**S-11**

# Penentuan Banyak Kelompok dalam Fuzzy C-Means Cluster Berdasarkan Proporsi Eigen Value Dari Matriks *Similarity* dan Indeks XB (Xie dan Beni)

**Anindya Apriliyanti Pravitasari**[3]

Email: dejafu_2008@yahoo.com

## Abstrak

Dalam analisis pengelompokkan (cluster), banyak kelompok menjadi suatu masalah yang berarti. Beberapa peneliti memilih banyak kelompok sesuai dengan kebutuhan dalam penelitiannya. Beberapa penelitian dalam analisis cluster lebih menitikberatkan pada struktur dan metode pengelompokkan yang terus berkembang dari waktu ke waktu. Metode terakhir yang sedang diminati adalah Fuzzy C-means Cluster. Fuzzy C-means Cluster melakukan pengelompokkan dengan prinsip meminimumkan fungsi objektif pengelompokkannya dimana salah satu parameternya adalah fungsi keanggotaan dalam fuzzy (sebagai pembobot) yang disebut juga dengan fuzzier (Klawonn dan Höppner, 2001). Makalah ini selain mengkaji metode pengelompokkan dengan Fuzzy C-means Cluster juga akan memilih banyak kelompok ideal dengan menggunakan indeks XB (Xie dan Beni). Untuk jumlah objek yang besar, indeks XB akan dihitung sebanyak objek yang dikelompokkan, maka hal ini tidaklah efektif. Untuk itu dicoba untuk membatasi banyak kelompok dengan menggunakan proporsi eigen value dari matriks kemiripan (*similarity*). Dengan membatasi banyak kelompok, perhitungan untuk mendapatkan kelompok ideal akan semakin cepat. Hal ini akan sangat berguna untuk efisiensi algoritma perhitungan indeks XB.

**Kata kunci** : analisis pengelompokkan, *cluster*, fuzzy c-means, indeks XB, proporsi, eigen value, matriks kemiripan, *similarity*.

[3] Dosen Jurusan Statistika, FMIPA, Universitas Padjadjaran Bandung

## 1. Pendahuluan

Analisis Cluster adalah salah satu analisis data eksploratori yang bertujuan untuk menentukan kelompok atau grup dari sekelompok data. Awal mulanya metode ini dikembangkan dengan menemukan struktur pengelompokkan diantara objek yang akan dikelompokkan. Paradigma data clustering mulai banyak diminati berbagai kalangan dan ditulis dalam berbagai paper dan jurnal (Shihab, 2000). Analisis cluster sempat disebut sebagai "*primary tool for so-called knowledge discovery*" (Fayyad *et al*, 1996) karena tingkat temuan struktur dan metode yang berkembang begitu pesat seiring dengan perkembangan paradigma lain diluar statistik, seperti fenomena data mining, *intelligent data analysis* (Liu, 2000), sampai *image processing* yang saat ini banyak diteliti. Faktanya, karena menggunakan data yang besar dan algoritma yang secara iteratif menentukan pengelompokkan, maka analisis cluster memiliki kepekaan akan kebutuhan yang tinggi dalam komputasi.

Perkembangan analisis cluster dimulai dari metode *hierarchical* yang secara garis besar membentuk sebuah *tree* diagram yang biasa disebut dengan dendogram yang mendeskripsikan pengelompokan berdasarkan jarak, *graph-theoritic* melihat objek sebagai *node* pada *network* terboboti, *mixture models* mengasumsikan suatu objek dihasilkan dari skala data yang berbeda-beda, *partitional* lebih dikenal dengan metode *non-hierarchy* termasuk didalamnya adalah metode K-means cluster. Perkembangan terakhir dari analisis cluster mempertimbangkan tingkat keanggotaan yang mencakup himpunan fuzzy sebagai dasar pembobotan bagi pengelompokan yang disebut dengan fuzzy clustering (Bezdek, 1981). Metode ini merupakan pengembangan dari metode partitional dengan pembobotan fuzzy yang memungkinkan pengelompokkan dimana kelompok data tidak terdistribusi secara jelas.

