Klasterisasi Data Kategorikal Berbasis Fuzzy K-Modes Dan Artificial Bee Colony

Khalid¹, Handayani Tjandrasa²
^{1,2}Jurusan Teknik Informatika, ITS, Surabaya, Indonesia
philllette@gmail.com

Abstrak

Fuzzy K-Modes (FKMO) merupakan metode klasterisasi data yang efektif untuk data kategorikal. Metode ini menggunakan metode fuzzy dan pencocokan ukuran ketidaksamaan (*dissimilarity measure*) yang sederhana untuk memutakhirkan titik pusat klaster dan mendapatkan solusi yang optimal. Meskipun demikian Fuzzy K-Modes memiliki kelemahan adanya kemungkinan berhenti dalam solusi lokal optimal.

Artificial Bee Colony (ABC) merupakan metode yang efektif untuk melakukan optimasi dan terbukti memiliki solusi global. Penelitian ini mengusulkan penggunaan algoritma Artificial Bee Colony untuk melakukan optimasi terhadap Fuzzy K-Modes (ABC-FKMO).

Implementasi Artificial Bee Colony untuk optimasi Fuzzy K-Modes terbukti mampu meningkatkan performance klasterisasi khususnya dalam aspek nilai Objective Function, F-Measure, dan Accuracy. Hasil pengujian dengan dataset Soybean Disease, Breast Cancer dan Congressional Voting Records dari UCI data repository, menunjukkan rata-rata accuracy sebesar 0.991, 0.615, dan 0.867. Objective Function lebih baik rata-rata sebesar 2,73 %, F-Measure lebih baik rata-rata sebesar 4,31 % dan Accuracy lebih baik rata-rata sebesar 5,16 %.

Kata Kunci : Data Kategorikal, Klasterisasi, Fuzzy K-Modes, Artificial Bee Colony

1. Pendahuluan

Dewasa ini, penelitian di bidang klasterisasi data kategorikal sudah mulai berkembang, walaupun perkembangannya masih jauh lebih sedikit dibanding klasterisasi pada tipe data numerik. Data kategorikal secara alami tidak bisa di perlakukan sebagai data numerik karena ada beberapa operasi dalam data numerik yang tidak bisa di lakukan dalam data kategorikal seperti mean dan median. Sebagai contoh atribut data kategorikal adalah atribut berdomain jenis kelamin (pria, wanita), domain agama (Islam, Kristen, Katolik, Hindu dan sebagainya), dan domain etnis (mongoloid, kaukasoid, negroid).

Klasterisasi merupakan alat utama dalam berbagai aplikasi dalam analisa data statistik, data mining, information retrieval, pengolahan citra dan sebagainya. Klasterisasi bertujuan melakukan pengelompokan obyek/data ke dalam beberapa klaster/kelompok sehingga obyek/data dalam satu klaster memiliki tingkat kesamaan yang maksimum dan data antar klaster memiliki kesamaan yang minimum (Tan, Steinbach dan kumar, 2006; Han dan Kamber, 2006).

Berdasar pendekatan dalam penetapan keanggotaan dalam klaster, Metode klasterisasi secara umum dapat dibagi menjadi dua yaitu *hard clustering* dan *fuzzy clustering* (Jain, Murty dan Flynn, 1999). Pada *Hard clustering*, tiap obyek/data hanya di alokasikan ke dalam satu satu klaster baik selama operasi klasterisasi maupun dalam output klasterisasi. Sedang pada

Fuzzy clustering, selama operasi klasterisasi tiap obyek/data di alokasikan ke dalam beberapa klaster dan di beri derajat keanggotaan dengan nilai antara 0 dan 1. Output *Fuzzy Clustering* dapat di ubah menjadi *hard clustering* dengan memilih nilai keanggotaan tertinggi.

