Optimasi Algoritma Fuzzy Clustering dengan Menggunakan Algoritma *Forest Optimization*

Edhi Prabowo[1], Robert Kurniawan[2]

* Corresponding author : edhi.bowo@gmail.com
[1,2] Politeknik Statistika (STIS) Jakarta
Jl. Otto Iskandardinata No.64C 1 4, Bidara Cina, Jatinegara, Daerah Khusus Ibukota Jakarta 13330

Abstract-- Fuzzy C - Means (FCM) is one of the most commonly used clustering techniques, but has weaknesses that are sensitive to local optimum and sensitive to early cluster centers. Forest Optimization Algortihm is able to overcome the weakness of FCM. FOFCM is built with 2 types of distance ie Euclidean and Mahalanobis. FOFCM has better performance than FCM, since most iterations of FOFCM are less than FCM. The Mahalanobis FOFCM produces the least objective function value in hyperspherical data distribution compared to Euclidean FOFCM and FCM. Therefore, it can be concluded that Mahalanobis FOFCM is suitable for hyperspherical data distribution.

***Keywords**: FCM, FOFCM, Forest Optimization Algortihm*

Abstrak-- *Fuzzy C-Means (FCM) adalah salah satu teknik clustering yang sering digunakan, tetapi memiliki kelemahan yaitu sensitif terhadap local optima dan sensitif terhadap pusat cluster awal. Forest Optimization Algortihm mampu mengatasi kelemahan dari FCM. FOFCM dibangun dengan 2 jenis jarak yaitu Euclidean dan Mahalanobis. FOFCM memiliki performa yang lebih baik dari FCM, karena sebagian besar iterasi FOFCM lebih sedikit dari FCM. FOFCM Mahalanobis menghasilkan nilai fungsi objektif paling kecil pada sebaran data hyperspherical dibandingkan dengan FOFCM Euclidean maupun FCM. Oleh karena itu, dapat disimpulkan bahwa FOFCM Mahalanobis cocok untuk data hyperspherical.*

***Kata kunci** : FCM, FOFCM, Forest Optimization Algortihm*

PENDAHULUAN

Analisis clustering adalah teknik pembelajaran unsupervised yang bertujuan untuk menemukan pengelompokan objek secara alami berdasarkan kemiripan karakteristik objek[1]. Pembelajaran unsupervized adalah sebuah metode untuk menemukan pola pada suatu dataset yang belum memiliki informasi atau label. Dalam beberapa dekade terakhir, clustering memainkan peran penting dalam berbagai bidang sains dan engineering, seperti analisis data, pengenalan pola[2], pembelajaran mesin (machine learning)[3], segmentasi gambar[4], pendeteksian kesalahan (error)[5].

Pada umumnya, metode clustering terbagi ke dalam dua kategori, yaitu hard clustering dan fuzzy clustering. Pada metode hard clustering, setiap objek memiliki tingkat keanggotaan bernilai 0 atau 1 untuk setiap cluster yang ada. Apabila suatu objek memiliki tingkat keanggotaan bernilai 1 pada suatu cluster, tingkat keanggotaan objek tersebut bernilai 0 untuk *cluster-cluster* lain[6]. Sedangkan metode *fuzzy clustering* cocok untuk digunakan pada *fuzzy dataset*. *Fuzzy dataset* adalah kumpulan objek yang memiliki serangkaian nilai tingkat keanggotaan. Dataset ini digolongkan berdasarkan fungsi keanggotaan yang akan

memberikan nilai tingkat keanggotaan setiap objek dalam rentang antara 0 dan 1 [7]. Salah satu algoritme yang populer untuk *fuzzy clustering* adalah *Fuzzy C-means.*

Fuzzy C-means (FCM) memiliki kelebihan yaitu metode ini bersifat *unsupervized* dan dapat mencapai pusat *cluster* yang konvergen. Tetapi, dalam kondisi tertentu FCM merupakan model *clustering* yang mempunyai ketangguhan jika dilihat dari nilai fungsi obyektifnya, jumlah iterasinya dan waktu yang diselesaikan, bila dibandingkan dengan FCM yang ditambahkan dengan fungsi optimasi[11]. Selain memiliki kelebihan tersebut, FCM juga memiliki keterbatasan. Beberapa diantaranya adalah mudah terjebak dalam *local optima*, dan sensitif terhadap pusat *cluster* awal[6]. Untuk mengatasi keterbatasan-keterbatasan yang ada pada FCM, banyak data scientist yang mengusulkan dan menerapkan algoritme-algoritme optimasi. Salah satu algoritme optimasi yang diusulkan adalah algoritme Forest Optimization[8].