Sejalan dengan perkembangan metode dalam analisis cluster, penentuan jumlah kelompok tetap dilakukan secara subjektif. Metode *hierarchi* (single linkage, complete linkage dan average linkage) membuat *cut off* dari dendogram kemudian menentukan jumlah kelompok yang ideal. Metode *non-hierarci* atau *partitional* menentukan terlebih dahulu jumlah cluster yang akan dibentuk, termasuk juga pembentukan kelompok dalam Fuzzy C-means Cluster. Penentuan jumlah kelompok biasanya disesuaikan dengan tujuan penelitian. Suatu masalah kemudian timbul, bagaimana jumlah kelompok ideal yang meminimumkan fungsi objektif sebagai dasar pengelompokkan. Jika dilakukan pemilihan jumlah kelompok dari satu kesatuan kelompok besar sampai sebanyak objek yang akan dikelompokkan, maka penyelesaiannya akan trivial. Karena bagaimanapun juga fungsi objektif akan optimal saat jumlah kelompok yang terbentuk sama dengan jumlah objek yang dikelompokkan. Hal ini dikarenakan tingkat kemiripan (*similarity*) yang tinggi terhadap objek itu sendiri.

Oleh karena itu dicoba untuk membatasi jumlah pengelompokkan berdasarkan proporsi eigen value dari matrik *similarity* objek yang akan dikelompokkan. Prinsipnya hampir sama dengan principal komponen dan analisis faktor yang menggunakan proporsi eigen value sebagai ukuran kontribusi yang dapat diberikan ketika mereduksi dimensi variable. Pembatasan jumlah pengelompokkan ini kemudian dikontrol dengan Indeks XB (Xie dan Beni), dimana kelompok hasil pemilihan dari proporsi eigen value yang memaksimumkan Indeks XB adalah ukuran kelompok yang terbaik.

## 2. Fuzzy C-Means Cluster

Secara umum teknik dari fuzzy cluster adalah meminimumkan fungsi objektif dimana parameter utamanya adalah fungsi keanggotaan dalam fuzzy (*membership function*) yang disebut juga dengan fuzzier (awonn dan Höppner, 2001). Klawonn secara khusus mendalami fuzzy clustering sebagai metode yang baik untuk digunakan dalam pengelompokkan data spasial dan *image analysis* (*Laboratorium of Data analysis and Pattern Recognition*). Oleh karena itu sebagian besar referensi dari tulisan ini didapatkan dari jurnal penelitian Klawonn bersama peneliti lainnya.

Fuzzy C-means cluster pertama kali dikemukakan oleh Dunn (1973) dan kemudian dikembangkan oleh Bezdek (1981) yang banyak digunakan dalam *pattern recognition*. Metode ini merupakan pengembangan dari metode non hierarki K-means Cluster, karena pada awalnya ditentukan dulu jumlah kelompok atau cluster yang akan dibentuk. Kemudian dilakukan iterasi sampai mendapatkan keanggotaan kelompok tersebut. Metode ini adalah metode yang paling digemari karena merupakan metode yang paling robust (( Klawonn dan Höppner, 2001) dan (Klawonn, 2000)) dan memberikan hasil yang smooth (halus) dengan toleransi relatif (Shihab, 2000).

Prinsip utama pengelompokkan dengan fuzzy c-means cluster adalah meminimumkan fungsi objektif

$$J_{FCM}\left(\mathbf{P}, \mathbf{U}, \mathbf{X}, c, m\right) = \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^{m} d_{ik}^{2}\left(\mathbf{x}_{k}, \mathbf{p}_{i}\right) \quad (1)$$

dengan asumsi *constraint*

$$\sum_{i=1}^{c} u_{ik} = 1, \text{ untuk } \forall k \in \{1, \ldots, N\}. \quad (2)$$

Keterangan:

**P** dan **U** adalah variabel yang kondisi optimalnya diharapkan, untuk matriks **U** kondisi optimalnya berarti konvergensi keanggotaan kelompok dalam FCM. **X**, *c*, *m* adalah parameter input dari $J_{FCM}$, dimana:

- *c* adalah jumlah cluster yang memenuhi **X** (jumlah cluster yang diinginkan, $2 \le c < N$ )
- $m \ge 1$ adalah tingkat ke-fuzzy-an dari hasil pengelompokkan. Parameter ini disebut dengan fuzzier, nilai dari *m* yang sering dipakai dan dianggap yang paling halus adalah *m*=2 (Klawonn dan Höppner, 2001)
- $u_{ik}$ adalah tingkat keanggotaan yang merupakan elemen dari matriks **U.**
- *N* jumlah observasi.
- $d_{ik}^2$ adalah jarak observasi yang dapat dirumuskan sebagai berikut:

$$d_{ik}^2(\mathbf{x}_k, \mathbf{p}_i) = \left\|\mathbf{x}_k - \mathbf{p}_i\right\|_A^2 = (\mathbf{x}_k - \mathbf{p}_i)^T A(\mathbf{x}_k - \mathbf{p}_i) \quad (3)$$

Jika A adalah matriks identitas maka $d_{ik}^2$ adalah jarak Euclid.

