *Adjusted Rand Index* (ARI) [24] dan *Normalized Mutual Information* (NMI) [25].

Tabel 2. Parameter untuk pengujian

Parameter	Algoritma	Keterangan
p = 2	DPC, DPC-IQR, dan GB-DPC	p digunakan untuk menghitung nilai radius d_c
k	K-means	k adalah jumlah klaster pada Tabel 1

Tabel 2 menjelaskan parameter-parameter yang digunakan dalam pengujian algoritma. Parameter pertama, $p = 2$, diterapkan pada algoritma DPC, DPC-IQR, dan GB-DPC untuk menghitung nilai radius d_c. Sedangkan *k* digunakan untuk algoritma K-means yang

menunjuukan jumlah klaster yang ada pada Tabel 1. Untuk algoritma DPC, pemilihan pusat klaster secara manual melalui grafik keputusan dilakukan oleh 2 orang responden. Tujuannya adalah untuk menghindari bias.

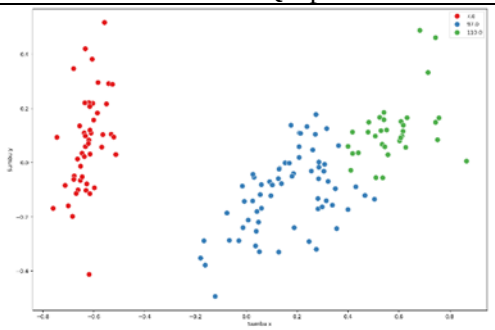
4. HASIL DAN PEMBAHASAN

4.1. DPC-IQR

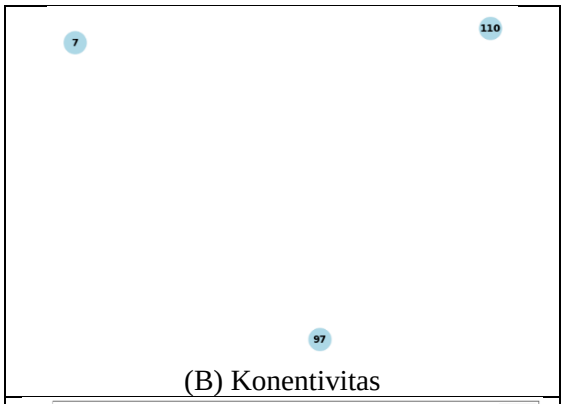
Algoritma 1: DPC-IQR
Input: Dataset: $x \in X_{i\times j}$ Parameter p Output: Label setiap data (*clusters*): $y \in X_{i\times 1}$ Algoritm: 1. Hitung nilai jarak Euclidean (Persamaan 1) 2. Hitung nilai d_c (Persamaan 3) 3. Hitung nilai ρ (Persamaan 2) 4. Hitung nilai δ (Persamaan 4) 5. Hitung nilai *gamma* (γ) (Persamaan 6) 6. Gunakan Persamaan 5 untuk mendapatkan pusat klaster: $c = \{c_1, c_2, \dots, c_t\}$. 7. Mengelompokkan data yang tersisa ke dalam klaster lokal (Persamaan 7). 8. Hitung konektivitas (Persamaan 8). 9. Filter konektivitas. 10. Perbaharui pusat klaster. 11. Jalankan langkah ke-7. 12. Return *y*.

Algoritma 1 menunjukkan langkah-langkah dari algoritma DPC-IQR dalam mengklasterkan sebuah dataset. Secara total terdapat 12 langkah yang harus dilakukan dan langkah ke-5 sampai ke-10 merupakah algoritma "Penentuan Pusat Otomatis" berbasis metode IQR yang diusulkan dalam penelitian ini. Penerapannya dalam kode program dapat diakses melalui tautan berikut: https://colab.research.google.com/drive/1Es8nP2g8rPUT_AEi0z6r3eM_dDzUfgdF?usp=sharing.

