Prosiding Pendidikan Matematika dan Matematika Volume 2 (2020)
e-ISSN No. 2721-6802

# Meningkatkan Performa *Fuzzy Clustering* dengan *Principal Component Analysis*

Joko Eliyanto, Sugiyarto
Magister Pendidikan Matematika Universitas Ahmad Dahlan
Matematika Universitas Ahmad Dahlan
jokoeliyanto@gmail.com

***Abstrak*—** Big data adalah data yang memiliki volume yang besar, jenis data yang beragam serta produksi data yang sangat cepat. Berbagai metode matematika dan statistika dikembangkan untuk menganalisis big data ini. Jumlah data yang begitu besar dan kompleks menjadi salah satu tantangan dalam analisis big data. Untuk mengatasi hal ini salah satu teknik yang digunakan adalah reduksi dimensi. Reduksi dimensi adalah pengurangan dimensi suatu dataset dengan pertimbangan bahwa informasi-informasi penting tetap dipertahankan. Dengan dimensi yang lebih rendah hasil analisis pada data hasil reduksi masih menghasilkan kesimpulan yang relevan. Metode yang popular digunakan untuk reduksi dimensi adalah *prinsipal component analysis*. *Principal component analysis* adalah metode reduksi dimensi didasarkan pada komponen utama yang diperoleh dari kombinasi linear data dengan vektor eigen matriks kovarian terbaiknya. *Fuzzy clustering* adalah metode pengelompokan data dengan sistem tidak tertuntun yang mengaplikasikan teori fuzzy pada proses pengelompokannya. Salah satu metode *fuzzy clustering adalah fuzzy c-means clustering.* Penerapan *principal component analysis* pada *fuzzy c-means clustering* terbukti menurunkan nilai fungsi objektif hingga tersisa 0.092903%, waktu komputasi yang tersisa sebesar 63.09063%. Di saat yang sama, kualitas klaster yang dihasilkan masih bisa dipertahankan dengan nilai di atas 90%.

***Kata kunci: Big Data, Clustering, Fuzzy-C-Means, Principal Component Analysis***

## I. PENDAHULUAN

Saat ini, hampir semua orang tidak bisa terlepas dari jejak digital yang mereka buat di perangkat milik mereka masing-masing, artinya jumlah data yang dibuat akan sangat besar. Perkembangan teknologi tidak hanya menjadikan data yang tersedia begitu banyak, namun juga cepat. Data tersebut pun berbentuk dalam berbagai macam jenis mulai dari teks, gambar, video, jumlah klik pada sebuah halaman web dan lain-lain. Data yang memiliki volume besar, dihasilkan dengan cepat dan dalam bentuk bermacam-macam disebut sebagai big data [1]. Jika pada masa lalu manusia fokus pada penelitian data sampel yang kecil untuk memahami populasi data yang sesungguhnya, pada era ini manusia dituntut untuk menggali dan memahami informasi pada data yang tersedia pada populasi yang besar.

Data yang tersedia dalam berbagai jenis dapat direpresentasikan sebagai variabel. Dalam sudut pandang matematika, variabel juga dipandang sebagai dimensi. Manusia memiliki keterbatasan dalam memvisualisasikan data dalam dimensi tertentu. Selain itu, kompleksitas proses komputasi untuk menganalis data hingga diperoleh ssebuah informasi penting sangat bergantung pada jumlah data dan variabel pada sebuah dataset. Semakin banyak record data dan variabel data, maka semakin kompleks pula proses komputasi yang haru dilakukan. Reduksi dimensi adalah cara untuk menurukan dimensi pada sebuah data set dengan tetap mempertahankan informasi yang penting pada dataset tersebut [2]. Reduksi dimensi berperan untuk menurunkan beban dan biaya komputasi [3]. *Principal component analisis* adalah salah satu metode reduksi dimensi yang popular diterapkan oleh para ilmuan. Misalkan data yang akan direduksi terdiri dari tupel atau vektor data yang dijelaskan oleh n variabel atau dimensi. *Principal component analysis* mencari vektor ortogonal berdimensi m yang paling baik digunakan untuk mewakili data, di mana $m \leq n$ [4]. Data asli tersebut dengan demikian diproyeksikan ke ruang yang jauh lebih kecil, menghasilkan reduksi dimensi .

