JURNAL MEDIA APLIKOM
ISSN : 2086 - 972X
Vol.9 No. 2 Desember 2017

ANALISIS ALGORITMA CLUSTERING

DALAM KASUS PENENTUAN JENIS BUNGA IRIS

Diwahana Mutiara Candrasari Hermanto

Program Studi Sistem Informasi , STIKOM Yos Sudarso

Jl. SMP 5 Karangklesem Purwokerto

Email : candrasari897@gmail.com

Abstract : *Clustering* is one process of data mining that aims to partition existing data into one or more cluster objects based on the characteristics it has. Data with the same characteristics are grouped in one cluster and data with different characteristics are grouped into another cluster. In this study will perform comparation and analyze the best algorithm for categorize flowers by using iris dataset. Clustering algorithm techniques used include *K-means, and K-medoids*,. The value of davies bouldin and number of clusters will be investigated using the rapidminer tool. The results show that the *K-Means* algorithm has the lowest davies bouldin value of 0.167, while *K-Medoids* yields davies bouldin value of 0.291, but among the three algorithms, the *K-Means* algorithm is the most dominant and best in the comparative process of grouping iris flowers.

Keywords: *K-means, K-medoids* , clustering

Abstrak: *Clustering* adalah salah satu proses dari data mining yang bertujuan untuk mempartisi data yang ada kedalam satu atau lebih cluster objek berdasarkan karakteristik yang dimilikinya. Data dengan karakteristik yang sama dikelompokkan dalam satu cluster dan data dengan karakteristik berbeda dikelompokkan kedalam cluster yang lain. Pada penelitian ini akan melakukan komparasi dan menganalisa algoritma yang paling baik untuk mengelompokkan jenis bunga dengan menggunakan dataset iris. Teknik algoritma clustering yang digunakan antara lain *K-Means dan K-Medoids*, .Nilai davies bouldin dan number of cluster akan diteliti menggunakan tool rapidminer. Hasil menunjukkan bahwa algoritma *K-Means* memiliki nilai davies bouldin paling rendah yaitu sebesar 0.167,sedangkan *K-Medoids* menghasilkan nilai davies bouldin sebesar 0.291, tetapi diantara ketiga algoritma tersebut, algoritma *K-Means* merupakan algoritma paling dominan dan paling baik dalam proses komparasi pengelompokkan bunga iris.

Keywords: *K-Means, K-Medoids*, clustering

I. PENDAHULUAN

Kemajuan teknologi informasi sudah semakin berkembang pesat disegala bidang kehidupan. Banyak sekali data yang dihasilkan oleh teknologi informasi yang canggih, mulai dari bidang industri, ekonomi, ilmu dan teknologi serta berbagai bidang kehidupan lainnya. Pada proses penentuan jenis bunga pada umumnya menggunakan beberapa faktor.

Dalam penelitian ini akan dilakukan proses penentuan jenis bunga iris, dalam penentuan bunga iris tersebut digunakan 4 faktor yang merupakan karakteristik sendiri dari bunga iris tersebut. Keempat faktor karakteristik bunga iris yang digunakan untuk penelitan yaitu *sepal length, sepal width, petal length, dan petal width*. Sehingga dengan adanya banyak faktor yang diteliti dan record data yang ada dalam penentuan jenis bunga iris , maka diperlukan sebuah metode yang dapat digunakan untuk menghasilkan secara tepat dan akurat dalam menentukan jenis bunga pada umumnya, termasuk penentuan jenis bunga iris pada khususnya.

Sudah banyak algoritma clustering yang dipakai dalam namun ada beberapa algoritma clustering yang populer digunakan dalam analisis komparasi seperti algoritma *K-Means, K-Medoids, C-Means, Hard C-Means , dan X*-Means. Namun, banyak penelitian yang hanya masih menggunakan satu algoritma clustering dalam penentuan jenis bunga iris. Dalam penelitian ini dilakukan perbandingan terhadap 2 algoritma clustering yaitu algoritma *K-Means,*dan *K-Medoids,* dengan menggunakan metode validasi euclidience divergen untuk training dan testing dataset, serta metode perbandingan uji beda parametrik t-test untuk membandingkan nilai davies bouldin dan number of cluster dari algoritma clustering, sehingga diperoleh algoritma yang memiliki sifat dominan dan yang memiliki nilai davie bouldin paling rendah dari kedua algoritma tersebut merupakan algoritma yang terbaik dalam penentuan jenis bunga iris .

