Evaluasi Hasil Klaster Pada Dataset Iris, Soybean-small, Wine Menggunakan Algoritma Fuzzy C-Means dan K-means++

Eko Budi Susanto*

Program Studi Teknik Informatika,STMIK Widya Pratama
Jl. Patriot 25 Pekalongan, Telp (0285)427816, www.stmik-wp.ac.id
email : eqo_bs@yahoo.com

Abstract

One technique in data mining is clustering. There are two types of clustering algorithms, namely soft and hard clustering clustering. K-Means++ is hard clustering algorithm which is an improvement of the algorithm K-Means. In K-Means++, center point selection is determined by the concept of probability do not choose at random like at the algorithm K-Means. Algorithm Fuzzy C-Means (FCM) is one of the improvements of the algorithm K-Means. FCM is soft clustering algorithm which applies fuzzy approach to determine the clusters based on the degree of membership. In this study will be evaluated on a cluster results of FCM algorithm and K-Means ++ at the dataset Iris, Wine and Soybean-Small. Results cluster of both algorithms will be compared and will look for the best. The test results of the cluster using the Confusion Matrix and Silhouette Coefficient. The result shows that the algorithm FCM and K-Means have almost similar performance. At the dataset Soybean-Small, Wine both algorithms have the same Silhoutte Coefficient, K-Means++ algorithm has an accuracy rate superior to the FCM algorithm.

Keywords: Data Mining; Fuzzy C-Means; K-Means++

1 PEDAHULUAN

Data Mining merupakan proses ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui dari suatu data. *Data Mining* bagian inti dari proses penggalian pengetahuan dalam *database(Knowledge Discovery in Database/KDD)*(Ian H. Witten, Frank Eibe 2011)(Maimon & Rokach 2010).*Data mining* mempunyai peranan yang sangat penting dalam pengolahan data untuk dijadikan sebuah informasi. Peran utama *data mining* diantaranya untuk estimasi, prediksi, klasifikasi, termasuk proses klasterisasi.

Pada proses klasterisasi terdapat dua macam teknik, yaitu *soft clustering* dan *hard clustering*(Nagamalai & Renault 2011). Pada teknik *soft clustering*, proses klasterisasi dilakukan berdasarkan nilai atau derajat keanggotaan. Beberapa algoritma yang termasuk ke dalam *soft clustering* antara lain Fuzzy Clustering, Fuzzy C-Means. Sedangkan pada teknik *hard clustering*, proses klasterisasi berdasarkan keanggotan bilangan *crisp* dan memungkinkan setiap titik data hanya menjadi milik satu *cluster* berdasarkan perhitungan jarak *Euclidean.* K-Medoids, K-Means termasuk dalam teknik *hard clustering*.

K-Means merupakan algoritma *clustering* yang banyak digunakan dalam teknik *clustering*(Bouveyron & Brunet-Saumard 2012)(Jain 2010), akan tetapi algoritma ini sangat sensitif terhadap penempatan awal pusat kluster(Celebi et al. 2013). Hal ini disebabkan pemilihan calon titik pusat dilakukan secara acak.Beberapa penelitian telah telah dilakukan untuk memperbaiki kekurangan algoritma K-Means, salah satunya Arthur, et al (2007) dalam penelitiannya mengusulkan aglortima K-Means++. K-Means++ diusulkan untuk memilih pusat klusterpada algoritma K-Means, bukan menghasilkan pusat kluster secara acak(Arthur & Vassilvitskii 2007).

Pada dasarnya algoritma Fuzzy C-Means merupakan perbaikan dari algoritma K-Means. Pada algoritma Fuzzy C-Means diterapkan pendekatan *fuzzy* untuk menentukan klaster berdasarkan derajat keanggotaan(Velmurugan 2012). Pada pendekatan fuzzy setiap elemen dari kumpulan data kemungkinan dimiliki oleh semua *cluster* tetapi dengan derajat keanggotaan yang berbeda. Fuzzy C-Means mampu menempatkan suatu data yang terletak diantara dua atau lebih *cluster* yang lain pada suatu cluster, sehingga data dapat menjadi anggota dari semua kelas atau *cluster* terbentuk dengan derajat atau tingkat keanggotan yang berbeda antara 0 dan 1.