Data telah banyak tersedia saat ini, namun informasi berdasarkan data-data ini masih sedikit. Salah satu informasi yang penting dari sebuah data adalah pola kelompok data atau *cluster*. Dalam bidang ekonomi, pengelompokan pelanggan sangat bermanfaat untuk menentukan pengembangan sebuah produk atau peningkatan pemasaran suatu

Prosiding Pendidikan Matematika dan Matematika Volume 2 (2020)
e-ISSN No. 2721-6802

produk. Pada bidang kesehatan pengelompokan dapat dilakukan untuk melakukan penagangan yang tepat pada kelompok-kelompok orang yang terdampak suatu penyakit tertentu. Pada data pemrosesan gambar, pengelompokan dapat digunakan untuk menemukan batas-batas suatu objek [5]. Memasukan sebuah data ke dalam suatu kelompok sesungguhnya dapat dilakukan dengan dua acara. Yaitu memasukkan sebuah data ke sebuah kelompok dengan kriteria tertentu(misal jarak terpendek) atau dengan menghitung derajat keanggotaan sebuah data terhadap klister-klaster yang ada. Cara yang kedua ini disebut sebagai *fuzzy clustering*. Dengan metode *fuzzy* kemungkinan keanggotaan sebuah data terhadap suatu kelompok dapat dipertimbangkan. Dalam *fuzzy clustering*, metode yang popular digunakan adalah *fuzzy c-means*. Pada metode *fuzzy c-means clustering* data yang diubah ke dalam bentuk *fuzzy* adalah jarak antara objek dengan pusat klaster yang diberikan, fungsi objektif dari *fuzzy c-means clustering* adalah hasil kali dari derajat keanggotaan data pada suatu klaster dengan kuadrat dari jarak titik data ke pusat klaster. Dengan meminimumkan fungsi ini maka akan diperoleh klaster yang anggota-anggota di dalamnya sangat mirip dan perbedaan antar klaster tinggi [6].

Beberapa penelitian terkait *fuzzy clustering* terkhusus *fuzzy c-means clustering* telah dilakukan untuk meningkatkan kualitas hasil klastering. Abas Majdi dan Morteza Beiki menerapkan algoritma optimisasi evolusioner yaitu *genetic alghorithm* dan *particle swam optimization* untuk merancang dan mengoptimasikan kalsterisasi *fuzzy c-means clustering* untuk selanjutnya digunakan untuk memprediksi modulus dari deformasi massa batuan. Metode baru ini menghasilkan hasil klastering yang lebih akurat jika dibandingan dengan *fuzzy c-means clustering* yang sudah ada [7]. *Fuzzy c-means clustering* memiliki kelemahan utama yang dapat terjebak pada beberapa optimum lokal. Untuk mengatasi kekurangan ini, Adil Baykasoğlu dkk menggunakan algoritma metaheuristik generasi baru. Weighted Superposition Attraction Algorithm adalah metode baru berbasis kecerdasan segerombolan yang menarik inspirasi dari prinsip superposisi fisika dalam kombinasi dengan gerakan agen yang tertarik. Karena kemampuan konvergensi dan kepraktisannya yang tinggi, algoritma ini digunakan untuk meningkatkan kinerja *fuzzy c-means clustering*. Hasilnya menunjukkan peningkatan signifikan atas algoritma *fuzzy c-means clustering* yang biasa [8].

Reduksi dimensi pada data yang akan diklaster akan menurunkan beban komputasi pada proses klasterisasi. Pada makalah ini akan dibahas peningkatan performa *fuzzy c-means clustering* menggunakan metode reduksi dimensi *principal component analysis*. *Principal component analysis* sebagai alat visualisasi data *fuzzy* pada dimensi tinggi telah dikaji oleh Zhao Y dan kawan-kawan [9]. *Principal component analysis* yang diterapkan bersama dengan *rapid centroid estimation* terbukti dapat meningkatkan hasil *k-means clustering* [10]. *Principal component analysis* dan analisis klaster diterapkan secara bersamaan telah diterapkan dalam beberapa bidang [11][12][13]. Optimisasi *fuzzy clustering* dengan strategi reduksi dimensi *principal component analysis* telah diteliti oleh Vijayavarajan R dan Muttan [14]. Penerapan *principal component analysis* untuk reduksi dimensi pada *k-means clustering* telah diaplikasikan untuk melakukan prediksi kanker payudara [15].

Berdasarkan penelitian-penelitian di atas makalah ini bertujuan untuk membahas peningkatan performa dari *fuzzy c-means clustering* menggunakan *principal component analysis* dengan lebih rinci. Dengan metode reduksi dimensi ini diharapkan diperoleh *algoritma fuzzy c-means* yang lebih cepat dan ringan beban komputasinya, namun di saat yang sama hasilnya tetap akurat. Untuk melihat hal tersebut, analisis dilakukan pada fungsi objektif, waktu komputasi dan kuaitas hasil klasterisasi yang dihasilkan dari metode baru yang diajukan. Manfaat dari penelitian ini adalah untuk memahami *fuzzy clustering* khususnya *fuzzy c-means clustering* dan *principal component analysis* secara lebih mendalam. Selain itu, manfaat selanjutnya adalah ingin mengetahui seberapa jauh peran dari *principal component analysis* untuk meningkatkan performa dari *fuzzy c-means clustering*.