Makalah ini terdiri dari bagian 1 pendahuluan yang berisi tentang latar belakang masalah dan tujuan penelitian. Bagian 2 tinjauan studi berisi tentang penelitian yang terkait dengan penelitian yang dilakukan saat ini. Di bagian 3, berisi tinjauan pustaka yang digunakan dalam penelitian. Bagian 4 berisi metode penelitian yang digunakan, hasil penelitian dijabarkan pada bagian 5. Dan pada bagian terakhir, ditarik kesimpulan dari hasil penelitian.

II. TINJAUAN STUDI

Analisis mengenai komparasi atau perbandingan algoritma clustering sebelumnya telah dilakukan oleh beberapa peneliti, antara lain Leela,V (V. Leela, K. Sakthi priya and R. Manikandan 2014) yang melakukan perbandingan algoritma clustering untuk penentuan bunga iris. Algoritma clustering yang digunakan dalam penelitiannya yaitu *Fuzzy C-Means,* dan Hard C-Means , dari hasil penelitian

komparasi untuk penentuan jenis bunga iris menyatakan bahwa *Fuzzy C-Means* merupakan algoritma yang paling baik dalam menentukan jenis bunga iris.

Penelitian lainnya dilakukan oleh Milatul Ulya (Ulya,2011) yang melakukan penelitian modifikasi *K-Means* berbasis *Ordered Weighted Averaging (OWA)* untuk kasus clustering. Penelitian ini dilakukan menggunakan sebuah algoritma clustering yaitu *K-Means*,dan *K-Medoids* yang dimodifikasi dengan OWA.Hasil yang didapatkan dari komparasi penelitian tersebut menyatakan bahwa dengan menggunakan algoritma *K-Means* yang dimodifikasi OWA mendapatkan hasil accurancy yang lebih tinggi yaitu 96,67% dibandingkan dengan hanya menggunakan *K-Means* yang mendapatkan nilai accurancy 89,33%. Bhaskara Srinivas [3] melakukan penelitian untuk mengefisiensi data menggunakan algoritma clustering pada dataset iris.

Dalam penelitian ini digunakan algoritma Genetic dan juga algoritma PAM (K-Medoids), penelitian ini bertujuan untuk menunjukkan optimalisasi proses clustering dan pengelompokkan dataset iris dalam cluster menggunakan dataset iris. Hasil penelitian ini menunjukkan apabila menggunakan genetik algoritma dihasilkan pengelompokkan data pada cluster 1 sebesar 132 item dan pengelompokkan data pada cluster 2 sebesar 18 items ,sedangkan apabila menggunakan PAM hasil yang didapatkan dalam pengelompokkan data *cluster* 1 sebesar 97 items dan pada cluster 2 sebesar 53 item.

III. TINJAUAN PUSTAKA

1. Data Mining

Data mining merupakan proses untuk menemukan pola (pattern) dari suatu data. Pola (pattern) yang ditemukan harus memiliki arti atau mengandung informasi penting [2].

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [1].

2. Pengelompokan Data Mining

Menurut Larose, data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat di lakukan, yaitu:

a. Deskripsi

Terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data.

b. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori.

c. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

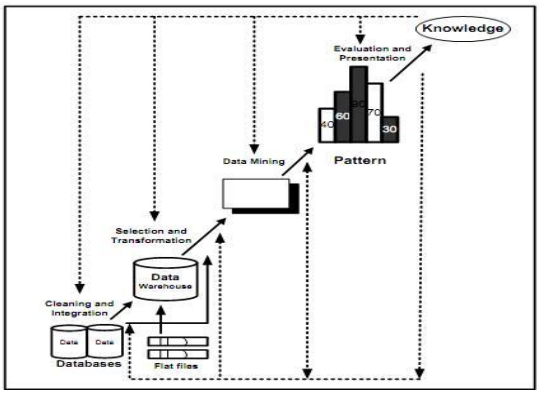
d. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Algoritma klasifikasi antara lain decision tree, naïve bayes, k-nearest neighbor, logistic linear, neural network, dan masih banyak lagi tipe algoritma klasifikasi yang sering digunakan.

e. Pengklusteran

Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (similarity) antara satu data dengan data yang lain. Asosiasi

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam suatu waktu. Untuk lebih jelas dalam proses data mining dapat melihat gambar 1 di bawah ini