Beberapa peneliti telah melakukan perbandingan kinerja antara algoritma *Hard Clustering* dan *Soft Clustering*, antara lain Shin dan Shon (2004) membandingkan kinerja algoritma K-

Means, SOM, dan Fuzzy C-Means, hasilnya, algoritma Fuzzy C-Means memiliki hasil segmentasi paling baik dari algoritma SOM dan K-Means(Shin & Sohn 2004). Danuta dan Jan (2005) membandingkan algoritma K-Means, DBSCAN, dan Two Step Clustering. Hasilnya algoritma K-Means memiliki kemampuan melakukan clustering dengan waktu yang tercepat (Zakrzewska & Murlewski 2005). Yohana (2011) membandingkan kinerja Fuzzy C-Means dengan Fuzzy Substractive Clustering, hasilnya Fuzzy C-Means memiliki tingkat validitas yang paling tinggi (Nugraheni 2011). Kemudian Asokan dan Mohanvalli (2011) juga membandingkan Fuzzy C-Means dengan K-Means, hasilnya Fuzzy C-Means memberikan hasil *clustering* yang lebih akurat dan efisien(Nagamalai & Renault 2011). Disis lain Kumar dan Wasan (2010) membandingkan varian dari algoritma K-Means, hasilnya K-Means++ memiliki tingkat konvergensi yang tinggi (Kumar 2010).

Berdasarkan penelitian terkait, algoritma clustering yang memiliki kinerja yang baik diantaranya K-Means++ dan Fuzzy C-Means. Untuk itu pada penelitian ini, kedua algoritma tersebut akan diterapkan untuk melakukan *clustering* pada dataset Wine, Soybean-Small, Iris dari UCI DATASET.Hasilnya akan dibandingkan dan ditentukan algoritma terbaik diantara keduanya. Proses pengujian dilakukan dengan metode *Silhoutee Coefficient*dan *Confusion Matrix*

2 KERANGKA TEORI

2.1 Data Mining.

Data mining merupakan inti dari proses *Knowledge Discovery in Database (*KDD*)*, yang melibatkan menyimpulkan algoritma yang mengeksplorasi data, mengembangkan model dan menemukan pola yang tidak diketahui sebelumnya. Model ini digunakan untuk memahami fenomena dari data, analisis dan prediksi. Aksebilitas dan banyaknya data membuat *Knowledge Discovery* dan *Data* Mining menjadi masalah yang cukup penting dan dibutuhkan(Maimon & Rokach 2010).

2.2 Clustering.

Clustering merupakan proses pembagian obyek ke dalam kelompok atau "*cluster*" sehingga obyek dalam suatu kelompok cenderung lebih mirip satu sama lain dibandingkan dengan obyek milik kelompok yang berbeda(Wu & Kumar 2009). *Clustering* juga disebut segmentasi data dalam beberapa aplikasi, karena membagi kelompok set data yang besar menjadi kelompok-kelompok yang memiliki kemiripan. Karakteristik tiap cluster tercermin pada setiap obyek yang kemiripan dalam cluster tersebut. Pada pendekatan *partitional clustering* terdapat istilah *Hard Clustering dan Fuzzy Clustering*. Pada dasarnya *cluster* dapat dilihat sebagai himpunan bagian dari himpunan data. Himpunan dapat berupa himpunan *fuzzy* atau *crisp (hard)*. Metode *hard clustering* berdasarkan teori himpunan klasik membagi *cluster* secara tegas yang mengharuskan obyek dapat menjadi anggota atau bukan anggota dari suatu *cluster*. Metode *Fuzzyclustering* membolehkan suatu objek untuk menjadi anggota dari beberapa *cluster* secara bersamaan, dengan derajat keanggotaan yang berbeda-beda. Derajat keanggotaan berada di antara rentang 0 dan 1. Di dalam situasi riil, *fuzzy clustering* memiliki hasil yang lebih natural dibandingkan dengan *hard clustering*(Jansen 2007).