## II. METODE PENELITIAN

Metode penelitian yang digunakan pada makalah ini adalah reduksi dimensi *principal component analysis* untuk meningkatkan performa. *Fuzzy c-means clustering.* Setelah data diklasterisasi kemudian, kualitas klaster dianalisis dengan menghitung nilai akurasi dan *purity*nya.

### *A. Fuzzy Clustering*

Klastering atau disebut juga analisis klaster merupakan metode pengelompokan dengan pembelajaran tidak tertuntun. Data dikelompokkan tanpa memberikan kriteria tertentu sebelumnya. Pada akhirnya, data-data pada klaster yang sama akan memiliki tingkat kemiripan yang tinggi. Sebaliknya, data-data pada klaster yang berbeda akan memiliki tingkat kemiripan yang rendah. *Fuzzy clustering* adalah metode klasterisasi yang mengimplementasikan logika fuzzy pada proses pengklasteran. Masing-masing datum memiliki probabilitas dengan tingkat tertentu untuk tergabung dalam suatu klaster. Hal inilah yang memungkinkan analis klaster yang dilakukan menjadi semakin detail.

Prosiding Pendidikan Matematika dan Matematika Volume 2 (2020)
e-ISSN No. 2721-6802

Diberikan himpunan $A = \{a_1, a_2, a_3, \cdots, a_n\}$ , suatu himpunan *fuzzy* dari himpunan A adalah himpunan dari derajat keanggotaan setiap objek antara 0 dan 1. Himpunan fuzzy H secara formal dapat dimodelkan sebagai berikut: $F_H = A \rightarrow [0,1]$. Selanjutnya diberikan objek-objek $o_1, o_2, o_3, \cdots, o_n$ , *fuzzy clustering* untuk sejumlah k klaster *fuzzy* $C_1, C_2, C_3, \cdots, C_k$ dapat ditampilkan dalam bentuk matriks partisi $M = [u_{ij}]$ dengan $(1 \leq i \leq n, 1 \leq j \leq k)$, yang mana $u_{ij}$ adalah derajat keanggotaan objek $o_i$ pada klaster *fuzzy* $C_j$. Matriks ini harus memenuhi kriteria sebagai berikut:

1. Untuk setiap objek $o_i$ dan klaster $C_j$ , berlaku syarat $0 \leq u_{ij} \leq 1$.(Syarat klaster fuzzy)
2. Untuk setiap objek $o_i$ , $\sum_{j=1}^{k} u_{ij} = 1$.

*B. Fuzzy C-Means Clustering*

*Fuzzy C-means* adalah salah satu teknik dalam *fuzzy clustering*. Misalkan diberikan data $a = \{a_i\}_{i=1}^{n}$, dengan $n > k$ yang mana n adalah banyaknya data dan k adalah banyaknya klaster pada titik data yang berbeda dalam dimensi *d*. Fungsi objektif *Fuzzy C-Means* adalah meminimumkan (1) sebagai berikut:

$$F(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{k} {u_{ij}}^{m} d_{ij} \quad (1)$$

dengan kendala

$$\sum_{j}^{k} u_{ij} = 1,$$
$$0 \leq u_{ij} \leq 1,$$
$$1 \leq i \leq n, 1 \leq j \leq k$$

Keterangan :

n : banyaknya data

k : banyaknya klaster

m : indeks *fuzziness*

$u_{ij}$ : derajat keanggotaan pada data ke-i dan klaster ke j

$d_{ij}$ : fungsi jarak untuk mengukur kemiripan antar datum dalam dataset

*C. Algoritma Fuzzy C-Means*

Algoritma *Fuzzy C-Means* adalah tahap-tahap pengklasteran dengan Teknik *Fuzzy C-Means*. Untuk melakukan klasterisasi dengan teknik ini, dilakukan langkah-langkah sebagai berikut:

1. Mengambil data yang mengandung variabel random $X = \{x_1, x_2, x_3, \cdots, x_n\}$, $Y = \{y_1, y_2, y_3, \cdots, y_j\}$ dan yang menyatakan objek dan atribut. Data berupa matriks yang berukuran $n \times j$ , yang mana *n* banyak data dan *j* adalah banyaknya atribut data.
2. Menentukan jumlah klaster = *k*, iterasi maksimal = *MaxIter*, error terkecil yang diharapkan = $\varepsilon$ dengan iterasi awal *t=1* dan nilai awal fungsi objektif $F(U,V)^{(0)} = 0$.
3. Menentukan matriks partisi awal ${u_{ij}}^{(0)}$ sebarang berukuran $k \times n$ yang terdiri dari bilangan random $0 \leq u_{ij} \leq 1$, sedemikian sehingga jumlah keseluruhan matriks partisi baru dalam kelas adalah sesuai dengan (1).
$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ u_{k1} & u_{k1} & \cdots & u_{kn} \end{bmatrix} \quad (3)$$
yang mana $u_{11} + u_{21} + \cdots + u_{k1} = 1$.
4. Menghitung pusat klaster $v_{ij}$.
$$v_{ij} = \frac{\sum_{i=1}^{k} ({u_{ij}}^{m} \times x_{ij})}{(\sum_{i=1}^{k} {u_{ij}}^{m})} \quad (4)$$
5. Menghitung nilai fungsi objektif $F(U,V)$ menggunakan (1).
6. Menghitung perubahan matriks partisi.
7. Memeriksa kondisi untuk berhenti yaitu jika $|F(U,V)^{t} - F(U,V)^{t-1}| < \varepsilon$ atau $t = MaxIter$ . Jika belum dipenuhi maka perhitungan terus dilakukan hingga salah satu di antara keduanya dipenuhi.

Prosiding Pendidikan Matematika dan Matematika Volume 2 (2020)
e-ISSN No. 2721-6802

*D. Evaluasi Klaster*

Untuk mengetahui tingkat akurasi algoritma klasterisasi dan label kelas yang tersedia maka dilakukan evaluasi pengklasteran . Pada makalah ini untuk mengevaluasi pengklasteran dilakukan dua uji, yaitu *purity* dan akurasi.

1. *Purity*
   Purity digunakan untuk menghitung kemurnian dari suatu klaster. Untuk menghitung *purity* setiap klaster yang diperoleh, diambil paling banyak dari objek yang masuk dalam klaster C yang mana $1 < i < C$ dan $C'$ adalah kelas asli ke-h dengan $1 < h < C'$. Kemudian untuk *puriy* keseluruhan C klaster yaitu dengan menjumlahkan setiap *purity* pada klaster ke-C kemudian dibagi dengan banyaknya objek yang didefinisikan sebagi berikut:

$$purity(P, C) = \frac{1}{n}\sum_i^k max_{i \le h \le C'} |P_i \cap C_h| \quad (5)$$

   yang mana $P = \{P_1, P_2, \cdots, P_k\}$ adalah himpunan klaster dan $C = \{C_1, C_2, \dots, C_k\}$ adalah himpunan kelas asli. Pengklasteran yang buruk memiliki nilai *purity* mendekati 0. Hal tersebut bermakna tidak ada hasil klaster yang sesuai kelas asli, sedangkan klaster yang baik memiliki nilai *purity* 1 . Sehingga, dapat diartikan hasil klaster sesuai dengan kelas asli.
2. Akurasi
   Akurasi dihitung dengan menjumlahkan banyaknya objek yang masuk dalam klaster ke-i, dimana $1 \le i \le k$ yang tepat pada kelas aslinya kemudian dibagi dengan banyaknya objek data. Akurasi didefinisikan sebagai berikut :

$$r = \frac{\sum_{i=1}^k a_i}{n} \quad (6)$$

   Keterangan:
   $a_i$ : jumlah objek pada klaster ke-i yang sesuai dengan kelas asli.
   $n$ : jumlah n objek.
   Hasil akurasi yang baik jika semua klaster sesuai dengan kelas asli dan kemudian dibagi dengan banyaknya data akan menghasilkan nilai maksimal 1.

*E. Principal Component Analysis*

Principal component analysis adalah metode untuk ekstraksi variabel, artinya variabel-variabel yang 'tidak penting' akan dibuang sehingga dimensi dapat direduksi. Variabel-variabel yang dibuang adalah variabel-variabel yang berkorelasi sehingga hasil dari adalah variabel-variabel yang independen. Ini mengubah satu dataset variabel yang saling terkait menjadi yang tidak berkorelasi yang disebut komponen utama. Jumlah komponen utama lebih kecil dari jumlah variabel dataset awal.