Forest Optimization Fuzzy C-Means (FOFCM) adalah sebuah metode Optimisasi FCM dengan algoritme Forest Optimization. Hasil penelitian FOFCM menunjukkan bahwa FOFCM lebih baik dibandingkan GGAFCM dan PSOFCM[6]. Pada penelitian tesebut, ukuran jarak yang digunakan adalah Euclidean.

Penelitian ini bertujuan untuk meningkatkan performa hasil *clustering* pada sebaran data h*yperspherical* dengan mengusulkan algoritme baru yaitu, algorime *Forest Optimization Fuzzy C-Means-Mahalanobis* (FOFCM-M). FOFCM-M merupakan algoritme FOFCM yang menggunakan ukuran jarak Mahalanobis dalam analisis *clustering*. Hasil *clustering* algoritme FOFCM-M tersebut akan dibandingkan dengan hasil pengelompokan algoritme *Fuzzy C-Means-Common Mahalanobis* (FCM-CM) yang merupakan algoritme FCM menggunakan ukuran jarak Mahalanobis.

Data yang digunakan dalam penelitian ini adalah data bangkitan *hyperspherical*. Hasil *clustering* terbaik dianalisis berdasarkan nilai fungsi objektif, dan iterasi.

Clustering

"Clustering is an unsupervised classification method when the only data available are unlabelled, and no structural information about it is available"[9]. *"Clustering is a mathematical tool that attempts to discover structures or certain patterns in a dataset, where the objects inside each cluster show a certain degree of similarity"* [10]. *"Clustering is a classification way for data analysis, which is utilized to classify a set of data or patterns commonly multidimensional into different groups according to a predefined measure, in order that items in the same group are more almost the same than those in different groups"* [6].

Berdasarkan pendapat-pendapat di atas maka dapat disimpulkan bahwa Clustering adalah suatu metode yang digunakan pada data yang tidak memiliki informasi label dan struktur untuk mengelompokan objek pada dataset ke dalam kelompok-kelompok/cluster-cluster berbeda, objek-objek yang berada di dalam kelompok yang sama lebih mirip satu sama lain dibandingkan dengan objek yang berada di dalam kelompok yang berbeda.

Fuzzy C-Means

Algoritme FCM pertama dikemukakan oleh Dunn (1973) dan disempurnakan oleh Bezdek (1981). Algoritme ini setara dengan algoritme k-means. Untuk menghasilkan struktur cluster terbaik dan hasil cluster yang optimal dapat dilakukan dengan meminimalkan fungsi objektif atau fungsi fitness. Fungsi objektif yang dipakai pada algoritma FCM didefinisikan sebagai berikut:

$$Jm = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^{m} d^2(y_j, z_i) \quad (1)$$

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad 1 \le i \le c; 1 \le j \le n \quad (2)$$

$$z_i = \frac{\sum_{j=1}^{n}(u_{ij})^{m} y_j}{\sum_{j=1}^{n}(u_{ij})^{m}} \quad (3)$$

Dengan Jm sebagai fungsi objektif FCM, y adalah matrik data dengan d dimensi, z adalah matriks pusat *cluster*, U adalah matriks partisi atau keanggotaan data, u_{ij} adalah nilai keanggotaan objek ke-j pada cluster ke-i; $u_{ij} \in [0,1]$, c adalah jumlah cluster, n adalah jumlah objek pada data,m adalah tingkat kekaburan atau fuzzifier, dan $d^2(y_j, z_i)$ adalah jarak antara objek y_j dengan pusat cluster z_i.