Algoritma pengelompokan Fuzzy C-means cluster diberikan sebagai berikut:

1. Menentukan *c* banyak cluster atau kelompok yang ingin dibuat.
2. Menentukan tingkat ke-fuzzy-an hasil pengelompokan (*m*).
3. Menghitung fuzzy cluster center (**P**) dengan persamaan (2)

$$\mathbf{p}_i^* = \frac{\sum_{k=1}^{N} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^m} \quad (4)$$

4. Update anggota matriks **U** dengan persamaan

$$u_{ik}^* = \frac{1}{\sum_{j=1}^{c} \left( \frac{d_{ik}^2}{d_{jk}^2} \right)^{1/(m-1)}} \quad (5)$$

5. Bandingkan nilai keanggotaan dalam matriks **U**, jika tidak banyak mengalami perubahan maka artinya sudah konvergen dan keanggotaannya sudah maksimal. Iterasi dihentikan dan didapatkan hasil pengelompokkan.

### 3. Penentuan Jumlah Kelompok

Penentuan jumlah kelompok dalam Fuzzy C-Means Cluster didasarkan pada dua hal. Yang pertama adalah dengan membatasi jumlah kelompok yang terbentuk melalui proporsi eigen value matriks *similarity* dari objek yang akan dikelompokkan. Yang kedua adalah melakukan kontrol dengan indeks XB. Tujuannya adalah untuk mengetahui apakah benar jumlah cluster terbaik bisa didapatkan diantara jumlah *cluster* yang dibatasi oleh proporsi eigen value matriks *similarity*. Berikut ini adalah ulasan mengenai proporsi eigen value matriks *similarity* dan Indeks XB.

**Proporsi Eigen Value Matriks *Similarity***

Analisis Eigen adalah salah satu teknik yang memberikan ringkasan struktur data yang direpresentasikan oleh matriks korelasi ataupun kovarians (Johnson dan Wichern, 2002). Proporsi dari eigen value menggambarkan seberapa besar struktur data yang dapat diwakili atau direpresentasikan oleh matriks korelasi atau kovarians tersebut. Dalam analisis komponen utama dan analisis faktor, proporsi dari eigen value memberikan interpretasi mengenai seberapa besar data dapat terwakili dalam dimensi yang telah direduksi.

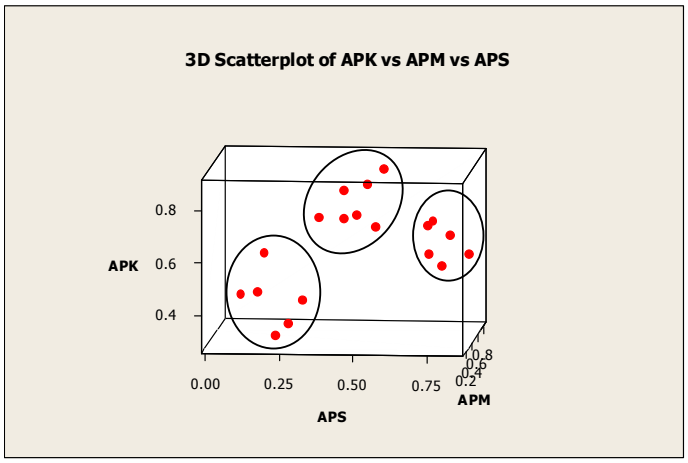
Pada kasus pengelompokkan dalam analisis cluster, matriks korelasi yang digunakan adalah matriks *similarity* dari objek yang akan dikelompokkan. Prinsipnya adalah semakin tinggi nilai similaritas antara objek satu dengan yang lain maka nilai pengamatan antar objek tersebut memiliki banyak kesamaan (berarti memungkinkan untuk menjadi satu kelompok). Proporsi eigen value untuk ilustrasi ini berarti memberikan informasi besarnya tingkat kesamaan antar objek. Proporsi eigen value 100 persen diberikan oleh semua eigen yang terbentuk yang banyaknya sama dengan banyak objek yang dikelompokkan. Berikut ini ilustrasi menggunakan data angka partisipasi pendidikan di Kabupaten Tuban (Angka partisipasi Murni, Angka Partisipasi Kasar dan Angka Partisipsi Sekolah). Pengelompokkan dilakukan untuk kecamatan-kecamatan pada jenjang pendidikan SMU. Hasil perhitungan eigen value dan kumulatif proporsi diberikan pada Tabel1.