Metode principal component analysis disebut juga metode reduksi dimensi. Sejumlah $n$ variabel data set akan direduksi menjadi $k$ variabel dengan $k < n$. Untuk mereduksi dimensi sebuah dataset pada metode ini dilakukan beberapa langkah berikut:

1. Menghitung matriks kovarian data.
2. Menghitung nilai eigen dan vektor eigen dari matriks kovarian.
3. Mengurutkan vektor eigen berdasarkan nilai eigen dari besar ke kecil.
4. Memilih sejumlah *k* nilai eigen.
5. Mengalikan dataset dengan vektor eigen yang bersesuaian dengan nilai eigen.
6. Diperoleh dataset baru dengan dimensi *k.*

## III. HASIL DAN PEMBAHASAN

*A. Algoritma Fuzzy C Means Clustering dengan Principal Component Analysis*

Untuk meningkatkan performa *fuzzy c-means clustering*, terlebih dahulu diterapkan metode reduksi dimensi *principal component analysis* pada dataset. Dengan dimensi yang lebih rendah, hal tersebut akan mengurangi beban dari algoritma *fuzzy c-means clustering*. Kemudian akan dilakukan uji *purity* dan akurasi untuk melihat kulitas hasil klasterisasi. Berikut adalah algoritma *fuzzy c-mans clustering* dengan *principal component analysis.*

1. Mengambil data yang mengandung variabel random $X = \{x_1, x_2, \cdots, x_n\}$ dan $Y = \{y_1, y_2, \cdots, y_j\}$ yang menyatakan objek dan atribut data. Data berupa matriks yang berukuran $n \times j$, yang mana n banyak data dan j adalah banyaknya atribut data.
2. Menerapkan metode *principal component analysis* sehingga jumlah variabel $Y = \{y_1, y_2, \cdots, y_j\}$ akan menjadi sejumlah variabel baru $Y = \{y_1, y_2, \cdots, y_p\}$, dengan $p < j$.
3. Menerapkan metode *fuzzy c-means clustering* sehingga diperoleh klaster-klaster data.
4. Menghitung nilai *purity* hasil klaster.
5. Menghitung akurasi hasil klaster.

*B. Simulasi Algoritma Fuzzy C-Means Clustering dengan Principal Component Analysis*

Algoritma *Fuzzy C Means dengan Principal Component Analysis* disimulasikan pada tiga dataset yang berbeda [16]. Pertama adalah dataset iris. Dataset ini berisi 4 variabel dengan 150 data masing-masing di dalamnya. Kelas asli dalam dataset berjumlah 3 kelas. Setiap kelas berisikan 50 data. Data iris terdiri dari 4 variabel atau atribut yaitu panjang sepal, lebar sepal, panjang petal dan lebar petal yang diukur dalam satuan dalam cm. Dataset yang kedua adalah dataset biji-hijian(*seeds)*. Dataset ini terdiri dari 7 variabel dan 210 baris data. Untuk memperoleh data ini, kelompok yang diperiksa terdiri dari biji-bijian milik tiga varietas gandum yang berbeda: Kama, Rosa dan Kanada, masing-masing 70 unsur, dipilih secara acak untuk percobaan. Visualisasi kualitas tinggi dari struktur kernel internal dideteksi menggunakan teknik sinar-X yang lembut. Ini tidak merusak dan jauh lebih murah daripada teknik pencitraan lain yang lebih canggih seperti pemindaian mikroskop atau teknologi laser. Gambar direkam pada piring KODAK X-ray 13x18 cm. Studi dilakukan dengan menggunakan gandum gandum dipanen yang berasal dari bidang eksperimental, dieksplorasi di Institute of Agrophysics dari Akademi Ilmu Pengetahuan Polandia di Lublin. Kemudian dataset yang terakhir adalah dataset *wine.* Data ini berisi 13 variabel dan 178 *record* data. Data ini adalah hasil analisis kimia dari anggur yang ditanam di wilayah yang sama di Italia tetapi berasal dari tiga kultivar yang berbeda. Algoritma *fuzzy c-means clustering* dan *fuzzy c-means clustering principal component analysis* disusun dalam m-file menggunakan perangkat lunak MATLAB 2019b dan kemudian dataset ini diolah menggunakan m-file tersebut. Ketiga dataset diklasterisasi dengan *fuzzy c-means clustering* dan *fuzzy c-means clustering principal component analysis*. Pada metode yang diajukan, dataset direduksi masing-masing menjadi dua variabel tersisa menggunakan *principal component analysis* kemudian diklasterisasi dengan *fuzzy c-means clustering*. Hasil klasterisasi dari ketiga dataset dengan metode *fuzzy c-means clustering* tanpa proses *principal component analysis* (FCM) ditampilkan dalam Tabel 1. Sedangkan hasil klasterisasi *fuzzy c-means clustering principal component analysis*(FCM PCA) ditampilkan pada Tabel 2.