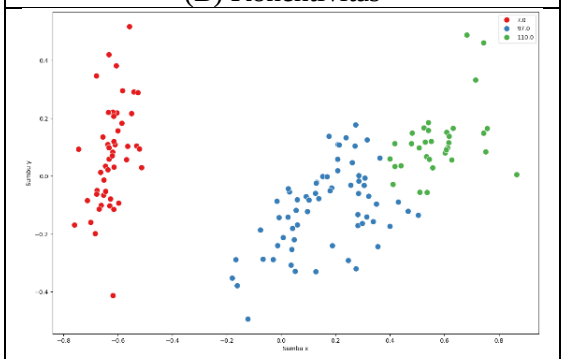
Tabel 3. Proses DPC-IQR pada dataset iris



(A) Klaster Awal



(B) Konentivitas

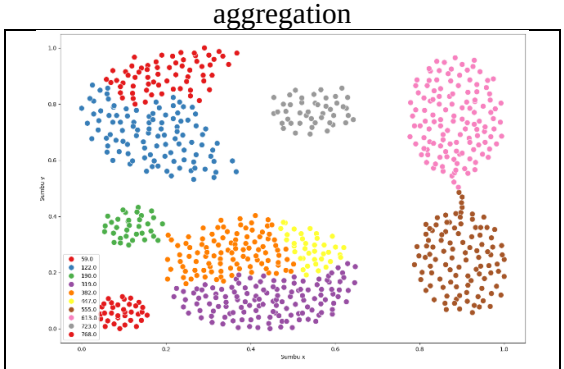


(C) Klaster Akhir

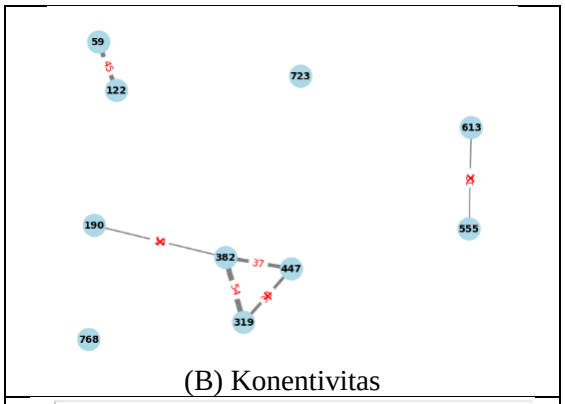
Tabel 3 menunjukkan proses algoritma DPC-IQR dalam mengklasterkan dataset iris. Gambar A pada Tabel 3 menunjukkan 3 jumlah klaster awal. Gambar B pada Tabel 3 menunjukkan bahwa tidak ada konektivitas antar titik (klaster). Oleh karena itu, tidak akan terjadi perubahan jumlah pusat (klaster) yang dihasilkan. Buktinya dapat dilihat pada Gambar A dan C pada Tabel 3 yang identik.

Klaster-klaster pada Gambar A di Tabel 3,4,5,6 diperoleh setelah menjalankan langkah ke-7 pada Algoritma DPC-IQR. Langkah ke-8 dan ke-9 pada Algoritma DPC-IQR akan menghasilkan data untuk Gambar B pada Tabel 3,4,5,6. Sedangkan, klaster-klaster pada Gambar C di Tabel 3,4,5,6 diperoleh dari langkah ke-11 Algoritma DPC-IQR.

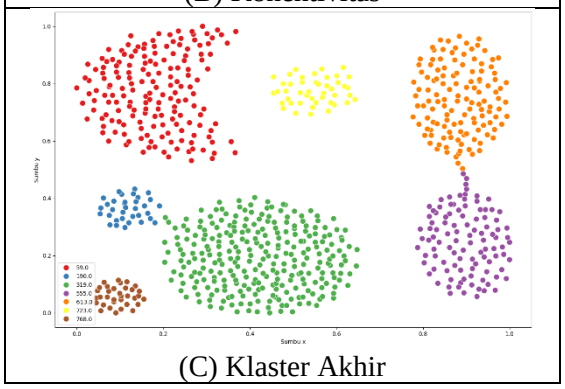
Tabel 4. Proses DPC-IQR pada dataset aggregation