Tabel 1. Eigen Value matrik *similarity* dan proporsi kumulatif

| i | Eigen | Proporsi | Kumulatif |
|---|---|---|---|
| 1 | 11.2358 | 0.517858 | 0.517858 |
| 2 | 6.2477 | 0.287956 | 0.805814 |
| 3 | 4.2132 | 0.194186 | 1 |

Dari nilai pada Tabel 1, dapat dilihat bahwa proporsi 100 persen dengan cepat diberikan oleh tiga eigen value. Maka kesimpulan sementara yang dapat diambil adalah pengelompokkan yang terbentuk sebanyak tiga cluster. Pada gambar 1 diberikan sebaran data angka partisipasi pendidikan dalam scatter plot 3D. Dari gambar 1 terlihat bahwa data berkumpul dalam tiga kelompok besar. Maka indikasi terbentuknya tiga cluster semakin kuat.

Selanjutnya akan dilakukan kontrol dengan perhitungan indeks XB. Untuk perhitungannya akan dilakukan untuk jumlah cluster 2, 3, 4 dan 5. sub bahasan selanjutnya akan membahas tentang indeks XB dan perhitungannya untuk contoh data yang sama.




Gambar 1. Scatter plot data partisipasi pendidikan.

### 3.1 Indeks XB (Xie dan Beni)

Sesuai dengan namanya Indeks XB ditemukan oleh Xie dan Beni yang pertama kali dikemukakan pada tahun 1991. Validitas dalam FCM ditentukan oleh banyak kelompok optimum melalui perhitungan Indeks validitas.

Formula dari Indeks XB diberikan pada (6). Formula ini mirip dengan Separation Index dengan nilai m yang dapat berubah-ubah, oleh karena itu indeks ini dapat digunakan untuk metode hard partition seperti K-means cluster maupun FCM. Kriterianya banyak kelompok optimum diberikan oleh nilai XB yang minimum pada lembah pertama.

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m d_{ik}^2(\mathbf{x}_k, \mathbf{p}_i)}{N \min_{i,j}\left\|\mathbf{p}_i, \mathbf{p}_j\right\|^2} \quad (6)$$

Dengan c menyatakan banyak cluster, uik adalah tingkat keanggotaan, $d_{ik}^2$ adalah jarak observasi dengan pusat cluster, pi adalah pusat cluster, N merupakan banyak objek yang akan dikelompokkan, $\min_{i,j}\left\|\mathbf{p}_i, \mathbf{p}_j\right\|^2$ menyatakan jarak minimum antara pusat cluster pi dan pj. Kriteria banyak cluster optimum diberikan oleh indeks XB yang minimum.

Rekomendasi untuk menggunakan indeks XB tertuang dalam penelitian Duo dkk (2007) yang menyatakan bahwa indeks XB memiliki ketepatan dan keandalan yang tinggi baik untuk memberikan banyak kelompok optimum pada metode hard partition seperti K-means cluster maupun pada FCM.

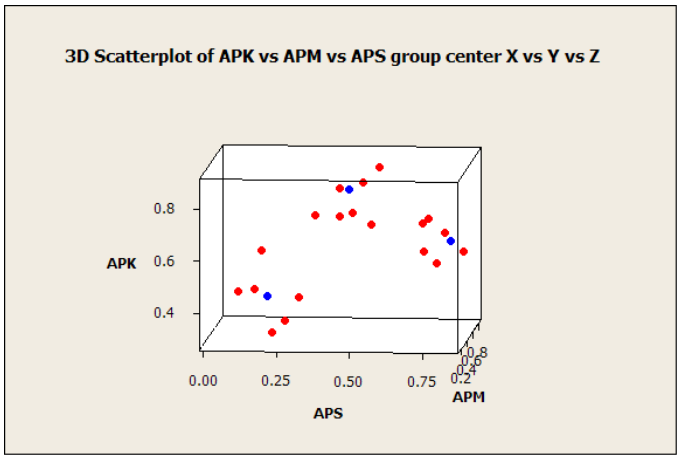
Hasil pengelompokkan Fuzzy C-means Cluster dengan bantuan matlab beserta indeks XB diberikan dalam Tabel 2. Dari Tabel 2 terlihat bahwa indeks XB maksimum diberikan oleh pengelompokkan dengan 3 cluster, maka kesimpulan awal yang diberikan sudah tepat, yaitu bahwa proporsi eigen value dapat dijadikan patokan awal untuk pemilihan banyaknya pengelompokkan dalam perhitungan indeks XB. Sehingga untuk ukuran data yang sangat besar, perhitungan indeks XB tidak perlu dilakukan sampai jumlah kelompok maksimum (sama dengan jumlah objek), namun perhitungannya dapat mengikuti informasi yang diberikan oleh banyaknya eigen dengan proporsi kumulatif yang cukup tinggi. Plot data beserta nilai pusat $\mathbf{p}_i$ dalam Fuzzy C-means Cluster diberikan oleh Gambar 2.