TABEL 1. HASIL KLASTERISASI FCM

| Data Set | Fungsi Objektif | Waktu Komputasi(detik) | Purity | Akurasi |
|---|---|---|---|---|
| IRIS | 123.1805 | 6.090096 | 0.98 | 0.98 |
| SEEDS | 381.9179 | 9.211472 | 0.919 | 0.919 |
| WINE | 1.21E+04 | 7.497974 | 0.6854 | 0.6854 |

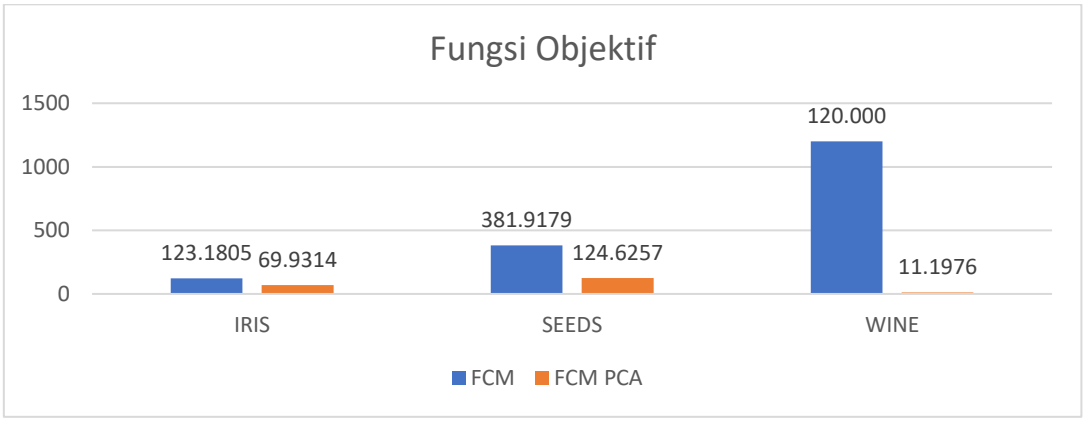
TABEL 2. HASIL KLASTERISASI FCM PCA

| Data Set | Fungsi Objektif | Waktu Komputasi(detik) | Purity | Akurasi |
|---|---|---|---|---|
| IRIS | 69.9314 | 3.84228 | 0.9067 | 0.9067 |
| SEEDS | 124.6257 | 8.599204 | 0.8619 | 0.8619 |
| WINE | 11.1976 | 7.409914 | 0.691 | 0.691 |

Prosiding Pendidikan Matematika dan Matematika Volume 2 (2020)
e-ISSN No. 2721-6802

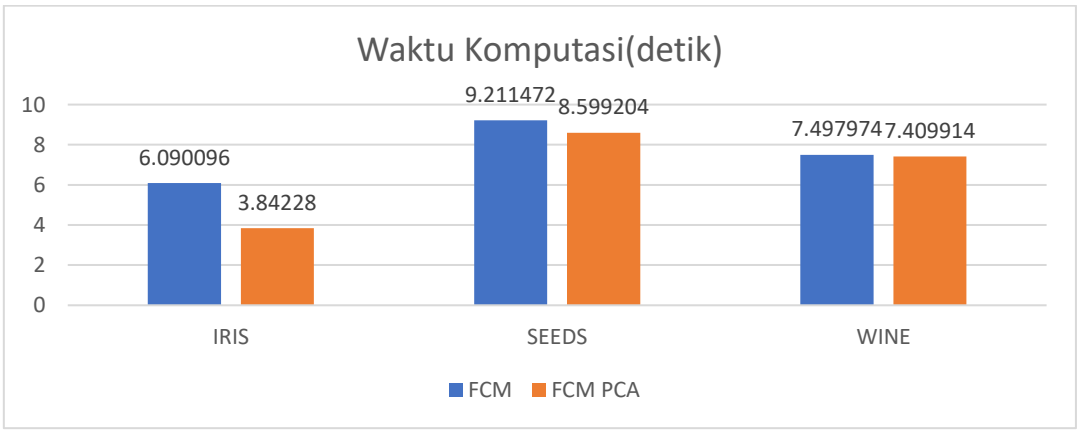
Berdasarkan Tabel 1 dan Tabel 2 maka akan diperbandingkan performa dari FCM PCA terhadap FCM biasa. Untuk melihat pengaruh PCA pada FCM maka dilihat beberapa indikator. Yang pertama adalah nilai fungsi objektifnya. Nilai fungsi objektif pada ketiga dataset baik dengan metode FCM dan FCM PCA ditunjukkan pada Gambar 1. Terlihat bahwa nilai fungsi objektif FCM PCA lebih rendah dari FCM. Tujuan awal dari FCM adalah meminimumkan fungsi objektif, sehingga semakin rendah nilai fungsi objektifnya maka semakin baik.