(A) Klaster Awal



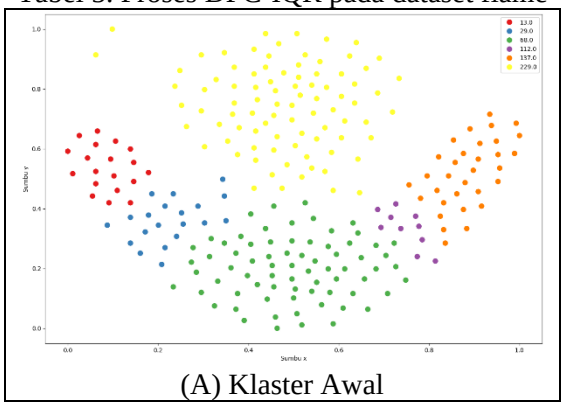
(B) Konentivitas



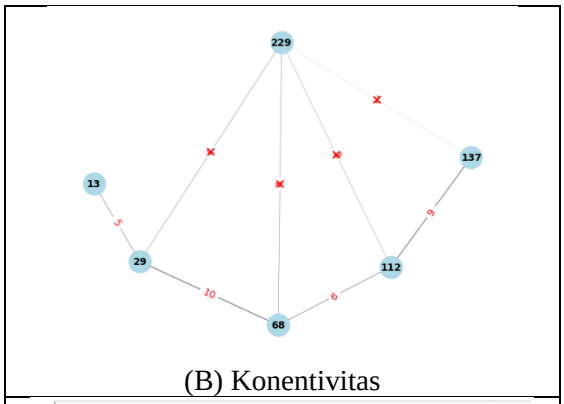
(C) Klaster Akhir

Tabel 4 menunjukkan proses algoritma DPC-IQR dalam mengklasterkan dataset aggregation. Pada Tabel 4 Gambar A terlihat terdapat 10 klaster awal yang dihaslikan oleh algoritma DPC-IQR. Titik pusat dari masing-masing klaster yaitu, 319, 613, 59, 723, 768, 382, 555, 190, 122, dan 447. Dari Tabel 4 Gambar 2, terlihat bahwa titik {59, 122} dan {319, 382, 387} memiliki konektivitas yang cukup. Sehingga, titik 122 akan digabungkan dengan titik 59 dan titik {382, 387} digabungkan dengan titik 319. Sehingga jumlah klaster atau pusatnya akan berubah menjadi 7 (Gambar C pada Tabel 4).

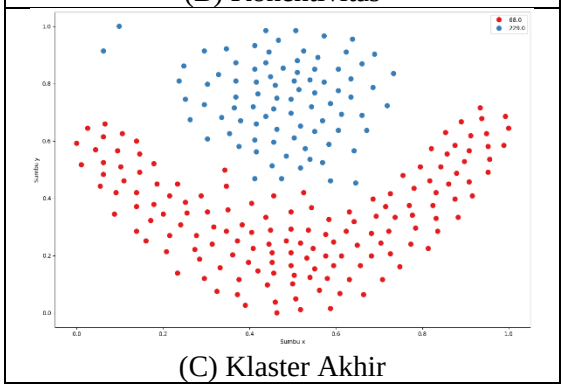
Tabel 5. Proses DPC-IQR pada dataset flame



(A) Klaster Awal



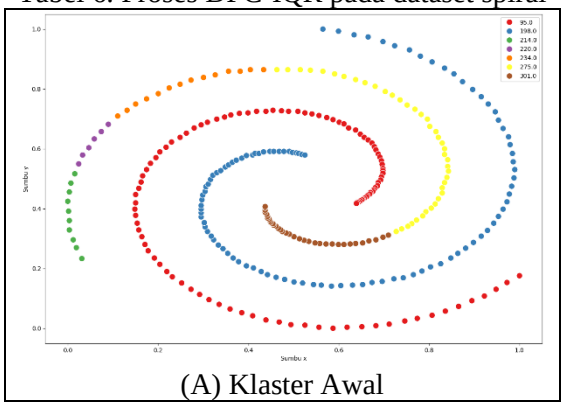
(B) Konentivitas



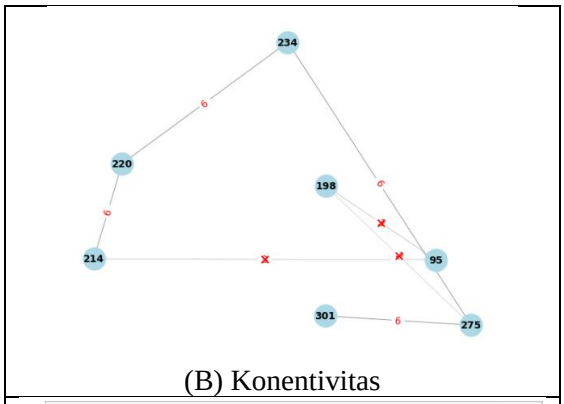
(C) Klaster Akhir

Tabel 5 menunjukkan proses algoritma DPC-IQR dalam mengklasterkan dataset flame. Pada Tabel 5 Gambar A terlihat terdapat 6 klaster awal yang dihaslikan oleh algoritma DPC-IQR. Titik pusat dari masing-masing klaster yaitu, 229, 68, 137, 13, 29, dan 112. Dari Tabel 5 Gambar 2, terlihat bahwa titik {68, 137, 13, 29, 112} memiliki garis (konektivitas) yang cukup. Sehingga, titik {137, 13, 29, 112} akan digabungkan dengan titik 68, karena 68 memiliki nilai kepadatan lokal yang lebih tinggi. Sehingga jumlah klaster atau pusatnya akan berubah menjadi 2 (Gambar C pada Tabel 5).