2.3 *Fuzzy* C-Means

FCM pertama kali diperkenalkan oleh James Bezdek pada tahun 1981. Fuzzy C-Means adalah salah satu teknik pengelompokkan data yang mana keberadaan tiap titik data dalam suatu kelompok (cluster) ditentukan oleh derajat keanggotan. Metode Fuzzy C-Means termasuk metode *unsupervised clustering* dimana jumlah pusat cluster ditentukan di dalam proses clustering. Algoritma dari fuzzy c-means adalah sebagai berikut : (Wang et al. 2010)

1) Masukan input data, dan tentukan parameter yang terlibat.
   a. matriks ukuran n x m (n = jml sampel data, m = atribut data),
   b. jml cluster (c),
   c. bobot/tingkat kekaburan/fuzzy (m),
   d. MaxIterasi, akurasi (ε),
   e. fungsi objektif (P0 = 0),
   f. iterasi awal t= 1.
2) Bangkitkan bilangan random sebagai derajat keanggotaan awal (μ).
3) Hitung pusat cluster tiap cluster.

$$V_{kj} = \frac{\sum_{j=1}^{n}(\mu_{ik}{}^{m} \times x_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{m}} \quad (1)$$

4) Hitung fungsi objektif pada iterasi ke – t.

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2\right](\mu_{ik})^m\right) \quad (2)$$

5) Update derajat keanggotan μ.

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(x_{ij}-v_{kj})^2\right]^{\frac{-2}{m-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(x_{ij}-v_{kj})^2\right]^{\frac{-2}{m-1}}} \quad (3)$$

6) Cek kondisi berhenti, jika
   a. Jika $P_t - P_{t-1} < \varepsilon$, atau
   b. t < MaxIterasi,

7) Jika tidak, ulangi langkah 3

2.4 K-Means++

Pada algoritma K-Means centroid awal dipilih secara acak dari kumpulan data. Meskipun pendekatan ini sederhana dan cepat, akan tetapi terkadang mengasilkan hasil yang jauh dari optimal, karena tidak ada jaminan akurasi. Pada tahun 2007 Arthur dan Vassilvitskii (Arthur & Vassilvitskii 2007), pada K-Means++ centroid awal dipilih dengan probabilitas tertentu, probabilitas pemilihan titik sebagai centroid sebanding dengan jarak centroid terdekat yang sudah dipilih(Karch 2010). Berikut adalah algoritma K-Means++:

1) Pilih pusat awal c_1 seragam secara acak dari X. Hitung vektor yang berisi *square distance* antara semua titik dalam dataset dan c_1,

$$D_i^2 = \|x_i - c_1\|^2 \quad (4)$$

2) Pilih pusat kedua c_2 dari X secara acak dari distribusi probabilitas

$$\frac{D_i^2}{\sum_j D_j^2} \quad (5)$$

3) Hitung ulang jarak vektor

$$D_i^2 = min(\|x_i - c_1\|^2, \|x_i - c_2\|^2) \quad (6)$$

4) Pilih pusat c_l berturut-turut dan menghitung ulang jarak vektor

$$D_i^2 = min(\|x_i - c_1\|^2, \ldots., \|x_i - c_l\|^2) \quad (7)$$

5) Jika k center telah dipilih, langkah selanjutnya standar algoritma K-Means.

2.5 Evaluasi Hasil Clustering

2.5.1 Confusion Matrix

Confusion Matrix merupakan salah satu cara untuk melihat kinerja *classifier/supervised learning*(Bramer 2007), dalam *unsupervised learning* biasa dikenal dengan istilah *matching matrix*. Setiap kolom dari matriks mewakili kelas yang diprediksi, sedangkan setiap baris mewakili kelas yang sebenarnya.

Tabel2-1 Confusion Matrix

		Predicted	
		Yes	No
Actual	Yes	TP	FN
	No	FP	TN

Keterangan:
TP : True Positive (prediksi benar sesuai dengan data empiris)
TN : True Negative (prediksi benar tidak sesuai dengan data empiris)
FP : False Positive (prediksi salah tidak sesuai dengan data empiris)
FN : False Negative (prediksi salah sesuai dengan data empiris)

2.5.2 Silhouette Coefficient

Kualitas dari hasil klaster (anggota klaster) dapat diketahui melalui nilai*Silhoutte Coefficient* (Carlo Vercellis 2009), untuk menghitungnya dapat menggunakan rumus jarak *EuclideanDistance*. Berikut langkah untuk menghitung nilai *Silhoutte Coefficient:*