GAMBAR 1. FUNGSI OBJEKTIF FCM DAN FCM PCA




Salah satu fungsi dari reduksi dimensi adalah menurunkan waktu komputasi. Pada penelitian ini PCA dapat meningkakan performa dari FCM dengan menurunkan waktu komputasi jika dibandingkan dengan FCM biasa. Hal ini ditunjukkan pada Gambar 2.

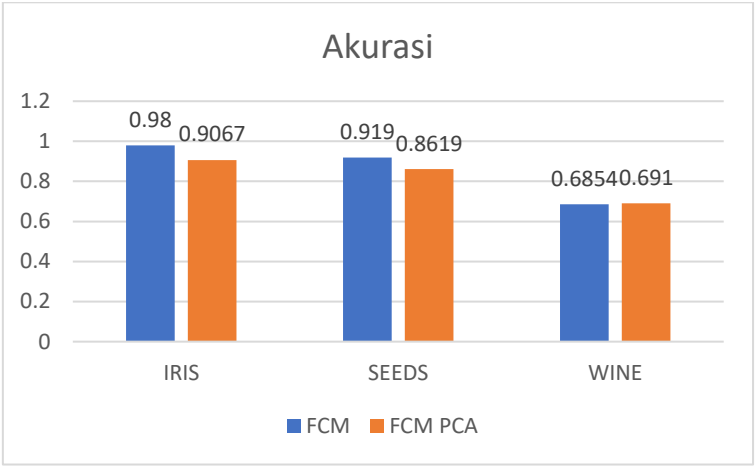
GAMBAR 2. FUNGSI OBJEKTIF FCM DAN FCM PCA




Kedua hasil di atas dapat merupakan dampak langsung dari reduksi dimensi PCA. Bagian yang menarik diamati lebih lanjut adalah hasil klasterisasi baik dari FCM dan FCM PCA. Kualitas hasil klasterisasi FCM dan FCM PCA dilihat dari dua aspek yaitu akurasi dan *purity*. Akurasi menunjukkan keakuratan metode klasterisasi untuk mengklaster dataset sesuai dengan klaster aslinya. Nilai akurasi FCM dan FCM PCA disajikan pada Gambar 3. Terlihat tingkat akurasi setelah penerapan PCA pada FCM tidak berubah signifikan. Keduanya memiliki tingkat

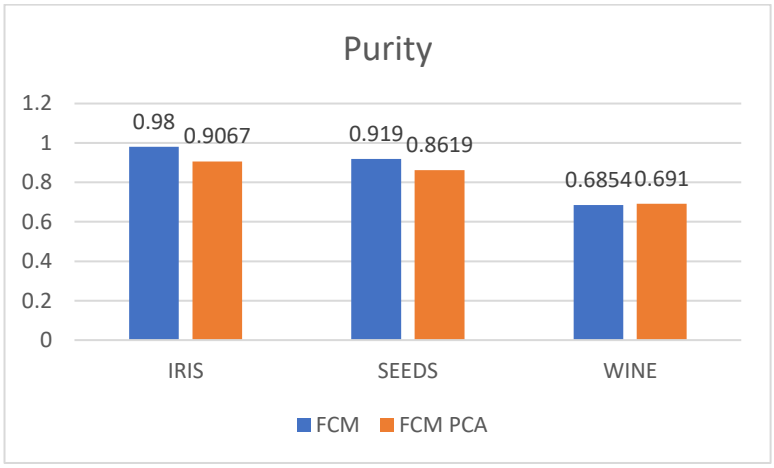
akurasi di atas 90%. Artinya FCM PCA masih mampu mempertahankan tingkat akurasi FCM di atas 90% untuk data set iris dan *seeds*. PCA yang dikombinasikan dengan FCM untuk mengkasterisasi dataset *wine* justru menunjukkan hasil yang berbeda. Tingkat akurasi tidak hanya dipertahankan namun justru dapat ditingkatkan meskipun relatif kecil yaitu dapat meningkatkan nilai akurasi hinga 0.056.