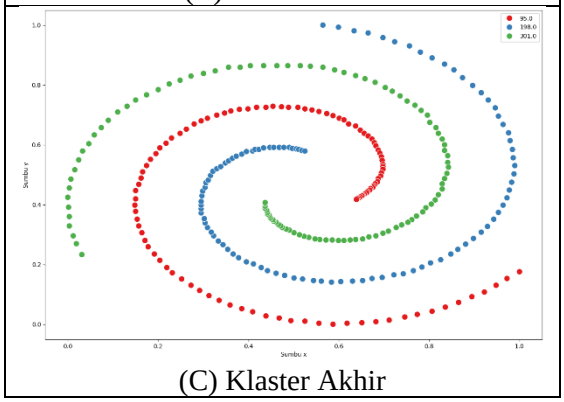
Tabel 6. Proses DPC-IQR pada dataset spiral



(A) Klaster Awal



(B) Konentivitas



(C) Klaster Akhir

Tabel 6 menunjukkan proses algoritma DPC-IQR dalam mengklasterkan dataset spiral. Pada Tabel 6 Gambar A terlihat terdapat 7 klaster awal yang dihaslikan oleh algoritma DPC-IQR. Titik pusat dari masing-masing klaster yaitu, 95, 301, 198, 220, 275, 214, dan 234. Dari Tabel 6 Gambar 2, terlihat bahwa titik {234, 220, 234, 275, 301} memiliki garis (konektivitas) yang cukup. Sehingga, titik {234, 220, 234, 275} akan digabungkan dengan titik 301, karena titik 301 memiliki nilai kepadatan lokal yang lebih tinggi. Sehingga jumlah klaster atau pusatnya akan berubah menjadi 3 (Gambar C pada Tabel 6).

4.2. Evaluasi Algoritma

Tabel 7. Hasil *clustering* menggunakan ARI

Dataset	DPC-IQR	GB-DPC	DPC	K-Means
Iris	**0,73**	0,45	0,56	0,72
Aggregation	**0,99**	0,6	0,71	0,73
Flame	**1**	0,4	0,45	0,46
Spiral	**1**	0,38	**1**	-0,01

Tabel 7 menunjukkan hasil evaluasi metode *clustering* menggunakan Adjusted Rand Index (ARI) pada berbagai dataset. ARI merupakan metrik yang mengukur kesesuaian hasil *clustering* dengan *ground truth*. Untuk dataset Iris, metode DPC-IQR memperoleh nilai ARI tertinggi sebesar 0,73, menunjukkan hasil *clustering* yang sangat baik. K-Means juga memberikan hasil yang baik dengan ARI 0,72, sedangkan metode DPC dan GB-DPC memiliki nilai yang lebih rendah, yaitu 0,56 dan 0,45. Pada dataset Aggregation, DPC-IQR mencatatkan performa terbaik dengan ARI mencapai 0,99, menandakan *clustering* yang sangat akurat. K-Means dan DPC mengikuti dengan nilai ARI 0,73 dan 0,71, sedangkan GB-DPC menunjukkan hasil terendah di 0,60. Untuk dataset Flame, DPC-IQR menunjukkan hasil *clustering* yang optimal dengan ARI 1,00, sedangkan K-Means dan DPC memperoleh nilai ARI masing-masing 0,46 dan 0,45, sementara GB-DPC memiliki ARI 0,40. Pada dataset Spiral, baik DPC-IQR maupun DPC mendapatkan ARI maksimal 1,00, menunjukkan performa *clustering* yang sangat baik. Sebaliknya, GB-DPC dan K-Means menunjukkan hasil yang kurang memuaskan dengan ARI masing-masing 0,38 dan -0,01. Secara umum, DPC-IQR tampil sebagai metode yang paling konsisten dalam memberikan hasil *clustering* terbaik di berbagai dataset. Hasil ini menunjukkan bahwa DPC-IQR cenderung lebih stabil dan akurat dalam berbagai kondisi dibandingkan metode lainnya

Tabel 8. Hasil *clustering* menggunakan NMI

Dataset	DPC-IQR	GB-DPC	DPC	K-Means
Iris	**0,79**	0,65	0,73	0,76
Aggregation	**0,99**	0,8	0,90	0,83
Flame	**1**	0,56	0,60	0,40
Spiral	**1**	0,43	**1**	0,03