1) Untuk setiap titik i, hitung rata-rata jarak obyek i dengan seluruh titik yang berada dalam satu klaster. Maka akan didapatkan rata-rata jarak anatar titik dalam satu klaster (a_i)
2) Untuk setiap titik i, hitung rata-rata jarak titik i dengan seluruh titik yang berada di klaster yang lain. Dari semua jarak rata-rata tersebut ambil nilai yang tekecil (b_i)
3) Setelah itu titik i akan memiliki nilai *Silhoutte Coefficient:*

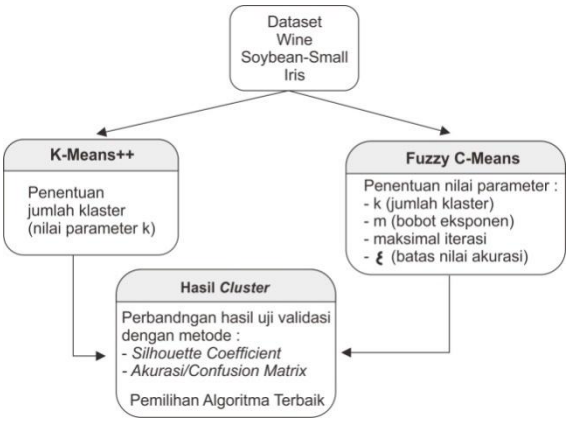
$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (8)$$

Hasil perhitungan *Silhoutte Coefficient* memiliki range antara -1 hingga 1. Dikatakan baik apabila bernilai positif, hal ini berarti titik sudah berada di klaster yang tepat. Sedangkan jika nilainya negatif ini menandakan terjadinya *overlapping* sehingga titik berada di klaster yang tidak tepat. Jika nilainya 0, ini berarti berada di antara dua klaster. Berikut ini adalah representasi dari nilai *Silhoutte Coefficient*:

Tabel 2-2 Representasi Kauffman dan Rousseeuw (1990)

SC	Representasi
0.71 – 1.00	Baik
0.51 – 0.70	Sedang
0.26 – 0.50	Buruk
≤ 0.25	Berada di klaster lain

2.5.3 Kerangka Pemikiran



Gambar2-1 Kerangka Pemikiran

Kedua algoritma clustering tersebut baik K-Meanss++ dan Fuzzy C-Means dipengaruhi oleh jumlah klaster yang ditentukan terlebih dahulu. Pada Fuzzy C-Means selain jumlah klaster, penentuan bobot eksponen (m), maksimal iterasi juga berpengaruh terhadap kinerjanya.

Dataset Wine, Soybean-Small, dan Iris diperoleh dari UCI Dataset. Ketiga dataset tersebut akan dikelompokkan menggunakan kedua algoritma tersebut. Hasilnya akan dibandingkan dan dicari mana yang terbaik diantara keduanya. Metode yang digunakan untuk melakukan uji validasi clustering yaitu *Silhoutte Coefficient*, dan *Confusion Matrix*.

2.6 Pustaka Rujukan

2.6.1 Penelitian Asokan, Mohanvalli

Penelitian yang dilakukan oleh Asokan dan Mohanvalli (2011) yang berjudul *Fuzzy Clustering for Effective Customer Relationship Management in Telecom Industry*(Nagamalai & Renault 2011)berfokuspada membandingkan dua pendekatan utama yaiut *soft clustering (K-Means)* dan *hard clustering (Fuzzy C-Means)* pada dataset telekomunikasi untuk menentukan rasio pelanggan *churn* sabagai upaya untuk meningkatkan *Customer Relationship Management (CRM).* Hasil *clustering* dari algoritma K-Means kemudian dianalisis dan ditemukan bahwa ada beberapa hasil *clustering* yang mengalami *missclassifications* (kesalahan klasifikasi). Sedangkan pada FCM tingkat kesalahan klasifikasi dapat berkurang. Bila dibandingkan dengan K-Means, FCM memberikan hasil *clustering* yang lebih akurat dan efisien.