Forest Optimization Algorithm (FOA)

Forest Optimization algorithm adalah sebuah algoritma yang terinspirasi dari proses alami siklus

hidup pohon. Di dalam hutan, terdapat pohon-pohon yang hidup selama beberapa dekade, akan tetapi tidak sedikit pohon yang memiliki kehidupan yang singkat. Algoritme ini mensimulasikan persebaran benih yang jatuh di sekitar pohon maupun yang menyebar jauh dari induk, misal terbawa angin atau jatuh ke sungai.

Di dalam FOA terdapat 3 tahap utama, yaitu:

1. *Local seeding*
   Di alam, saat pembenihan terjadi, banyak benih yang jatuh disekitar pohon induknya dan tumbuh menjadi pohon-pohon muda. Kemudian pohon-pohon muda saling berkompetisi dan pohon yang memiliki lingkungan tumbuh yang lebih baik, seperti cahaya matahari yang cukup, akan menjadi pemenang kompetisi dan dapat bertahan hidup. FOA mensimulasikannya dengan menggandakan pohon-pohon yang berumur 0, kemudian mengubah satu nilai variabel secara acak pada pohon baru. Jumlah banyaknya pohon-pohon baru yang tercipta dari induk ditentukan dengan parameter *Local Seeding Change (LSC).*
2. *Population Limiting*
   Dengan terbatasnya tanah dan nutrisi di hutan, maka tidak sedikit pohon-pohon yang tidak dapat bertahan hidup. Keterbatasan ini disimulasikan dengan menghilangkan pohon-pohon yang tidak dapat bertahan hidup di dalam hutan. Dalam FOA terdapat parameter sebagai pembatas populasi di dalam hutan, yaitu *Life Time* (batas umur pohon) dan Area Limit (batas populasi pohon). Apabila pohon melampaui batas umur (*Age)* maka pohon tersebut akan dihilangkan dari populasi hutan. Selain itu, akibat adanya pembenihan maka populasi pohon di hutan akan selalu bertambah, maka pohon-pohon akan dihilangkan dari populasi apabila populasi di hutan lebih besar dari *Area Limit.* Pohon-pohon diurutkan berdasarkan nilai fungsi objektif (jm), kemudian pohon dengan jm terbesar dihilangkan dari populasi hutan. Pohon yang dihilangkan dari populasi hutan masuk ke dalam populasi kandidat.
3. *Global Seeding*
   Selain persebaran benih di sekitar induk, benih-benih juga menyebar ke berbagai penjuru hutan. Untuk melakukan simulasi ini, pohon-pohon di dalam kandidat populasi dipilih secara acak. Kemudian pohon-pohon terpilih diubah nilai-nilai variabelnya. Jumlah variabel yang diubah ditentukan oleh parameter *Global Seeding Change (GSC).*

Forest Optimization Fuzzy C-Means (FOFCM)

FOFCM adalah pengaplikasian FOA pada algoritme FCM. Berikut adalah tahapan analisis *clustering* data menggunakan FOFCM:

a. Menginput parameter
   a. Data yang akan dikelompokkan.
   b. Jumlah cluster (c)
   c. Nilai fuzzifier (w)
   d. Jumlah populasi pohon yang diinginkan (area limit)
   e. Nilai batas maksimal umur (life time)
   f. Nilai Local Seeding Changes (LSC)
   g. Nilai Global Seeding Changes (GSC)
   h. Nilai Transfer rate
   i. Nilai loop Gradient method
b. Membangkitkan pohon sejumlah populasi (area limit) secara acak dari data.

