J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 12, No. 1, 30-35, Juli 2015
ISBN:1858-1382

Pengklasteran dengan Algoritma Fuzzy *C*-Means

Muhammad Abdy*

Abstrak

Pengklasteran merupakan proses mengelompokan data berdasarkan kemiripan atau kedekatannya. *Hard-clustering* akan mengelompokan data ke dalam klaster-klaster dimana setiap titik data akan berada dalam tepat satu klaster, sementara *soft-clustering* akan mengelompokan data dalam klaster dimana setiap data dapat menjadi anggota dari beberapa klaster dengan derajat keanggotaan yang berbeda-beda. Salah satu soft-clustering yang sangat populer adalah fuzzy *c*-mean, yaitu sutu algoritma pengklasteran yang mencari pusat-pusat klaster dengan meminimumkan fungsi ketidakmiripan. Pada tulisan ini akan dibahas algoritma fuzzy *c*-means dan akan diberikan contoh data simulasi.

Kata Kunci: Pengklasteran, hard-clustering, soft-clustering, Fuzzy C-Means.

1. Pendahuluan

Pengklasteran merupakan suatu proses untuk membuat pengelompokan dari sekumpulan objek berdasarkan kemiripan (*similarity*) atau kedekatan (*proximity*). Hasil dari pengklasteran adalah klaster-klaster yang merupakan bagian-bagian dari sekumpulan objek yang memiliki kemiripan dalam klaster yang sama dan memiliki ketakmiripan (*dissimilarity*) dengan objek yang lain dalam klaster yang berbeda. Pada pengklasteran konvensional (*hard-clustering*), objek-objek akan terpartisi ke dalam klaster-klaster dimana suatu objek akan menjadi anggota dari tepat satu klaster (*hard-partition*). Secara formal didefinisikan sebagai berikut.

Definisi 1.
Misalkan X adalah suatu himpunan data dan $x_i \in X$.
Suatu partisi $P = \{C_1, C_2, ..., C_L\}$ *dari X adalah hard-partition jika dan hanya jika memenuhi:*

(i) $\forall x_i \in X, \exists C_j \in P \ni x_i \in C_j$

(ii) $\forall x_i \in X, x_i \in C_j \Rightarrow x_i \in C_k$ *dimana* $k \neq j$, $C_j \in P$

Syarat pertama dari definisi tersebut menjamin bahwa semua titik data *X* akan menjadi anggota dari suatu klaster, dan syarat kedua menjamin bahwa semua klaster adalah *mutually exclusive*. Banyak permasalahan pengklasteran dalam kehidupan sehari-hari yang tidak sesuai dengan hard-clustering. Sebagai contoh, misalnya suatu kabupaten dibagi menjadi tiga klaster berdasarkan komoditi pertanian yang dihasilkan, yaitu klaster I untuk komoditi sayur-sayuran, klaster II untuk komoditi padi, dan klaster III untuk komoditi jagung. Apabila digunakan hard-clustering, maka kecamatan A yang 50% hasil pertaniannya adalah padi dan 30% adalah sayur-sayuran, akan masuk dalam klaster II, padahal kecamatan A tersebut dapat juga berada di klaster I dengan tingkat keanggotaan yang berbeda di dalam klaster II. Untuk mengatasi permasalahan demikian, diperkenalkanlah soft-clustering atau fuzzy-clustering, sehingga objek-objek akan terpartisi ke dalam klaster-klaster dimana suatu objek dapat menjadi anggota dari beberapa klaster dengan derajat kenggotaan tertentu (*soft-partition*). Secara formal didefinisikan sebagai berikut.

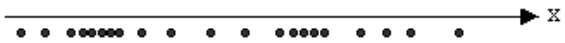
* *Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar, e-mail: abdy02@yahoo.com.*

Definisi 2.
Misalkan X adalah suatu himpunan data dan $x_i \in X$.
Suatu partisi $P = \{C_1, C_2, ..., C_L\}$ *dari X adalah soft-partition jika dan hanya jika memenuhi*