2.6.2 Penelitian Kumar, Wasan

Penelitian yang dilakukan oleh Kumar dan Wasan (2010) yang berjudul *Compartive Analysis of k-mean Based Algortihms*(Kumar 2010). membandingkan kinerja dari varian dari algoritma K-Means seperti Global K-Means, Efficient K-Means, K-Means++, dan X-Means pada leukemia cancer dataset dan colon cancer dataset. Pada penelitian tersebut membandingkan tingkat akurasi dan konvergen dari masing-masing algoritma. Hasil dari peneliitan tersebut menyatakan bahawa secara keseluruan tingkat konvergensi yang paling tinggi adalah K-Means++ dan Global K-Means dibandingkan dengan varian algoritma K-Means lainnya. Begitu juga dengan tingkat akurasinya, K-Means++ dan Global K-Means memiliki tingkat akurasi yang paling tinggi dibandingkan dengan varian algoritma K-Means yang lain.

3 METODE PENELITIAN

3.1 Pengumpulan Data

Data diperoleh dari UCI Dataset. Dataset yang digunakan yaitu Wine, Soybean-Small, dan Iris.

3.2 Eksperimen

Eksperimen pada tahap ini dilakukan simulasi segmentasi pelanggan dengan menggunakan *Software Matlab 2014b*, dengan memanfaatkan *functionsource code*algoritma K-Means++ dan Fuzzy C-Means yang sudah tersedia di Matlab 2014b. Hasil proses *clustering* dilakukan uji validasi hasil klaster dengan menggunakan metode *Silhouette Coefficient dan Confusion Matrix*. Eksperimen dilakukan untuk mengetahui perbandingan dari uji validasi dari kedua algoritma tersebut untuk mendapatkan algoritma terbaik.

3.3 Uji Validasi

Pada tahap ini akan dilakukan perbandingan uji validasi hasil klaster pada algoritma K-Means++ dan FCM, dengan menggunakan metode *Silhouette Coefficient*untuk mengetahui kualitas dari hasil klaster yang dihasilkan. Metode *Confusion Matrix*untuk mengetahui tingkat akurasi.

4 HASIL DAN PEMBAHASAN

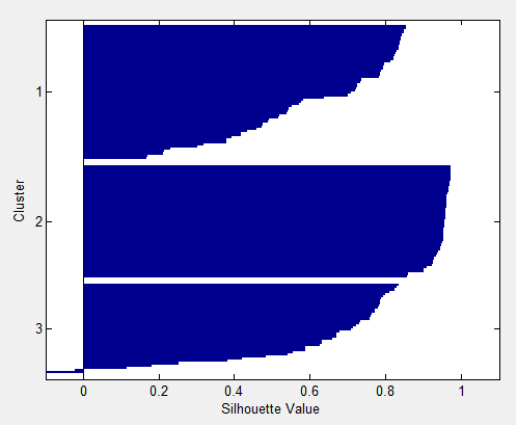
4.1.1 Pengujian pada dataset Iris

Dataset Iris *terdiri* 4 variabel dan 3 klaster yaitu setosa sebanyak 50 *records*, versicolor sebanyak 50 *records*, virginica sebanyak 50 *records*. Pada tabel 4-1 merupakan hasil pengujian akurasi dengan *confusion matrix* pada algoritma FCM

Tabel4-1 Hasil Confusion Matrix Algoritma FCM pada Dataset Iris

		aktual		
	cluster	setosa	versicolor	virginica
prediksi	setosa	50	0	0
	versicolor	0	47	13
	virginica	0	3	37

Dari tabel tersebut diatas dapat disimpulkan bahwa tingkat akurasi algoritma FCM pada dataset iris yaitu 89%. Sedangkan nilai *Silhouette Coefficient* dari hasil klaster tersebut yaitu 0.731. Berikut adalah grafik dari *Silhouette Coefficient* pada matlab 2014b.