$X_{(1,1)}$	$X_{(1,...)}$	$X_{(1,D)}$	$X_{(2,1)}$	$X_{(2,...)}$	$X_{(2,D)}$	$X_{(.,...,.)}$	$X_{(C,D)}$	*Age*	*Jm*

Gambar 1. *Ilustrasi sebuah tree*

Dengan C adalah jumlah cluster, D adalah jumlah dimensi atau variabel data. Age adalah variabel yang menyimpan nilai umur pohon, pada inisialisasi awal bernilai 0. Jm adalah variabel yang menyimpan nilai objektif dari pohon tersebut.

c. Jalankan alogritme FOFCM:
   a. Lakukan *local seeding*
      Lakukan local seeding terhadap pohon-pohon yang berumur 0. Local seeding dilakukan dengan menciptakan pohon-pohon baru berdasarkan pohon yang berumur 0 dengan mengubah salah satu variabel. Variabel dipilih secara acak dan nilai variabel terpilih diubah menjadi x+r, x adalah nilai variabel sebelum diubah dan r adalah nilai acak yang berada dalam rentang $[-x_{min}, x_{min}]$, $-x_{min}$ adalah nilai minimum dari variabel terpilih. Parameter LSC digunakan untuk menentukan berapa banyak pohon baru yang dapat terbentuk dari 1 pohon lama. Pohon-pohon yang baru tercipta berumur 0, sedangkan umur pohon-pohon yang masuk ke tahap local seeding bertambah 1.

2. Hilangkan beberapa pohon dari popolasi utama
   a. Hilangkan pohon yang memiliki umur sama dengan life time.
   b. Hitung nilai Jm(dengan persamaan (1)) dari setiap pohon. Urutkan pohon berdasarkan nilai Jm, dimulai dari nilai Jm terkecil.
   c. Apabila pohon-pohon di populasi utama melebihi area limit, hilangkan pohon-pohon yang memiliki nilai Jm terbesar hingga jumlah pohon di populasi utama sama dengan are limit.
   d. Pohon-pohon yang dihilangkan di tahap ini akan disimpan ke dalam kandidat populasi.
3. Lakukan *global seeding*
   Pilih pohon sejumlah transfer rate dari kandidat populasi secara acak. Lakukan global seeding pada pohon-pohon terpilih. Global seeding dilakukan dengan mengubah nilai variabel pohon terpilih, variabel dipilih secara acak sebanyak GSC. Nilai variabel diubah menjadi r. Pohon-pohon tersebut kemudian dimasukan ke dalam populasi utama dengan umur 0.
4. Hilangkan beberapa pohon dari popolasi utama
   a. Hilangkan pohon yang memiliki umur sama dengan life time.
   b. Hitung nilai Jm dari setiap pohon. Urutkan pohon berdasarkan nilai Jm, dimulai dari nilai Jm terkecil.
   c. Apabila pohon-pohon di populasi utama melebihi area limit, hilangkan pohon-pohon yang memiliki nilai Jm terbesar hingga jumlah pohon di populasi utama sama dengan are limit.
5. Pilih pohon terbaik
   a. Pada iterasi pertama pohon terbaik dipilih dari pohon dengan nilai Jm terkecil.
   b. Pada iterasi selanjutnya pohon terbaik dibandingkan dengan pohon terbaik pada iterasi sebelumnya, dipilih pohon dengan nilai Jm yang lebih kecil.
6. Jalankan metode *Gradient*.
   a. Jalankan metode gradient pada pohon terbaik. Apabila output metode gradient lebih baik dibandingkan pohon terbaik berdasarkan nilai Jm, output metode gradient dijadikan pohon terbaik.
   b. Pohon terbaik dimasukan kembali ke dalam populasi utama dengan umur 0.
7. Hentikan proses apabila
   a. Pohon terbaik tidak berubah selama beberapa iterasi, atau
   b. Nilai akurasi tercapai, atau
   c. Iterasi maksimal tercapai
4. Pohon terbaik didapatkan. Pohon terbaik tanpa variabel age dan Jm dijadikan pusat cluster.
5. Jalankan algoritme FCM:
   b. Masukan pusat cluster yang didapat.
   c. Gunakan 1 iterasi.
6. Hasil *cluster* FOFCM didapatkan.

Distance (Jarak)

a. Euclidean

Euclidean distance atau *Ruler distance* atau *Pythagorean metric* merupakan ukuran jarak yang sering digunakan. Ukuran jarak ini juga merupakan ukuran jarak default dari FCM dan FOFCM.