## 4. Kesimpulan

Penentuan jumlah cluster yang ideal dapat dilakukan dengan perhitungan indeks XB. Namun untuk jumlah data yang besar, maka perhitungan indeks XB akan dilakukan sampai jumlah pengelompokkan maksimum, yaitu sebanyak jumlah objek itu sendiri. Hal ini kurang efektif, maka direkomendasikan untuk menentukan banyaknya cluster yang mungkin terbentuk dengan memperhatikan proporsi kumulatif eigen value matriks *similarity* dari objek dalam pengelompokkan.

Tabel 2. Hasil Cluster dan Indeks XB

| Kecamatan | Ukuran Cluster | | | |
|---|---|---|---|---|
| | 2 | 3 | 4 | 5 |
| KENDURUAN | 2 | 3 | 3 | 1 |
| BANGILAN | 2 | 3 | 4 | 5 |
| SENORI | 1 | 1 | 2 | 2 |
| SINGGAHAN | 1 | 1 | 2 | 2 |
| MONTONG | 2 | 2 | 1 | 3 |
| PARENGAN | 2 | 2 | 3 | 3 |
| SOKO | 1 | 1 | 2 | 2 |
| RENGEL | 2 | 3 | 4 | 5 |
| PLUMPANG | 1 | 1 | 2 | 2 |
| WIDANG | 2 | 2 | 1 | 4 |
| PALANG | 1 | 1 | 2 | 2 |
| SEMANDING | 2 | 2 | 1 | 3 |
| TUBAN | 2 | 3 | 3 | 1 |
| JENU | 1 | 1 | 2 | 2 |
| MERAKURAK | 2 | 2 | 1 | 3 |
| KEREK | 2 | 3 | 4 | 4 |
| TAMBAKBOYO | 2 | 3 | 3 | 1 |
| JATIROGO | 2 | 3 | 4 | 5 |
| BANCAR | 2 | 2 | 1 | 3 |
| *Indeks XB* | 2.26 | 3.98* | 3.36 | 1.39 |




Gambar 2. Scatter Plot dengan pusat $\mathbf{p}_i$

**Referensi :**

Bezdek, James., 1981. *Pattern Recognition with Fuzzy Objective Function Algorith*, Plenum Press, New York.

Calinski and Harabasz, (1974), "A Dendrite Method for Cluster Analysis". Communication in Statistics 3, 1-27.

Dunn, J.C., (1973), "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact well-Separated Cluster", Journal of Cybernetic 3, 32-57.

Fayyad, U, M., Piatetsky-Saphiro, G., Smith., (1996). Advance and Knowledge discovery and data mining, Part 2.33, http://AAIPress.com/AdvanceKnowledgedisc-Fayyadetal//

Johnson, Wichern, (2002), *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Klawonn, Frank., (2000), "Fuzzy Clustering: Insight and a New Approach", Science Journal, http://public.rz.fh-wolfenbuettel.de/klawonn.

Klawonn dand Höppner, (2001), "What is Fuzzy about Fuzzy Clustering? Understanding and Improving the Concept of the Fuzzier". Science Journal, http://public.rz.fh-wolfenbuettel.de/klawonn.

Klawonn dan Keller, (1997), "Fuzzy Clustering and Fuzzy Rules", Science Journal, http://public.rz.fh-wolfenbuettel.de/klawonn.

Klawonn dan Klementida, (1997), "Matematical Analysis of Fuzzy Clasifiers", Science Journal, http://public.rz.fh-wolfenbuettel.de/klawonn.

Klawonn dan Kruse, (1995), "Clustering Method in Fuzzy Control", Science Journal, http://public.rz.fh-wolfenbuettel.de/klawonn.

Sharma, S, (1996), *Applied Multivariate Techniques,* John Wiley and Sons, Inc, New York.

Shihab, A.I., (2000) "Fuzzy Clustering Algorithm and Their Application to Medical Image Analysis". Dissertation, University of London, London.

Pickert, Klawonn, dan Wingender., (1997), "Fuzzy Cluster Analysis for Identification of Gene Regulation Region". Science Journal, http://public.rz.fh-wolfenbuettel.de/klawonn.

Zadeh, Lotfi. A., (1965), Fuzzy Sets. *Information Control*, vol 8, 338-353.