Tabel 8 menyajikan hasil evaluasi metode *clustering* menggunakan Normalized Mutual Information (NMI) pada beberapa dataset. NMI adalah metrik yang digunakan untuk mengukur seberapa baik hasil *clustering* dapat mengidentifikasi struktur yang benar dalam data. Pada dataset Iris, metode DPC-IQR menunjukkan performa terbaik dengan nilai NMI sebesar 0,79, diikuti oleh K-Means dengan nilai 0,76. Metode DPC memperoleh nilai NMI 0,73, sedangkan GB-DPC memiliki nilai terendah di 0,65. Untuk dataset Aggregation, DPC-IQR juga mencatatkan hasil yang sangat baik dengan nilai NMI 0,99, menandakan kemampuan *clustering* yang sangat tinggi dalam mengidentifikasi struktur data. Metode DPC dan K-Means menunjukkan nilai NMI

masing-masing 0,90 dan 0,83, sementara GB-DPC berada di posisi terendah dengan nilai 0,80. Pada dataset Flame, DPC-IQR menunjukkan hasil *clustering* yang optimal dengan nilai NMI 1,00, menunjukkan kesesuaian yang sempurna dengan struktur data. Sebaliknya, K-Means hanya mencapai nilai NMI 0,40, dengan DPC dan GB-DPC berada di nilai 0,60 dan 0,56. Untuk dataset Spiral, baik DPC-IQR maupun DPC memperoleh nilai NMI maksimal 1,00, menunjukkan bahwa keduanya sangat efektif dalam mengidentifikasi struktur data yang benar. K-Means, di sisi lain, menunjukkan hasil yang sangat rendah dengan nilai NMI 0,03, sedangkan GB-DPC memiliki nilai 0,43. Secara keseluruhan, algoritma DPC-IQR secara konsisten menunjukkan performa yang sangat baik di berbagai dataset, dengan nilai NMI tertinggi pada sebagian besar kasus, mengindikasikan kemampuannya dalam menghasilkan *clustering* yang lebih akurat dibandingkan metode lainnya.

5. KESIMPULAN

Dalam penelitian ini, diusulkan sebuah algoritma baru yang dinamakan "Penentuan Pusat Otomatis" (DPC-IQR) untuk mengatasi kekurangan yang ada pada algoritma Density Peaks Clustering (DPC). Algoritma ini menggunakan metode Inter Quartile Range (IQR) sebagai basis untuk memilih pusat klaster secara otomatis, sehingga menghilangkan subjektivitas dan kesalahan yang timbul dari pemilihan pusat secara manual (campur tangan manusia). Hasil evaluasi menunjukkan bahwa algoritma DPC-IQR menghasilkan klaster-klaster yang lebih akurat pada berbagai dataset dibandingkan dengan algoritma DPC, Gab-Based Density Peaks Clustering (GB-DPC), dan K-Means. Berikut adalah hasil evaluasi berdasarkan nilai Adjusted Rand Index (ARI) dan Normalized Mutual Information (NMI) pada berbagai dataset: Dataset Iris memiliki nilai ARI sebesar 0,73 dan NMI sebesar 0,79, dataset Aggregation memiliki nilai ARI sebesar 0,99 dan NMI sebesar 0,99, dataset Flame memiliki nilai ARI sebesar 1 dan NMI sebesar 1, serta dataset Spiral memiliki nilai ARI sebesar 1 dan NMI sebesar 1. Hasil ini menunjukkan bahwa algoritma DPC-IQR memiliki kinerja yang baik dalam menentukan pusat klaster dan menghasilkan kalster-klaster yang lebih akurat dibandingkan metode lainnya.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] A. Ghosal, A. Nandy, A. K. Das, and S. G. and M. Panday, "A Short Review on Different Clustering Techniques and Their Applications," *Springer Link*, pp. 69–83, 2019.

[2] F. Duarte, "Amount of Data Created Daily (2024)."

[3] K. G. Flores and S. E. Garza, "Density Peaks Clustering With Gap-Based Automatic Center Detection," *Knowl Based Syst*, 2020.

[4] X. Xu, S. Ding, L. Wang, and Y. Wang, "A Robust Density Peaks Clustering Algorithm With Density-Sensitive Similarity," *Knowl Based Syst*, 2020.

[5] F. Salsabila, T. Ridwan, and H. H, "Analisa Volume Penyebaran Sampah Di Karawang Menggunakan Algoritma K-Means Clustering," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4226.