Jarak Mahalanobis dapat dihitung menggunakan rumus yang dinotasikan sebagai berikut:

$$dist_{euclidean}(x,y) = \sqrt{\sum_{k=1}^{d}|x_k - y_k|^2} \quad (4)$$

Dengan (x, y) adalah koordinat objek dan *d* adalah dimensi objek.

b. Mahalanobis

Jarak Mahalanobis dapat dihitung menggunakan rumus yang dinotasikan sebagai berikut:

$$dist_{\text{Mahalanobis}}(x,y) = \sqrt{(x-y)A^{-1}(x-y)^T} \quad (5)$$

Dengan (x, y) adalah koordinat objek, *A* adalah kovarian dari matriks x dan y, dan *T* adalah transpose.

c. Common Mahalanobis

Jarak Mahalanobis dapat dihitung menggunakan rumus yang dinotasikan sebagai berikut:

$$dist_{\text{Common Mahalanobis}}(x,y) = \sqrt{(x-y)A^{-1}(x-y)^T - \ln|A^{-1}|} \quad (6)$$

Dengan (x, y) adalah koordinat objek, *A* adalah kovarian dari matriks x dan y, dan *T* adalah transpose.

METODOLOGI PENELITIAN

Data Bangkitan

Data bangkitan yang digunakan dalam penelitian ini diperoleh dengan software R runif{stat} untuk sebaran data *hypersphericall*. Jumlah data yang dibentuk adalah sebanyak 20 dan 200. Masing-masing jumlah data dibangkitkan menjadi 2 kelompok dengan jumlah variabel 2, 3, dan 4.

Metode analisis

a. Melakukan pengelompokan menggunakan FCM, FOFCM, FCM-CM dan FOFCM-M pada data.
b. Menghitung nilai fungsi objektif (jm) dan iterasi dari hasil pengelompokan metode FCM, FOFCM, FCM-CM dan FOFCM-M.
c. Membandingkan hasil pengelompokan FCM, FOFCM, FCM-CM dan FOFCM-M.

HASIL DAN PEMBAHASAN

Parameter yang digunakan

Berikut adalah parameter yang digunakan dalam penelitian ini:
Untuk FOFCM dan FOFCM-M:
Seed : 50
Jumlah *cluster* : [2;5]
Fuzzifier : [1,5;4]
Area Limit : 20
Life Time : 15
LSC : 20% dari jumlah variabel
GSC : 10% dari jumlah variabel
Transfer rate : 10%
Loop gradient method : 3
Nochange : 3

Untuk FCM dan FCM-CM:
Seed : 50
Jumlah *cluster* : [2;5]
Fuzzifier : [1,5;4]
Error : 1e-5
Loop Max: 1000

Perbandingan Pengelompokan FOFCM dan FCM pada Data

Dari input parameter-parameter di atas, dilakukan percobaan berulang kali kemudian didapatkan bahwa jumlah cluster dan fuzzifier terbaik untuk data 20 observasi adalah 5 cluster dengan fuzzifier 1,5, sedangkan untuk data 200 observasi adalah 2 cluster dan fuzzifier 1,5. Berikut adalah hasil percobaan yang telah dilakukan.

Table 1. *Perbandingan nilai fungsi objektif (jm)*

No	Jumlah observasi	Variabel	Jumlah Cluster (c)	Fuzifier (w)	**FCM**	**FOFCM**	**FCM-CM**	**FOFCM-M**
1	20	2	5	1.5	13.581811	13.805559	9.6203622	4.949602
2	20	3	5	1.5	24.470333	24.465888	19.222588	13.86114
3	20	4	5	1.5	29.913511	29.913511	17.565311	21.738221
4	200	2	2	1.5	393.533	393.533	399.5435	212.5477
5	200	3	2	1.5	573.2086	573.2086	585.1206	374.086
6	200	4	2	1.5	775.1077	775.1077	903.0477	533.0542