(i) $\forall x_i \in X, \exists C_j \in P \ni 0 \leq \mu_{C_j}(x_i) \leq 1$

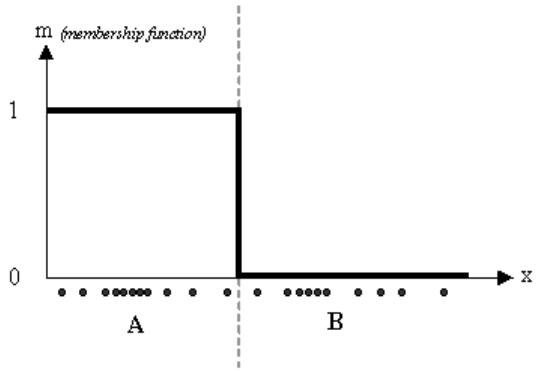
(ii) $\forall x_i \in X, C_j \in P \ni \mu_{C_j}(x_i) > 0$, *dimana* $\mu_{C_j}(x_i)$ *adalah derajat keanggotaan* x_i *dalam klaster* C_j.

Berikut diberikan suatu ilustrasi tentang hard-clustering dan soft-clustering dengan menggunakan data fiktif yang berdistribusi pada suatu sumbu, seperti diperlihatkan pada Gambar 1.



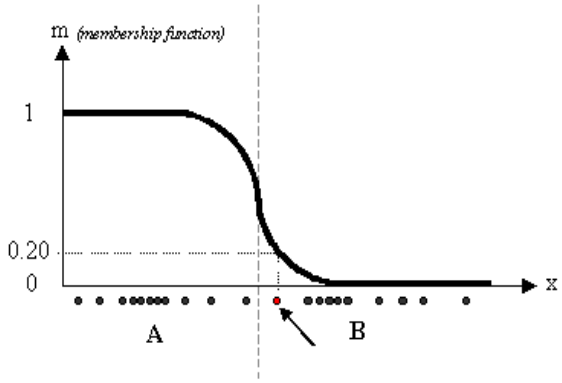
Gambar 1. Titik-titik Data.

Misalkan data tersebut dibagi menjadi dua klaster (A dan B), maka dengan menggunakan hard-clustering, derajat keanggotaan titik-titik data di dalam klaster A dan B adalah seperti pada Gambar 2.



Gambar 2. Fungsi Keanggotaan Hard-clustering.

Pada Gambar 2, semua titik-titik data dalam klaster B mempunyai derajat keanggotaan nol dalam klaster A, dan demikian sebaliknya. Akan tetapi, jika digunakan fuzzy-clustering maka suatu titik data dapat menjadi anggota dari kedua klaster dengan tingkat keanggotaan yang berbeda. Pada Gambar 3, titik data yang bertanda merah merupakan anggota dari klaster B dengan derajat keanggotaan 0.8, tetapi titik data tersebut juga merupakan anggota dari klaster A dengan derajat keanggotaan 0.2.



Gambar 3. Fungsi Keanggotaan Soft-clustering.

2. Fuzzy *C*-Means

Suatu *fuzzy-partition* yang memenuhi syarat tambahan $\sum_j \mu_{C_j}(x_i)=1$ disebut soft-partisi terkendala. Algoritma fuzzy *c*-means merupakan salah satu algoritma pengklasteran fuzzy yang menghasilkan soft-partisi terkendala dan merupakan suatu metode partisi iteratif yang bertujuan menemukan pusat klaster yang meminimumkan fungsi ketidakmiripan sehingga menghasilkan c-partisi optimal. Metode tersebut menghitung pusat klaster (*centroid*) dan membangkitkan matriks kelas keanggotaan. Metode ini dikembangkan oleh Dunn (1973) dan diperbaiki oleh Bezdek (1981). Metode ini juga sering digunakan dalam pengenalan pola.