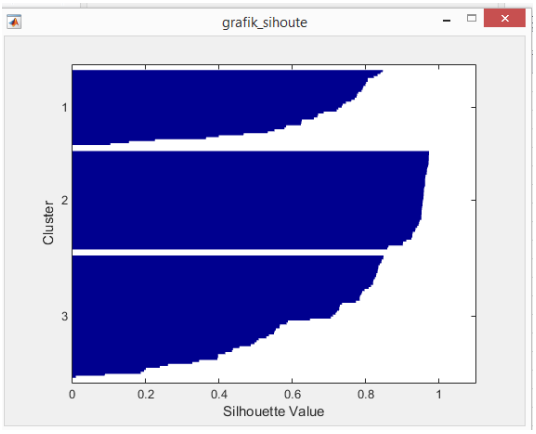


Gambar4-1 Grafik Silhoutte Coefficient Algoritm FCM pada data iris

Sedangkan hasil pengujian klaster pada algoritma K-Means++ dengan metode *confusion matrix* mempunyai tingkat akurasi 89.33%. Hasilnya dapat dilihat pada tabel 4-2. Sedangkan pada gambar 4-2 menunjukkan grafik dari *Silhouette Coefficient*. Nilai *Silhouette Coefficient* yang dihasilkan adalah 0.735.

Tabel4-2 Hasil Confusion Matrix Algoritma K-Means++ pada Dataset Iris

		aktual		
	cluster	setosa	versicolor	virginica
prediksi	setosa	50	0	0
	versicolor	0	48	14
	virginica	0	2	36



Gambar4-2 Grafik Silhoutte Coefficient Algoritm K-Means++ pada Dataset Iris

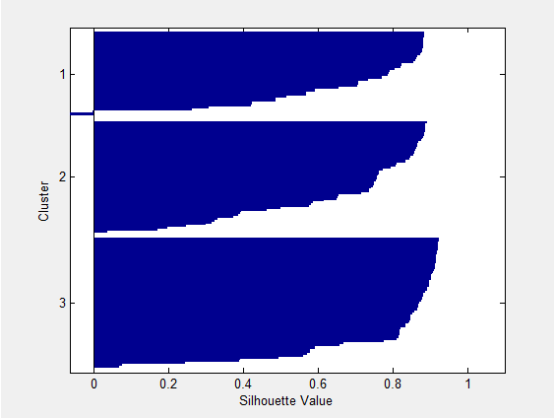
4.1.2 Pengujian pada dataset Wine

Pada dataset wine terdiri dari 13 variabel, dan 3 klaster yaitu klaster pertama terdiri dari 59 *records*, klaster kedua terdiri dari 71 *records*, klaster ketiga terdiri dari 48 *records*. Pada algoritma FCM pengujian dengan *Confusion Matrix* dapat dilihat pada tabe 4-3, dengan tingkat akurasi adalah 68.54%.

Tabel4-3 Hasil Confusion Matrix Algoritma FCM pada Dataset Wine

		aktual		
	cluster	Satu	Dua	Tiga
prediksi	Satu	45	1	0
	Dua	0	50	21
	Tiga	14	20	27

Pada gambar 4-3 dapat dilihat hasil dari grafik *Silhouette Coefficien*, dan nilai *Silhouette Coefficient* yang dihasilkan adalah 0.728



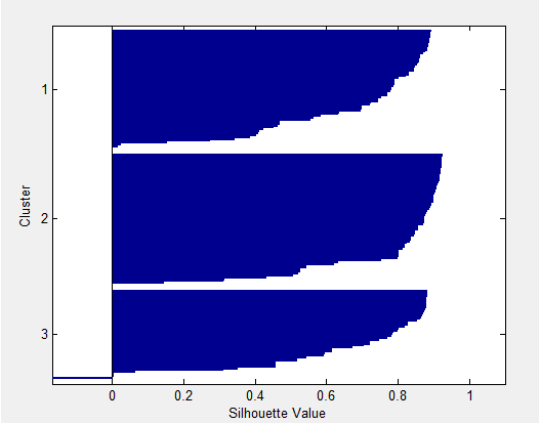
Gambar4-3 Grafik Silhoutte Coefficient Algoritm FCM pada Dataset Wine

Sedangkan hasil pengujian *Confusion Matrix* pada algoritma K-Means++ dapat dilihat pada tabel 4-4. Tingkat akurasi yang dihasilkan adalah 70.22%

Tabel4-4 Hasil Confusion Matrix Algoritma K-Means++ pada Dataset Wine

		aktual		
	cluster	Satu	Dua	Tiga
prediksi	Satu	46	1	0
	Dua	0	50	19
	Tiga	13	20	29

Dan nilai *Silhouette Coefficient* yang dihasilkan adalah 0.728. Pada gambar merupakan representasi dari *Silhouette Coefficient* pada algoritma K-Means++