[6] D. N. Batubara, A. P. Windarto, A. Wanto, D. Hartama, and E. Irawan, "Penerapan Datamining Klastering Pada Perusahaan Industri Mikro di Indonesia," *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, pp. 330–335, 2020.

[7] M. Hu, Y. Zhong, S. Xie, H. Lv, and Z. Lv, "Fuzzy System Based Medical Image Processing for Brain Disease Prediction," *Front Neurosci*, vol. 15, 2021.

[8] Y. Zhang and H. Kiryu, "MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes," *Brief Bioinform*, vol. 23, no. 6, 2022.

[9] A. Saputra, B. Mulyawan, and T. Sutrisno, "Rekomendasi Lokasi Wisata Kuliner Di Jakarta Menggunakan Metode K-Means Clustering Dan Simple Additive Weighting," *Jurnal Ilmu Komputer Dan Sistem Informasi*, vol. 7, 2019.

[10] Z. Zhang *et al.*, "Density decay graph-based density peak clustering," *Knowl Based Syst*, 2021.

[11] P. Bhattacharjee and P. Mitra, "A survey of density based clustering algorithms," *Front Comput Sci*, vol. 15, no. 1, p. 151308, 2020, doi: 10.1007/s11704-019-9059-3.

[12] R. J. G. B. Campello, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *WIREs*

Data Mining and Knowledge Discovery, vol. 10, no. 2, p. e1343, Mar. 2020, doi: https://doi.org/10.1002/widm.1343.

[13] Y. Wang *et al.*, "Density peak clustering algorithms: A review on the decade 2014–2023," *Expert Syst Appl*, vol. 238, p. 121860, 2024, doi: https://doi.org/10.1016/j.eswa.2023.121860.

[14] Z. Wang and Y. Wang, "A New Density Peak Clustering Algorithm for Automatically Determining Clustering Centers," in *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, 2020, pp. 128–134. doi: 10.1109/IWECAI50956.2020.00034.

[15] T. Fan, Z. Yao, L. Han, B. Liu, and L. Lv, "Density peaks clustering based on k-nearest neighbors sharing," *Concurr Comput*, vol. 33, no. 5, p. e5993, Mar. 2021, doi: https://doi.org/10.1002/cpe.5993.

[16] A. Lotf, P. Moradi, and H. Beigy, "Density Peaks Clustering Based on Density Backbone and Fuzzy Neighborhood," *Pattern Recognit*, vol. 107, pp. 1–30, 2020.

[17] L. Sun, T. Tao, X. Zheng, S. Bao, and Y. Luo, "Combining density peaks clustering and gravitational search method to enhance data clustering," *Eng Appl Artif Intell*, vol. 85, pp. 865–873, 2019, doi: https://doi.org/10.1016/j.engappai.2019.08.012.

[18] W. Tong, S. Liu, and X.-Z. Gao, "A density-peak-based clustering algorithm of automatically determining the number of clusters," *Neurocomputing*, vol. 458, pp. 655–666, 2021, doi: https://doi.org/10.1016/j.neucom.2020.03.125.

[19] J.-L. Lin, J.-C. Kuo, and H.-W. Chuang, "Improving Density Peak Clustering by Automatic Peak Selection and Single Linkage Clustering," *Symmetry (Basel)*, vol. 12, no. 7, 2020, doi: 10.3390/sym12071168.

[20] R. Siringoringo, R. P. Angin, and B. Rumahorbo, "MODEL KLASIFIKASI GENETIC-XGBOOST DENGAN T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING PADA PERAMALAN PASAR," *JTM*, vol. 11, no. 1, pp. 30–36, Aug. 2022.

[21] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data," *Jurnal Gaussian*, vol. 9, no. 3, pp. 227–236, Aug. 2020.

[22] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science (1979)*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[23] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics (Basel)*, vol. 9, no. 8, 2020, doi: 10.3390/electronics9081295.

[24] V. Robert, Y. Vasseur, and V. Brault, "Comparing High-Dimensional Partitions with the Co-clustering Adjusted Rand Index," *J Classif*, vol. 38, no. 1, pp. 158–186, 2021, doi: 10.1007/s00357-020-09379-w.

[25] A. Mahmoudi and D. Jemielniak, "Proof of biased behavior of Normalized Mutual Information," *Sci Rep*, vol. 14, no. 1, p. 9021, 2024, doi: 10.1038/s41598-024-59073-9.