Misalkan $X=\{x_k\}_{k\in[1,\,n]}$ adalah suatu himpunan berhingga. M_{cxn} adalah matriks yang entri-entrinya ada dalam interval [0,1], dan c ($2 < c < n$) adalah suatu bilangan bulat. Matriks $U = (\mu_{ik})_{(i,\,k)\in[1,\,c]x[1,\,n]} \in M_{cxn}$ disebut fuzzy *c*-partisi dari X jika memenuhi syarat berikut:

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

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

Lokasi dari suatu klaster diwakili oleh pusatnya, $v_i = (v_{ij})_{j\in[1,\,p]} \in R^p$, disekitar objek-objeknya berkosentrasi. Untuk memperbaiki partisi awal, digunakan kriteria variansi yang mengukur ketidakmiripan di antara titik-titik dalam suatu klaster dan pusat klasternya. Kriteria yang digunakan adalah jarak Euclidean $d_{ik} = d(x_k, v_i)$, dimana

$$d(x_k,v_i) = \left\| x_k - v_i \right\| = \left[\sum_{j=1}^{p} (x_{kj} - v_{ij})^2 \right]^{1/2}. \qquad (2)$$

Fungsi ketidakmiripan (fungsi tujuan) yang digunakan dalam fuzzy *c*-mean adalah

$$J(U, v_1, v_2, ..., v_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}{}^m d_{ij}^2 . \qquad (3)$$

dengan $\mu_{ij}\in[0,1]$, v_i adalah pusat klaster ke-*i*, d_{ij} adalah jarak Euclidean antar pusat klaster ke-*i* dan data ke-*j*, $m \in [1,\infty)$ adalah suatu exponen pembobot yang menentukan tingkat kekaburan klaster (*fuzziness cluster*). Untuk m=1, maka klastering akan menjadi hard-clustering. Fuzzy *c*-mean memperoleh partisi yang baik dengan mencari prototype atau pusat klaster v_i yang

meminimumkan fungsi tujuan. Dengan mendifferensialkan fungsi tujuan pada persamaan (3) terhadap v_i (U konstan) dan terhadap μ_{ij} (v konstan) dengan kendala $\sum_{i=1}^{c} \mu_{ij} = 1$, maka fungsi tujuan akan minimum jika dan hanya jika

$$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m x_j}{\sum_{j=1}^{n} \mu_{ij}^m}, \quad (4)$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}}, \quad (5)$$

Secara detail, algoritma fuzzy *c*-means mempunyai tahapan-tahapan sebagai berikut:
Tahap 1 : Pilih suatu nilai untuk parameter fuzziness klaster *m*, dengan $m > 1$;
Tahap 2 : Pilih suatu nilai untuk kriteria penghentian iterasi ε (yaitu $\varepsilon = 0.0001$ memberikan suatu konvergensi yang layak);
Tahap 3 : Pilih suatu ukuran jarak dalam variabel–space (yaitu jarak Euclidean);
Tahap 4 : Pilih banyaknya kelas atau grup *c*, dengan $c = 2, 3, ..., n\text{-}1$;
Tahap 5 : Inisialisasi secara acak matriks keanggotaan (U) dengan kendala $\sum_{i=1}^{c} \mu_{ij} = 1$, untuk setiap $j = 1, 2, ..., n$;
Tahap 6: Hitung pusat klaster (v_i) dengan menggunakan persamaan (4);
Tahap 7: Hitung ketidakmiripan diantara pusat klaster dan titik data dengan menggunakan persamaan (3). Hentikan iterasi jika $\left\| U^{[k+1]} - U^{[k]} \right\| < \varepsilon$.
Tahap 8: Hitung suatu U baru dengan persamaan (5). Lanjutkan ke tahap 6.

Algoritma fuzzy *c*-means mempartisi suatu himpunan data ke dalam sejumlah klaster *c* yang telah ditentukan sebelumnya secara bebas. Oleh karena itu, diperlukan suatu kriteria untuk menentukan banyaknya klaster optimal dalam data, yang biasa disebut masalah validitas klaster (Fauziah, 2008).

3. Validitas Klaster

Validitas klaster merupakan suatu masalah krusial dalam aplikasi tehnik fuzzy-clustering. Sebagaimana diketahui bahwa tujuan dari pengklasteran adalah mengelompokan objek-objek dalam klaster-klaster sedemikian rupa sehingga asosiasi atau kemiripan dari objek-objek dalam klaster yang sama adalah besar, dan kecil untuk objek dalam klaster yang berbeda. Oleh karena itu, kekompakan (*compatcness*) dan keterpisahan (*separation*) merupakan ukuran yang layak untuk menilai kebaikan (*goodness*) dari klaster yang dihasilkan.

Validitas fungsional yang paling banyak digunakan adalah koefisien partisi, entropi partisi dan exponen partisi (Fauziah, 2008). Koefisien partisi dan entropi partisi merupakan indeks yang dihitung hanya dengan menggunakan elemen matriks keanggotaan.