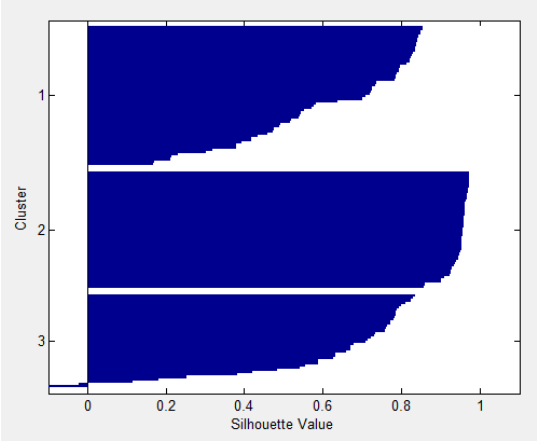
Gambar4-4 Grafik Silhoutte Coefficient Algoritm K-Means++ pada Dataset Wine

4.1.3 Pengujian pada dataset Soybean-Small

Dataset Soybean-Small terdiri dari 35 variabel dan 4 klaster, yaitu klaster D1 terdiri dari 10 *records*, klaster D2 terdiri dari 10 *records*, D3 terdiri dari 10 *record*, D4 terdiri dari 17 *records*. Hasil pengujian dengan *Confusion Matrix* dapat dilihat pada tabel 4-5, dengan tingkat akurasi adalah 70.21%.

Tabel4-5 Hasil Confusion Matrix Algoritma FCM pada Dataset Soybean-Small

		aktual			
	klaster	D1	D2	D3	D4
prediksi	D1	10	0	0	0
	D2	0	10	0	0
	D3	0	0	5	9
	D4	0	0	5	8



Gambar4-5 Grafik Silhoutte Coefficient Algoritm FCM pada Dataset Soybean-Small

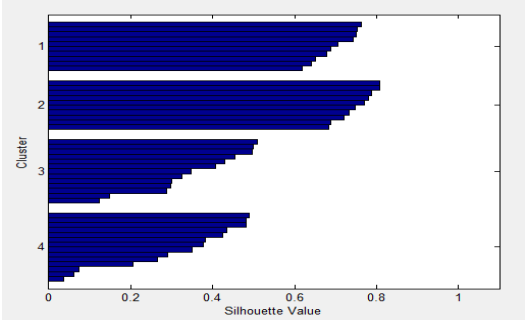
Gambar 4-5 menunjukkan grafik dari *Silhouette Coefficient*. Nilai *Silhouette Coefficient* yang dihasilkan adalah 0.499.

Sedangkan hasil pengujian *Confusion Matrix* pada algoritma K-Means++ dapat dilihat pada tabel 4-6. Tingkat akurasi yang dihasilkan adalah 72.34%.

Tabel4-6 Hasil Confusion Matrix Algoritma K-Means++ pada Dataset Soybean-Small

		aktual			
	klaster	D1	D2	D3	D4
prediksi	D1	10	0	0	0
	D2	0	10	0	0
	D3	0	0	5	8
	D4	0	0	5	9

Gambar 4-6 menunjukkan grafik dari *Silhouette Coefficient*. Nilai *Silhouette Coefficient* yang dihasilkan adalah 0.499



Gambar4-6 Grafik Silhoutte Coefficient Algoritma K-Means++ pada Dataset Soybean-Small

5 KESIMPULAN

Secara *keseluruhan* algoritma FCM dan K-Means++ hampir mempunyai performa yang sama. Pada dataset Iris keduanya mempunyai tingkat akurasi yang sama dan mempunyai nilai *Silhoutte Coefficient* yang hampir sama.

Pada *dataset* Soybean-Small dan Wine keduanya mempunyai nilai*Silhoutte Coefficient* yang sama. Akan tetapi algoritma K-Means++ memiliki tingkat akurasi yang lebih unggul dibandingkan algoritma FCM pada kedua dataset tersebut.