Dari tabel 1 diatas, dapat diketahui bahwa nilai jm dari FOFCM-M lebih baik pada data selain nomor 3 dibandingkan dengan metode FCM, FOFCM dan FCM-CM. Sedangkan pada data nomor 3 FOFCM-M lebih baik dibandingkan FCM dan FOFCM.
Selain nilai jm, dibandingkan pula iterasi yang diperlukan dalam menghasilkan pengelompokkan terbaik. Berikut adalah tabel yang menunjukkan jumlah iterasi yang diperlukan:

KESIMPULAN

Berdasarkan hasil eksperimen di atas diperoleh kesimpulan bahwa penanganan ketidakseimbangan kelas memberikan dampak

Table 2. *Perbandingan jumlah iterasi*

No.	Jumlah observasi	Variabel	Jumlah Cluster (c)	Fuzzifier (w)	**FCM**	**FOFCM**	**FCM-CM**	**FOFCM-M**
1	20	2	5	1.5	38	9	20	12
2	20	3	5	1.5	27	18	25	13
3	20	4	5	1.5	149	20	42	24

4	200	2	2	1.5	9	7	27	13
5	200	3	2	1.5	9	6	54	13
6	200	4	2	1.5	8	6	267	16

Dari tabel 2 diatas, dapat dilihat bahwa FOFCM dan FOFCM-M memiliki jumlah yang lebih sedikit dibandingkan dengan FCM dan FCM-CM. FOFCM-M memiliki jumlah iterasi yang lebih sedikit dibaningkan FCM-CM.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, algoritme FOFCM-M menghasilkan nilai jm lebih kecil dan lebih baik dibandingkan FCM, FOFCM, dan FCM-CM. Sedangkan dalam jumlah iterasi FOFCM lebih baik dari FOFCM-M, akan tetapi FOFCM-M lebih baik dari FCM dan FCM-CM. Dari hasil penelitian dapat disimpulkan bahwa FOFCM-M lebih baik dalam melakukan analisis clustering pada data bersebaran *hyperspherical*.

Untuk penelitian selanjutnya, disarankan untuk menggunakan ukuran jarak yang lain, misalnya seperti jarak Bray-Curtis dan Manhattan. Dengan penggunaan ukuran jarak yang lain diharapkan dapat meningkatkan performa dari FOFCM.

DAFTAR PUSTAKA

[1] A. José-García and W. Gómez-Flores, "Automatic clustering using nature-inspired metaheuristics: A survey," *Appl. Soft Comput. J.*, vol. 41, pp. 192–213, 2016.

[2] D. Biswas, A. Cranny, N. Gupta, K. Maharatna, J. Achner, J. Klemke, M. Jöbges, and S. Ortmann, "Recognizing upper limb movements with wrist worn inertial sensors using k-means clustering classification," *Hum. Mov. Sci.*, vol. 40, pp. 59–76, 2015.

[3] C. J. Lu and L. J. Kao, "A clustering-based sales forecasting scheme by using extreme learning machine and ensembling linkage methods with applications to computer server," *Eng. Appl. Artif. Intell.*, vol. 55, pp. 231–238, 2016.

[4] F. Zhao, J. Fan, and H. Liu, "Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local spatial information for image segmentation," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4083–4093, 2014.

[5]Z. Du, B. Fan, X. Jin, and J. Chi, *Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis*, vol. 73. Elsevier Ltd, 2014.

[6] A. Chaghari, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "Fuzzy clustering based on Forest optimization algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, 2016.

[7] L. a. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[8] M. Ghaemi and M. R. Feizi-Derakhshi, "Forest optimization algorithm," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6676–6687, 2014.

[9] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets Syst.*, vol. 158, no. 19, pp. 2095–2117, 2007.

[10] R. Suganya and R. Shanthi, "Fuzzy C-Means Algorithm-A Review," *Int. J. Sci. Res. Publ.*, vol. 2, no. 11, pp. 2250–3153, 2012.

[11] B. I. Nasution and R. Kurniawan, "Robustness of classical fuzzy C-means (FCM)," *2018 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, 2018, pp. 321-325. doi: 10.1109/ICOIACT.2018.8350729