3.1. Koefisien Partisi

Misalkan $U \in M_{cxn}$ adalah fuzzy *c*-partisi dari *n* titik data. Koefisien partisi (Bezdek, 1981) dari *U* adalah

$$F(U, c) = \sum_{k=1}^{n} \sum_{i=1}^{c} \frac{u_{ik}^2}{n} \quad (6)$$

Misalkan bahwa Ω_c adalah hasil pengklasteran, maka pemilihan optimal dari *c* adalah

$$\max_c \left\{ \max_{\Omega_c} F(U,c) \right\}, \; c = 1, 2, 3, ..., n \quad (7)$$

3.2. Entropi Partisi

Bezdek (1981) menyatakan bahwa entropi partisi, dari sebarang fuzzy c-partition $U \in M_{cxn}$ dari *X*, dimana $|X| = n$, $1 \le c \le n$ adalah

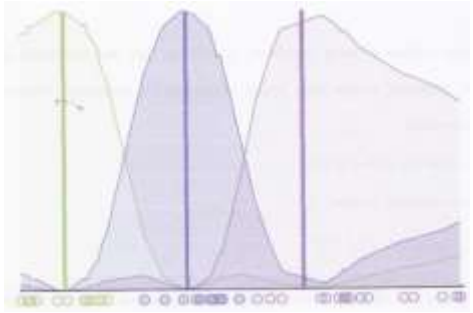
$$H(U,c) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik} \ln(\mu_{ik}) \quad (8)$$

Pemilihan optimal *c* adalah

$$\min_c \left\{ \min_{\Omega_c} H(U,c) \right\}, \; c = 1, 2, 3, ..., n \quad (9)$$

4. Data Simulasi

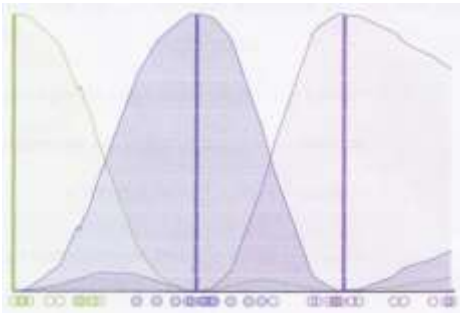
Pada bagian ini diberikan data simulasi sederhana yang akan diklaster dengan fuzzy *c*-means. dimana algoritma fuzzy *c*-means pada bagian 3 disusun dalam program Matlab. Empat puluh data dibangkitkan, dan dipilih tiga klaster. Kriteria penghentian yang dipilih adalah 0.0001 dan fuzziness klaster $m = 2$.



Gambar 4. Fungsi Keanggotaan Titik Data dalam Tiap Klaster pada Keadaan Awal ($U^{[0]}$).

Gambar 4 memperlihatkan keadaan awal sebelum iterasi (matriks inisial $U^{[0]}$). Garis vertikal merupakan pusat klaster awal yang ditentukan. Setelah dilakukan iterasi sampai $\left\| U^{[k+1]} - U^{[k]} \right\| < 0.0001$, maka posisi pusat klaster berubah, seperti yang ditunjukkan pada

Gambar 5. Pemilihan pusat klaster dan ε yang akurat pada keadaan awal akan menentukan panjangnya iterasi.



Gambar 5. Fungsi Keanggotaan Titik Data dalam Tiap Klaster Setelah Iterasi ($U^{[k]}$).

Daftar Pustaka

Bezdek, J.C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum, New York.

Cuaves, E. *et al.*, 2004. Fuzzy segmentation applied to face segmentation. *Technical Report*, B-04-09. Institut Fur Informatik, Freie Universitat Berlin, Germany.

Demko, C., 1995. Image understanding using fuzzy isomorphism of fuzzy structures. *Proceeding of the FUZZ-IEEE/IFES '95 Confrence*, Japan, Yokohama.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting well separated cluster. *Journal of Cybernetics*, 3, 32 – 57.

Fauziah, Z., 2008. Dynamic profiling of EEG during zeizure using fuzzy information space. *PhD Thesis*, Faculty of Science, UTM Malaysia