GAMBAR 3. AKURASI HASIL KLASTERISASI FCM DAN FCM PCA




Selain akurasi, tingkat kemurnian klaster atau *purity* juga menunjukkan kualitas hasil klasterisasi. Semakin tinggi tingkat *purity* hal ini berarti semakin mirip hasil klasterisasi dengan kelas asli. Mirip seperti nilai akurasi, nilai *purity* pada kedua metode ini relatif tidak berubah begitu jauh. Hal ini ditunjukkan pada Gambar 4. Untuk dataset iris dan *seeds* nilai *purity* dapat dipertahankan dan pada dataset *wine* dapat ditingkatkan dengan nilai yang sama yaitu 0.056.

GAMBAR 4. PURITY HASIL KLASTERISASI FCM DAN FCM PCA




Yang menarik adalah, FCM PCA tetap masih mempertahankan kualitas hasil klaster. Terbukti dengan nilai *purity* dan akurasi yang masih bisa dipertahankan hingga di atas 90%, bahkan pada dataset *wine*, kedua nilai evaluasi tersebut dapat ditingkatkan. Berdasarkan beberapa hal di atas dapat dikatakan PCA mampu meningkatkan performa FCM.

## IV. SIMPULAN DAN SARAN

*Principal component analysis* mampu meningkatkan performa dari *fuzzy c-means clustering* yang merupakan salah satu dari metode *fuzzy clustering*. Hal ini didasarkan pada penurunan nilai fungsi objektif dan penurunan waktu komputasi. Meskipun demikian, kualitas hasil klaster masih dapat dipertahankan. Untuk meningkatkan klasterisasi *fuzzy clustering* terdapat metode reduksi dimensi lain, seperti *multidimentional reduction*, *core and reduct*, *factor analysis* dan lain-lain.

Prosiding Pendidikan Matematika dan Matematika Volume 2 (2020)
e-ISSN No. 2721-6802

## DAFTAR PUSTAKA

[1] B. Marr, "Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results", John Wiley & Sons, 2006.
[2] K.P. Singh, A. Malik, ,D. Mohan, and S. Sinha, "Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study". Water research, vol.38, pp. 3980-3992, 2004.
[3] P. Paokanta, N. Harnpornchai, S. Srichairatanakool, and M. Ceccarelli, "The Knowledge Discovery of [beta]-Thalassemia Using Principal Components Analysis: PCA and Machine Learning Techniques", International Journal of e-Education, e-Business, e-Management and e-Learning, pp. 169, 2011.
[4] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis", Journal of Signal and Information Processing, pp.173, 2013
[5] J. Han, J. Pei, and M. Kamber, "Data mining: concepts and techniques", Elsevier, 2011.
[6] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm", Computers & Geosciences, vol 10, p. 191, 1984.
[7] A. Majdi, M. Beiki, "Applying evolutionary optimization algorithms for improving fuzzy C-mean clustering performance to predict the deformation modulus of rock mass", International Journal of Rock Mechanics and Mining Sciences, vol. 113 pp.172-182, 2019.
[8] A. Baykasoğlu, İ. Gölcük, F. B. Özsoydan, "Improving fuzzy c-means clustering via quantum-enhanced weighted superposition attraction algorithm", Hacettepe Journal of Mathematics and Statistics, vol. 48, pp. 859-882, 2018.
[9] X. Wu, J. Zhu, B. Wu, C. Zhao, J. Sun, C. Dai, "Discrimination of Chinese Liquors Based on Electronic Nose and Fuzzy Discriminant Principal Component Analysis". Foods, vol. 8, pp.38, 2019.
[10] Sapriadi, Sutarman, E. B. Nababan, "Improvement of K-Means Performance Using a Combination of Principal Component Analysis and Rapid Centroid Estimation", Journal of Physics: Conference Series 1230, 2019.
[11] S. C. Chin, X. Ji, W.L. Woo, T. J. Kwee, W. Yang, "Modified multiple generalized regression neural network models using fuzzy C-means with principal component analysis for noise prediction of offshore platform", Neural Computing and Applications, vol.31, pp.1127-1142, 2019.
[12] Hamed, Mohamed, "Application of Surface Water Quality Classification Models Using Principal Components Analysis and Cluster Analysis", SSRN Electronic Journal, vol. 10, pp. 2139, 2019.
[13] M. Premasundari and C. Yamini, "A violent crime analysis using fuzzy c-means clustering approach", ICTACT Journal on Soft Computing, vol. 9, pp. 1939-1944, 2019.
[14] Vijayarajan R, Muttan S Fuzzy C-Means Clustering Based Principal Component Averaging Fusion", International Journal of Fuzzy Systems, vol. 16, 2014.
[15] A. Jamal, A. Handayani, A. Septiandri, E. Ripmiatin, dan Y. Effendi, "Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction", Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, vol. 09, pp. 192-201, 2018.
[16] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.