Gambar 1. Proses Data Mining

3. Metode Clustering

Proses pengelompokan sekumpulan obyek kedalam kelas-kelas obyek yang sama disebut clustering pengelompokan [1]. Pengklasteran merupakan satu dari sekian banyak fungsi proses data mining untuk menemukan kelompok atau identifikasi kelompok obyek yang hampir sama. Analisis kluster *(Clustering)* merupakan usaha untuk mengidentifikasi kelompok obyek yang mirip-mirip dan membantu menemukan pola penyebaran dan pola hubungan dalam sekumpulan data yang besar. Hal penting dalam proses pengklasteran adalah menyatakan sekumpulan pola ke kelompok yang sesuai yang berguna untuk menemukan kesamaan dan perbedaan sehingga dapat menghasilkan kesimpulan yang berharga. .

4. Kategori *Clustering*

Metode klastering yang umumnya digunakan *K-Means clustering* yang termasuk metode partitioning clustering, yakni memilah-milah data ke dalam *cluster* yang ada.:

a. Fuzzy c-Means

b. X-Means

c. *K-Medoids*, dan

d. Kernel K-Means

5. Algoritma *K-Means*

Algoritma *K-Means* merupakan salah satu metode data klustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster /kelompok. Metode ini mempartisi ke dalam cluster / kelompok sehingga data yang memiliki karakteristik yang sama (High intra class similarity) dikelompokkan ke dalam satu cluster yang sama dan yang memiliki karakteristik yang berbeda (Law inter class similarity) dikelompokkan pada kelompok yang lain .[3]

Proses klustering dimulai dengan mengidentifikasi data yang akan dikluster, **Xij** (i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan dikluster dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap kluster ditetapkan secara bebas (sembarang), **Ckj** (k=1,...,k; j=1,...,m). Kemudian

dihitung jarak antara setiap data dengan setiap pusat kluster. Untuk melakukan penghitungan jarak data ke-i pada pusat kluster ke-k (ck), diberi nama (**dik**).

Menurut Hasn & Kamber (Larose,2005) algoritma *K-Means* bekerja dengan membagi data ke dalam k buah cluster yang telah ditentukan. Beberapa cara penghitungan jarak yang biasa digunakan yaitu:

a. Euclidean distance

Formula jarak antar dua titik dalam satu, dua dan tiga dimensi secara berurutan ditunjukkan pada formula 1, 2, 3 berikut ini :

$$\sqrt{(x-y)^2} = |x-y| \quad (2\text{-}1)$$

$$d(p,q) = \sqrt{(p1-q1)^2 + (p2-q2)^2} \quad (2\text{-}2)$$

$$d(p,q) = \sqrt{(p1-q1)^2 + (p2-q2)^2 + (p3-q3)^2} \quad (2\text{-}3)$$

b. Manhattan Distance

Manhattan distance disebut juga taxicab distance. (2-4)

$$d1(p,q) = ||p-q||1 = \sum_{i=1}^{n} |pi-q1| \quad (2\text{-}4)$$

6. Algoritma *K-Medoids*

Algoritma K-Medoids dikembangkan oleh Leonard Kaufman dan Peter J. Rousseeuw, dan algoritma ini sangat mirip dengan *K-Means*, karena keduanya algoritma partitional, terutama dengan kata lain, keduanya memecah dataset menjadi kelompok-kelompok, dan keduanya bekerja berusaha untuk meminimalkan kesalahan, tapi algoritma *K-Medoids* bekerja dengan medoids, yang merupakan entitas dari dataset yang mewakili kelompok di mana ia dimasukkan, dan *K-Means* bekerja dengan Sentroid, yang artifisial diciptakan entitas yang mewakili cluster. Algoritma *K-Medoids* atau dikenal pula dengan PAM (Partitioning Around Medoids) menggunakan metode partisi clustering untuk mengelompokkan sekumpulan n objek menjadi sejumlah k cluster. Algoritma ini menggunakan objek pada kumpulan objek untuk mewakili sebuah cluster. Objek yang terpilih untuk mewakili sebuah cluster disebut medoid. Cluster dibangun dengan menghitung kedekatan yang dimiliki antara medoid dengan objek non-medoids.

Algoritma *K-Medoids* adalah sebagai berikut [2].

1. Secara acak pilih *k* objek pada sekumpulan *n* objek sebagai *medoid*.
2. Ulangi:
3. Tempatkan objek *non-medoid* ke dalam *cluster* yang paling dekat dengan *medoid*.
4. Secara acak pilih oacak: sebuah objek *non-medoid*.
5. Hitung total biaya, *S*, dari pertukaran medoid oj dengan orandom.
6. Jika $S < 0$ maka tukar oj dengan oacak untuk membentuk sekumpulan *k* objek baru sebagai *medoid*.
7. Hingga tidak ada perubahan.