Dilihat dari kualitas hasil klaster yang dihasilkan berdasarkan nilai *Silhoutte Coefficient* rata-rata kedua algoritma tersebut mempunyai kualitas klaster yang baik. Hanya saja pada dataset Soybean-Small yang memiliki jumlah variabel paling banyak dibandingkan dengan dataset Wine dan Iris yaitu sebanyak 13, kualitas klaster yang dihasilkan dari kedua algoritma tersebut masuk ke dalam kategori buruk.

Tabel5-1 Hasil Pengujian Algoritma FCM dan K-Means++

DATASET		Algoritma	
		FCM	K-Means++
Iris	Akurasi	89.33%	89.33%
	Sillhoute	0.731	0.735
Wine	Akurasi	68.54%	70.22%
	Sillhoute	0.728	0.728
Soybean-Small	Akurasi	70.21%	72.34%
	Sillhoute	0.499	0.499

DaftarPustaka

Arthur, D. & Vassilvitskii, S., 2007. k-means ++ : The Advantages of Careful Seeding. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 8, pp.1–11.

Bouveyron, C. & Brunet-Saumard, C., 2012. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167947312004422 [Accessed March 9, 2013].

Bramer, M., 2007. *Principles of Data Mining*, London: Springer.

Carlo Vercellis, 2009. *Business Intelligence : Data Mining and Optimization for Decision Making* First Edit., Southern Gate: John Wiley & Sons, Ltd.

Celebi, M.E., Kingravi, H.A. & Vela, P.A., 2013. Expert Systems with Applications A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems With Applications*, 40(1), pp.200–210. Available at: http://dx.doi.org/10.1016/j.eswa.2012.07.021.

Ian H. Witten, Frank Eibe, M.A.H., 2011. *Data mining: Practical Machine Learning Tools and Techniques 3rd Edition* 3rd ed., United States of America: Elsevier.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp.651–666. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167865509002323 [Accessed March 19, 2014].

Jansen, S.M.H., 2007. *Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior Acknowledgments*. University of Maastricht (UM). Available at: https://dke.maastrichtuniversity.nl/westra/PhDMaBa-teaching/GraduationStudents/StephanJansen2007/Stephan_Jansen2007.pdf.

Karch, G., 2010. *GPU-based acceleration of selected clustering techniques*. Silesian University of Technology in Gliwice. Available at: http://www.visus.uni-stuttgart.de/uploads/tx_vispublications/Grzegorz_Karch_-_GPU-based_acceleration_of_selected_clustering_techniques.pdf.

Kumar, P., 2010. Comparative Analysis of k-mean Based Algorithms. *International Journal of Computer Science and Network Security*, 10(4), pp.314–318. Available at: http://paper.ijcsns.org/07_book/201004/20100447.pdf.

Maimon, O. & Rokach, L., 2010. *Data Mining and Knowledge Discovery Handbook* O. Maimon & L. Rokach, eds., Boston, MA: Springer US. Available at: http://www.springerlink.com/index/10.1007/978-0-387-09823-4 [Accessed May 22, 2013].

Nagamalai, D. & Renault, E., 2011. Trends in Computer Science ,. In *Trends in Computer Science, Engineering and Information Technology*. Springer Berlin Heidelberg.

Nugraheni, Y., 2011. *Data Mining dengan Metode Fuzzy untuk Customer Relationship Management (CRM) pada Perusahaan Retail*. Udayana Denpasar.

Shin, H.W. & Sohn, S.Y., 2004. Segmentation of stock trading customers according to potential value. *Expert Systems with Applications*, 27, pp.27–33.

Velmurugan, T., 2012. Evaluation of k-Medoids and Fuzzy C-Means clustering algorithms for clustering telecommunication data. *2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET)*, pp.115–120. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6513891.

Wang, H., Li, D. & Chu, Y., 2010. A New Scalability of Hybrid Fuzzy C-Means Algorithm. *2010 International Conference on Artificial Intelligence and Computational Intelligence*, (3), pp.55–58. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5655161 [Accessed April 3, 2013].

Wu, X. & Kumar, V., 2009. *The Top Ten Algorithms in Data Mining*, Minneapolis, Minnesota USA: Taylor & Francis Group, LLC.

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1578784.

Zakrzewska, D. & Murlewski, J., 2005. Clustering algorithms for bank customer segmentation. *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp.197–202. Available at: