PRISMA 3 (2020): 58-63
PRISMA, PROSIDING SEMINAR NASIONAL MATEMATIKA
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

Clustering Data dengan Algoritme *Fuzzy c-Means* Berbasis Indeks Validitas *Partition Coefficient and Exponential Separation* (PCAES)

Dewi Syifaur Rohmah[a,*], Dewi Retno Sari Saputro[b]

[a,b] *Program Studi Matematika FMIPA Universitas Sebelas Maret, Jl. Ir. Sutami 36A, Surakarta, Indonesia*

**Alamat Surel: dewisyifaur@student.uns.ac.id*

Abstrak

Clustering merupakan proses pengelompokkan data menjadi beberapa *cluster* atau kelompok. Hasil dari *clustering* diperoleh data dengan tingkat kemiripan yang tinggi dalam satu *cluster* dan berbeda dengan *cluster* lainnya. Pengelompokkan data tidak dilakukan secara manual melainkan dengan algoritme *clustering*. Salah satu algoritme tersebut adalah *fuzzy c-means* (FCM). FCM mengelompokkan data ke dalam suatu *cluster* berdasarkan derajat keanggotaan dari setiap data tersebut. FCM juga memiliki tingkat akurasi yang tinggi dan waktu komputasi yang cepat. Banyaknya *cluster* menjadi hal penting dalam proses *clustering*. Diperlukan suatu metode yang dapat digunakan untuk memeroleh *cluster* yang optimal sehingga hasil *clustering* dikatakan valid. Validitas tersebut dapat dilakukan dengan menentukan indeks validitas yang mempunyai nilai maksimum pada uji validitas. PCAES merupakan salah satu indeks validitas *cluster* yang menggabungkan dua faktor yaitu koefisien partisi yang dinormalisasi dan ukuran pemisahan eksponensial untuk setiap *cluster*. Pada penelitian ini dibahas teori *clustering* dengan metode FCM berbasis indeks validitas PCAES sebagai penentu banyak kelompok dalam proses *clustering*.

Kata kunci:

clustering, *fuzzy c-means*, indeks validitas, PCAES.

1. Pendahuluan

Data mining merupakan proses pencarian pola-pola yang menarik dan tersembunyi (hidden pattern) dari suatu data yang berukuran besar yang tersimpan dalam suatu basis data, data warehouse, atau tempat penyimpanan data lainnya (Tan *et al*, 2016). Pada data mining, clustering menjadi salah satu proses yang penting untuk mengolah data. Clustering sering digunakan sebagai langkah awal dalam proses analisis data. Dalam proses clustering data dikelompokkan menjadi beberapa cluster. Output proses clustering diperoleh data yang mempunyai karakteristik sama dengan data lainnya dalam satu cluster dan berbeda dengan data pada cluster yang lain. Hasil clustering yang berupa cluster-cluster selanjutnya dapat digunakan sebagai input dalam suatu teknik pengolahan data lainnya. Proses clustering ini digunakan algoritme yang secara ekstensif tidak hanya untuk mengorganisasikan dan mengkategorikan data, namun juga untuk kompresi data dan konstruksi model. Melalui pencarian kesamaan data, data dengan karakteristik sama dapat direpresentasikan dengan lebih sedikit simbol.

Terdapat beberapa algoritme dalam clustering yang digunakan yakni yang berbasis metode statistik, berbasis fuzzy, berbasis neural network, dan metode lain untuk optimasi centroid atau lebar cluster. Algoritme clustering berbasis fuzzy merupakan generalisasi dari metode partitional cluster dengan memperbolehkan suatu individu diklasifikasi secara parsial ke dalam lebih dari satu cluster. Salah satu algoritme clustering berbasis fuzzy yang sering digunakan adalah fuzzy c-means (FCM). FCM menggunakan model pengelompokkan fuzzy sehingga data dapat menjadi anggota dari semua kelas atau

To cite this article:
Rohmah, D.S. & Saputro, D.R.S. (2020). Clustering Data dengan Algoritme Fuzzy c-Means Berbasis Indeks Validitas Partition Coefficient and Exponential Separation (PCAES). *PRISMA, Prosiding Seminar Nasional Matematika 3*, 58-63

cluster. Pada algoritme FCM keberadaan data dalam suatu cluster ditentukan dari derajat keanggotaan setiap data. Pada algoritme FCM keberadaan data dalam suatu cluster ditentukan dari derajat keanggotaan setiap data.

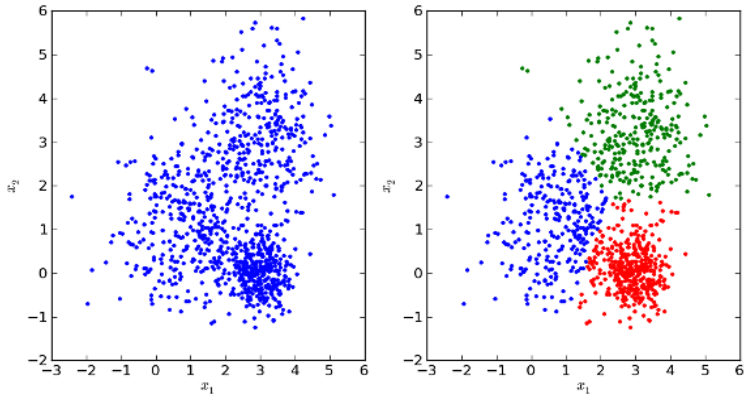
Pada proses clustering penentuan banyaknya cluster menjadi hal yang penting. Penentuan ini dapat berpengaruh pada tingkat validitas hasil clustering. Banyaknya cluster pada FCM ditentukan saat inisialisasi awal proses clustering. Guna memeroleh cluster yang optimum, dapat ditentukan indeks validitas yang maksimum pada uji validitas. Penelitian yang dilakukan oleh Wu dan Yang (2004) merekomendasikan indeks validitas untuk fuzzy clustering yaitu partition coefficient and exponential separation (PCAES). Pada artikel ini diuraikan teori clustering dengan FCM berbasis indeks validitas PCAES.

2. Pembahasan

2.1. Clustering

Clustering adalah proses pengelompokkan objek data ke dalam sejumlah *cluster* atau kelompok tertentu dengan kata lain *clustering* melakukan pemisahan atau segmentasi data ke dalam sejumlah *cluster* menurut karakteristik tertentu. *Clustering* banyak digunakan dalam berbagai bidang aplikasi seperti pada bidang teknik, ilmu komputer, medis, astronomi, sosial, dan ekonomi.

Menurut Kusumadewi, (2010) terdapat dua tahapan dalam melakukan *clustering* yaitu memutuskan apakah jumlah *cluster* digunakan atau tidak dan menentukan algoritme yang akan digunakan. Dalam melakukan analisis *clustering* terdapat dua pendekatan yang dapat digunakan yaitu *hard clustering* dan *soft clustering* (Gan *et al*, 2007). Hasil *clustering* yang baik akan menghasilkan data dengan tingkat kemiripan yang tinggi dalam satu *cluster* begitupun sebaliknya. Perhatikan Gambar 1, yang menunjukkan pencaran data sebelum di-*cluster* (a) dan sesudah di-*cluster* (b).



(a) (b)

Gambar 1. (a) pencaran data sebelum di-cluster (data asli) dan (b) hasil *cluster*

Sumber gambar : https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

2.2. Himpunan Fuzzy

Himpunan *fuzzy* (HF) merupakan hal penting dalam perkembangan konsep ketidakpastian, diperkenalkan oleh Lotfi A. Zadeh pada tahun 1965 (Klir & Yuan, 1995). Munculnya HF, mematahkan anggapan bahwa teori probabilitas sebagai satu-satunya alat untuk memecahkan masalah yang mengandung unsur ketidakpastian. Himpunan *fuzzy* didasarkan pada gagasan untuk memperluas jangkauan karakteristik suatu amatan sedemikian hingga karakteristik tersebut akan mencakup bilangan riil pada interval [0,1]. Keanggotaan himpunan *fuzzy* ditentukan oleh derajat keanggotaan yang menentukan tingkat kesesuaian setiap anggota dengan fungsi keanggotaan yang telah ditentukan dalam himpunan *fuzzy*. Terdapat beberapa fungsi keanggotaan *Fuzzy* yakni fungsi linear, fungsi segitiga, fungsi trapesium, fungsi Gauss, fungsi lonceng, fungsi Sigmoid dan sebagainya.

2.3. Fuzzy C-means

Fuzzy C-Means (FCM) merupakan salah satu metode *cluster* dengan memertimbangkan keberadaan data dalam suatu *cluster* ditentukan oleh keanggotaan yang mencakup himpunan *fuzzy*. Hal ini dapat membantu peneliti untuk mengetahui kesamaan dari setiap objek data pada satu kelompok atau *cluster*. FCM termasuk metode *supervised clustering* dengan jumlah *cluster* ditentukan dalam proses *cluster*. Algoritme yang biasa digunakan dalam *fuzzy clustering* adalah *fuzzy c-means* (FCM).

FCM adalah teknik *clustering* data dimana keberadaan tiap-tiap titik data dalam suatu *cluster* ditentukan oleh derajat keanggotaannya. Teknik ini pertama kali diperkenalkan oleh Jin Bezdek pada tahun 1981 (Kusumadewi, 2006). Konsep dasar FCM pertama kali adalah menentukan pusat *cluster* yang akan menandai lokasi rata-arata untuk tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Dengan cara memperbaiki nilai keanggotaan tiap-tiap data secara berulang, meminimisasi fungsi objektif, dan pusat *cluster* maka akan terlihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimisasi fungsi objektif yang menggambarkan jarak dari titik data yang diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik tersebut.

Fungsi objektif yang digunakan pada algoritme FCM adalah

$$J_w(U,V) = \sum_{i=1}^{n}\sum_{k=1}^{c}(\mu_{ik})^w(d_{ik})^2 \; ; \; U = \begin{bmatrix} \mu_{11}(x_1) & \dots & \mu_{1m}(x_k) \\ \vdots & & \vdots \\ \mu_{n1}(x_1) & \dots & \mu_{nm}(x_k) \end{bmatrix}$$

$$V = \begin{bmatrix} v_{11} & \dots & v_{1m} \\ \vdots & & \vdots \\ v_{n1} & \dots & v_{nm} \end{bmatrix} \; ; \; d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^{m}(x_{kj} - x_{ij})\right]^2$$

dengan

$J_w(U,V)$	: fungsi objektif terhadap U dan V
c	: jumlah *cluster*
n	: jumlah data yang digunakan
w	: pangkat pembobot, dengan $w \in [1,\infty)$
U	: matriks partisi awal
V	: matriks pusat *cluster*
μ_{ik}	: elemen-elemen dari matriks partisi U atau fungsi keanggotaan data ke-*k* (*k* = 1,2,3,...,n) pada *cluster* ke-*i* (*i* = 1,2,3,...,c)
d_{ik}	: fungsi jarak setiap data terhadap setiap pusat *cluster*

nilai J_w terkecil adalah yang terbaik sehingga $J_w{}^*(U^*,V^*) = \min J_w(U,V)$. Jika $d_{ik} > 0, \forall i,k; w > 1$ dan X setidaknya memiliki m elemen, maka $(u,v) \in M_{fm} > 0, \forall i,k; w$ dapat meminimasi J_w hanya jika

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}$$

dan

$$V_{kj} = \frac{\sum_{i=1}^{n}(\mu_{ik})^w x_{ij}}{\sum_{i=1}^{n}(\mu_{ik})^w}; 1 \le i \le m; 1 \le J \le m$$

algoritme FCM diuraikan sebagai berikut.

1. Menentukan data yang akan di-*cluster* $\boldsymbol{X}$ yang merupakan matriks berukuran $n \times p$ (*n:* jumlah sampel data, *p:* atribut setiap data). X_{ij} = data sampel ke-i (i : $1,2,\dots,n$), atribut ke-j ($j = 1,2,\dots,m$).
2. Menentukan (inisialisasi) nilai dari variabel berikut,
   a. jumlah *cluster* : c
   b. pangkat : *w*
   c. maksimum iterasi : MaxIter
   d. error terkecil yang diharapkan : ε

e. fungsi objektif awal : P_0 dengan $P_0 = 0$
f. iterasi awal : t dengan $t = 1$

3. Membangkitkan bilangan random sebagai elemen-elemen matriks partisi awal U

$$\boldsymbol{U} = \begin{bmatrix} \mu_{11}(x_1) & \dots & \mu_{1m}(x_k) \\ \vdots & \vdots & \vdots \\ \mu_{n1}(x_1) & \dots & \mu_{nm}(x_k) \end{bmatrix}$$

matriks partisi pada *fuzzy clustering* harus memenuhi kondisi sebagai berikut.

$$\mu_{ik} = [1,0]; (1 \le i \le c; 1 \le k \le n) ; \sum_{i=1}^{n} \mu_{ik} = 1; 1 \le i \le c$$

$$0 < \sum_{i=1}^{c} \mu_{ik} < c; 1 \le k \le n$$

4. Menghitung jumlah setiap kolom (atribut) dengan rumus yang ditulis sebagai

$$Q_j = \sum_{k=1}^{c} \mu_{ik}$$

dengan $j = 1,2,\dots,n.$

5. Melakukan normalisasi data dengan menghitung persamaan berikut

$$\mu_{ik} = \frac{\mu_{ik}}{Q_j}$$

6. Menghitung pusat *cluster* ke-k : V_{kj} , dengan $k = 1,2,\dots,c;$ dan $j = 1,2,\dots,m.$

$$V_{kj} = \frac{\sum_{i=1}^{n} (\mu_{ik})^w x_{ij}}{\sum_{i=1}^{n} (\mu_{ik})^w} \; ; \; \boldsymbol{V} = \begin{bmatrix} v_{11} & \dots & v_{1m} \\ \vdots & & \vdots \\ v_{n1} & \dots & v_{nm} \end{bmatrix}$$

7. Menghitung fungsi objektif pada iterasi ke-t dengan rumus

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

8. Menghitung perubahan matriks partisi dengan menggunakan rumus yang ditulis sebagai

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{p} \left(X_{ij} - V_{kj} \right)^2 \right]^{\frac{-1}{p-1}}}{\sum_{k=1}^{c} \left[\sum_{j=1}^{p} \left(X_{ij} - V_{kj} \right)^2 \right]^{\frac{-1}{p-1}}}$$

9. Mengecek kondisi berhenti yaitu apabila ($|P_t - P_{t-1}| < \varepsilon$) atau ($t <$ MaxIter) maka perhitungan berhenti. Sebaliknya, jika tidak demikian maka $t = t + 1$ dan mengulangi langkah ke-4.

2.4. Indeks Validitas

Algoritme *FCM* sering digunakan karena memiliki tingkat keakuratan yang tinggi dan waktu komputasi yang cepat (Sutoyo &Sumpala, 2015), namun kelemahan FCM adalah dalam penentuan jumlah *cluster* yang optimal. Jumlah *cluster* harus ditentukan terlebih dahulu pada inisialisasi awal sebelum dilakukan penelitian. Penentuan jumlah *cluster* yang tidak sesuai berdampak pada hasil *clustering* yang tidak optimal sehingga secara lebih luas berdampak pada saat pengambilan keputusan. Guna meminimalisir hal tersebut serta mengoptimalkan hasil *clustering*, perlu dilakukan validasi jumlah *cluster*. Validasi jumlah *cluster* dilakukan dengan menggunakan *cluster validity index* atau indeks validitas. Indeks validitas merupakan ukuran validitas untuk memeroleh jumlah *cluster* optimal yang sepenuhnya dapat menjelaskan struktur data (Zhao & Franti, 2014). Indeks validitas mengukur derajat kekompakan dan separasi struktur data pada

seluruh klaster dan menemukan jumlah klaster optimal yang kompak dan terpisah dari klaster yang lain (Wu & Yang, 2004).

2.5. Parttition Coefficient and Exponential Separation

Partition coefficient and exponential separation (PCAES) merupakan salah satu indeks validitas yang dapat digunakan untuk menentukan banyaknya *cluster* yang optimum. PCAES direkomendasikan oleh Wu dan Yang (2004), pada indeks PCAES ini dipertimbangkan dua faktor untuk memvalidasi setiap *cluster* yaitu koefisien partisi dan ukuran pemisahan eksponensial. Faktor koefisien partisi digunakan untuk mengukur ketidakselarasan suatu *cluster i* terhadap *cluster* yang paling selaras dengan ukuran kekompakkan μ_M. Ukuran pemisahan eksponensial telah dikaji oleh Wu dan Yang (2004) bahwa suatu jarak tipe eksponensial memberikan properti kuat berdasarkan analisis fungsi pengaruh. Selain itu ukuran pemisahan eksponensial juga digunakan untuk membuat ukuran pemisahan berada pada interval (0,1].

Indeks PCAES untuk *cluster i* ditulis sebagai

$$\text{PCAES}_\text{i} = \sum_{j=1}^{n} \frac{{\mu_{ij}}^2}{\mu_M} - \exp\left(-\min_{k\neq i}\left\{\frac{\|a_i - a_k\|^2}{\beta_T}\right\}\right) \quad (1)$$

dengan $\mu_M = \min_{1\leq i\leq c}\left\{\sum_{j=1}^{n} {\mu_{ij}}^2\right\}$ dan $\beta_T = \frac{\sum_{l=1}^{c}\|a_l - \bar{a}\|^2}{c}$, sementara istilah koefisien partisi yang dinormalisasi dituliskan sebagai

$$\sum_{j=1}^{n} \frac{{\mu_{ij}}^2}{\mu_M}$$

dengan $0 < \sum_{j=1}^{n} \frac{{\mu_{ij}}^2}{\mu_M} \leq 1$, dan ukuran pemisahan eksponensial dituliskan sebagai berikut

$$exp\left(-min_{k\neq i}\left[\frac{\|\alpha_i - \alpha_k\|^2}{\beta_T}\right]\right)$$

dengan $0 < exp\left(-min_{k\neq i}\left[\frac{\|\alpha_i - \alpha_k\|^2}{\beta_T}\right]\right) \leq 1$.

Pada persamaan (1) terlihat bahwa PCAES_i *cluster* terdeteksi menggunakan dua ukuran yaitu koefisien partisi dan ukuran pemisahan. Semakin maksimum nilai PCAES_i yang diperoleh berarti bahwa *cluster i* selaras dan terpisah dengan *cluster* lainnya. Jika nilai PCAES_i yang diperoleh minimum atau negatif maka menunjukkan bahwa *cluster i* tidak teridentifikasi dengan baik.

Menurut Wu dan Yang (2004), PCAES_i digunakan terlebih dahulu untuk mengukur keselarasan dan ukuran pemisahan untuk setiap *cluster*, selanjutnya menjumlahkan seluruh PCAES_i sebagai PCAES untuk mengukur keselarasan dan ukuran pemisahan struktur data. Sehingga total ukuran keselarasan dari data set diukur berdasarkan persamaan berikut

$$\sum_{i=1}^{c}\sum_{k=1}^{n} \frac{{u_{ik}}^2}{u_M}$$

yang merupakan koefisien partisi yang dinormalisasi. Total ukuran pemisahan dataset diukur berdasarkan persamaan berikut

$$\sum_{i=1}^{c} exp\left(-min_{k\neq i}\left[\frac{\|\alpha_i - \alpha_k\|^2}{\beta_T}\right]\right).$$

Indeks validitas PCAES ditulis sebagai berikut

$$\text{PCAES(c)} = \sum_{i=1}^{c}\sum_{k=1}^{n} \frac{{u_{ik}}^2}{u_M} - \sum_{i=1}^{c} exp\left(-min_{k\neq i}\left[\frac{\|\alpha_i - \alpha_k\|^2}{\beta_T}\right]\right)$$

Nilai PCAES yang maksimum menunjukkan bahwa setiap *cluster* dari *c cluster* selaras dan terpisah dari *cluster* lain sebaliknya, nilai PCAES yang minimum menunjukkan bahwa beberapa dari *c cluster* tidak selaras atau tidak terpisah dari *cluster* lainnya.

3. Simpulan

Berdasarkan pembahasan diperoleh simpulan bahwa pada algoritme FCM penentuan banyaknya cluster menjadi fokus utama permasalahan pada penelitian ini, untuk menyelesaikan masalah tersebut dipilih suatu indeks validitas yang dapat mengukur tingkat kevalidan suatu cluster. Indeks partition coefficient and exponential separation (PCAES) menjadi salah satu solusi untuk menentukan tingkat kevalidan dari suatu cluster. Semakin maksimum nilai PCAES menunjukkan bahwa setiap cluster dari c cluster selaras dan terpisah dari cluster lain, sebaliknya nilai PCAES yang minimum menunjukkan bahwa beberapa dari c cluster tidak selaras atau tidak terpisah dari cluster lainnya.

Daftar Pustaka

Gan, G., Chouqun, M., & Wu, J. (2007). *Data Clustering*. United State Of America : The America Statistic Association.

Klir, G.J. & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic, Theory and Applications.* Prentice Hall Inc., Upper Saddle River.

Kusumadewi & Purnomo. (2010). *Aplikasi Logika Fuzzy untuk Pendukung Keputusan*. Edisi 2. Yogyakarta : Penerbit Graha Ilmu.

Kusumadewi, S., Hartati S., Harjoko S., & Wrdoyo R. (2006). *Fuzzy Multi-attribute Decision Making*. Yogyakarta : Penerbit Graha Ilmu.

Sutoyo, M., S., & Sumpala, A., T., (2015). Penerapan Fuzzy C-Means untuk Deteksi Dini Kemampuan Penalaran Matematis. *Scientific Journal of Informatics*, 2(15), 129-136.

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

Wu, K-L. & Yang, M-S. (2004). A cluster validity Index for Fuzzy Clustering. *Pattern Recognition Letters*, 26(9), 1275-1291.

Zhao, Q., & Franti, P. (2014). WB-index : A Sum-of-Squares Based Index for Cluster *Validity, Data & Knowledge Engineering. Elsevier B.V.,* 92, 77–89