7. Pengujian Menggunakan Sample T-Test T-test

Dengan dua sampel digunakan untuk menentukan apakah rata-rata dua populasi adalah sama [1].

Jenis-jenis dari t-test antara lain :

a. One Sample T-Test

One sample t-test merupakan teknik analisis untuk membandingkan satu variabel bebas. Teknik ini digunakan untuk menguji apakah nilai tertentu berbeda secara signifikan atau tidak dengan rata-rata sebuah sampel.

b. Paired Sample T-Test

Analisis Paired sample t-test merupakan prosedur yang digunakan untuk membandingkan rata-rata dua variabel dalam satu grup. Artinya analisis ini berguna untuk melakukan pengujian terhadap satu sampel yang mendapatkan suatu treatment yang kemudian akan dibandingkan rata-rata dari sampel tersebut antara sebelum dan sesudah treatment.

c. Independent Sample T-Test Independent sample t-test adalah uji yang digunakan untuk menentukan apakah dua sampel yang tidak berhubungan memiliki rata-rata yang berbeda. Jadi tujuan metode ini aalah membandingkan rata-rata dua grup yang tidak memiliki hubungan satu sama lain.

IV. METODE PENELITIAN

1. Dataset

Dalam hal ini data set yang digunakan dalam proses analisi ini merupakan dataset yang bersifat public yang diperoleh dari uci yaitu dataset bunga iris. Pada Tabel 1 , akan dijelaskan mengenai struktur dataset yang digunakan .

Tabel 1. Struktur Dataset

No	Nama Atribut	Tipe Data	Ket
1	No_Id	Numeric	Attribute
2	Sepal Length (a1)	Numeric	Attribute
3	Sepal Width (a2)	Numeric	Attribute

Dalam Tabel 2. Akan dijelaskan mengenai record data yang digunakan dalam komparasi algoritma klasifikasi.

Tabel 2. Record Data Iris

No	ID	a1	A2	a3	a4
1	id_1	5,1	3,5	1,4	0,2
2	id_2	4,9	3,0	1,4	0,2

3	id_3	4,7	3,2	1,3	0,2
4	id_4	4,6	3,1	1,5	0,2
5	id_5	5,0	3,6	1,4	0,2

2. Pengolahan dan Validasi Dataset dengan Menggunakan Algoritma *Clustering*.

Dalam analisis ini, menggunakan tool *RapidMiner* yang digunakan untuk melakukan validasi dan pengujian dataset pada data mining . *RapidMiner* adalah suatu aplikasi open source yang digunakan untuk melakukan data mining. Perbedaan utama antara tool-tool komersial seperti Enterprise Miner, PASW, Statistika, dan tool-tool gratis seperti Weka dan RapidMiner, adalah efisiensi komputasionalnya. Dalam proses komparasi algoritma clustering dilakukan beberapa proses untuk mendapatkan hasil yang lebih akurat. Proses tersebut penginputan dataset, komparasi clustering, evaluasi, validasi, dan pembandingan.

V. HASIL DAN PEMBAHASAN

Hasil dari proses komparasi algoritma clustering yang digunakan yaitu algoritma *K-Means*, dan K-*Medoids* dengan menggunakan data set iris dibagi menjadi 2 yaitu hasil dari evaluasi masing-masing algoritma clustering dan juga hasil dari proses t-test yang dilakukan.

1. Proses Clustering Distance Performance

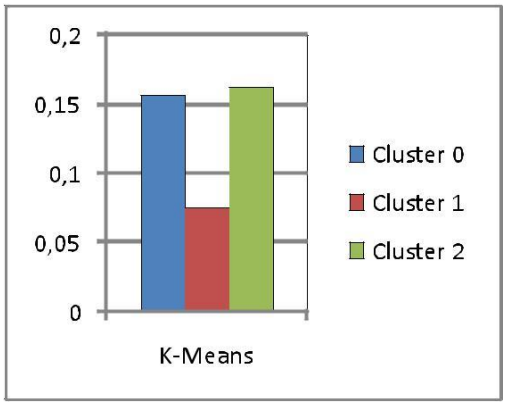
a. Algoritma *K-Means*

Dari hasil pengolahan dataset menggunakan algoritma clustering *K-Means* diperoleh hasil avg jarak antar centroid untuk cluster 0 sebesar 0,157, cluster 1 sebesar 0,076, dan cluster 2 sebesar 0,163 Untuk lebih jelasnya lagi dapat memperhatikan tabel 3.

Tabel 3. Hasil dari AVG Within Centroid Distance k-Means

Algoritma	Cluster 0	Cluster 1	Cluster 2
K-MEANS	0.157	0.076	0.163

Sehingga apabila digambarkan dalam grafik mengenai hasil AVG Within Distance algoritma *K-Means* pada gambar 2.



Gambar 2. Grafik Nilai AVG Within Centroid Distance k-Means

Selain mendapatkan hasil mengenai jarak rata-rata antar centroid proses cluster distance performa juga mendapatk\an nilai elegan values untuk masing-masing attribute. Pada tabel 4 dan 5 akan dijelaskan mengenai hasil yang didapatkan dengan menggunakan algoritma *K-Means* dalam proses cluster distance performa.

b. Algoritma *K-Medoids*

Dari hasil pengolahan dataset menggunakan algoritma clustering *K-Medoids* diperoleh hasil avg jarak antar centroid untuk cluster 1 sebesar 0,33, dan cluster 2 sebesar 0,867 Untuk lebih jelasnya lagi dapat memperhatikan tabel 6.

Tabel 6. Hasil dari AVG Within Centroid Distance k-Medoids

Algoritma	Cluster 1	Cluster 2
k-Medoids	0.33	0.867

3. Proses Clustering Map On Label

Dari hasil proses cluster map on label dari masing-masing algoritma clustering yang digunakan untuk komparasi dapat disimpulkan dalam tabel 7.

Tabel 7. Hasik Rekapitulasi Komparasi Algoritma Clustering

Indikator	K-means	K-medoids
Example Distribution	0.998	0.999

4. Hasil Proses Komparasi Algoritma

Clustering dengan Menggunakan T-Test. Dalam komparasi algoritma clustering menggunakan T-Test yaitu Paired T-Test. Dalam hal ini proses pembandingan T-Test dilakukan pada dua proses komparasi yaitu T-Test untuk proses cluster distance performance dan T-Test untuk proses cluster count

performance. Setelah melalui proses pemandingan dengan T-Test diperoleh hasil pembandingan dalam proses cluster distance performance dapat dilihat pada tabel 8.

Tabel 8. Hasil Proses T-Test Cluster Distance Performance

Algoritma		K-Means	K-Medoids
	Hasil	0.167	0.291
K-Means	0.167		1.000
K-Medoids	0.291		

VI. PENUTUP

Penelitian dengan membandingkan dua algoritma clustering yaitu *K-Means* dan *K-Medoids*, untuk menentukan jenis bunga iris dengan membandingkan nilai Davies Bouldin dan nilai dari number of cluster yang dihasilkan masing-masing algoritma komparasi tersebut serta dilakukan uji beda terhadap akurasi masing-masing algoritma dengan uji beda parametrik t-test menunjukkan bahwa algoritma *K-Medoids* memiliki nilai *Davies Bouldin* sebesar 0.291 tetapi nilai number of cluster paling sedikit 2.000 , sedangkan sifatnya tidak dominan dengan lainnya. Sedangkan algoritma *K-Means* memiliki nilai *Davies Bouldin* paling rendah yaitu 0.167 , tetapi memiliki sifat paling dominan diantara ke dua algoritma lainnya. Berikutnya ada algoritma *K-Medoids*. Sehingga dalam hal ini algoritma yang dapat digunakan untuk menentukan jenis dari bunga iris sendiri dapat menggunakan algoritma *K-Means* untuk mempermudah dalam proses clustering.

VII. DAFTAR RUJUKAN

[1] Han, Jiawei; & Kamber, Micheline. (2001). *Data Mining Concepts and Technique Second Edition*. San Francisco: Morgan Kauffman.

[2] Leela,V, Sakthi, P.K, dan Manikandan, R .(2014), Comparative Study Of Clustering Techniques Iris Datasets. World Applied Sciences Journal, 29 (Data Mining and Soft Computing Techniques): 24-29.

[3] Srinivas, Bhaskara A, Vardhan, Visnu B, dkk (2013). *An Efficient Data Clustering Algorithm Over Iris Dataset.* Journal Of Advanced Research In Computer Science and Softwre Engineering.