Terdapat berbagai macam metode klasterisasi data kategorikal, diantaranya ROCK (Guha, Rastogi dan Shim, 2000), LIMBO (Andritsos, et al. 2004), K-Modes, Fuzzy K-Modes, dan lain lain. Salah satu metode yang paling populer adalah K-Modes Clustering (Huang, 1998). K-Modes termasuk dalam *Hard Clustering*. K-Modes merupakan pengembangan dari metode K-Means agar dapat di gunakan untuk klasterisasi data kategorikal. K-Modes menggunakan sebuah ukuran jarak (*dissimilarity*) berupa kecocokan suatu nilai atribut tiap dimensi terhadap titik pusat klaster, menggantikan mean dengan modus, dan menggunakan metode berbasis frekuensi untuk memutakhirkan modus dalam proses meminimalkan jarak (*dissimilarity*) dari seluruh data ke pusat klaster masing masing. Karena K-Modes yang dikembangkan Huang merupakan pengembangan K-Means, maka K-Modes mempunyai karakteristik dan kelemahan yang sama dengan K-Means. Kelemahan tersebut adalah keakuratan hasil klaster sangat tergantung dari penentuan titik awal pusat klaster sehingga sensitif terhadap penentuan titik awal. Dan memungkinkan hasil klaster konvergen pada lokal optimal (Funderlic, et al. 2004; Gan, Yang dan Wu, 2005).

Untuk mengatasi masalah yang terjadi pada inialisasi pusat kluster dan keanggotaan kluster dalam metode K-Modes, Huang dan Ng mengusulkan sebuah metode baru yang di beri nama Fuzzy K-Modes (Huang dan Ng, 1999). Fuzzy K-Modes termasuk dalam *Fuzzy Clustering*. Merupakan pengembangan Fuzzy C-Means dengan menggunakan generate *fuzzy partition matrix* dari data kategorikal. Metode Fuzzy K-Modes juga masih mungkin terjadi permasalahan lokal optimal (Gan, Yang dan Wu, 2009). Untuk mengatasi masalah lokal optimal ini, diperlukan suatu metode yang mampu mencari sebuah solusi global optimal dan menghindari kemungkinan solusi lokal optimal. Artificial Bee Colony (ABC) merupakan suatu metode *metaheuristik* yang mengadopsi perilaku koloni lebah madu dalam mencari makan (*foraging behavior*). Metode ABC telah terbukti memiliki kemampuan untuk menangani permasalahan lokal optimal dan memiliki kualitas yang lebih baik dibandingkan dengan metode *metaheuristik* lainnya seperti Algoritma Genetika, Particle Swarm Optimization, dan Differential Evolution (Karaboga dan Akay, 2009). Dalam penelitian ini diusulkan sebuah metode baru untuk klasterisasi data kategorikal berbasis pada hibridasi metode Fuzzy K-Modes dan Artificial Bee Colony (ABC) yang diberi nama ABC-FKMO. ABC di gunakan untuk membantu Fuzzy K-Modes agar dapat keluar dari permasalahan lokal optimal sehingga dengan metode ABC-FKMO ini di harapkan mampu mengoptimalkan posisi titik pusat kluster yang mengarah pada solusi global optimal.

2. Artificial Bee Colony dan Fuzzy K-Modes untuk Klasterisasi Data Kategorikal.

Gambar 1 Blok Diagram Penelitian

Gambaran tentang metode yang diusulkan dalam penelitian ini bisa dijelaskan dalam Gambar 1. Dua tahap penting dalam penelitian ini adalah *Preprocessing*, dan klasterisasi. *Preprocessing* dilakukan untuk mengubah nilai *missing value* dengan modus tiap fitur. Berikutnya data yang sudah dipreprocessing diklasterisasi dengan Fuzzy K-modes dan dilakukan optimasi terhadap Fuzzy K-Modes menggunakan Artificial Bee Colony. Setelah tahap klasterisasi, parameter uji bisa dihitung untuk mengevaluasi kinerja metode yang diusulkan. Parameter yang digunakan adalah *Objective Function*, *Accuracy*, *F-Measure*, dan waktu komputasi.

2.1 Fuzzy K-Modes

Metode klasterisasi Fuzzy K-Modes merupakan versi fuzzy dari K-Modes dan juga Merupakan pengembangan dan modifikasi dari metode Fuzzy C-Means agar dapat di gunakan untuk klasterisasi data kategorikal. Fuzzy K-Modes pertama kali dikenalkan oleh Huang (Huang dan Ng,1999). Secara prinsip, perbedaan antara K-Modes dan Fuzzy K-Modes terletak pada proses penentuan pusat kluster dan Penentuan membership tiap titik.

Berikut ini akan di berikan gambaran metode Klasterisasi Fuzzy K-Modes. Anggap $D=\{X,Y\}$ adalah sebuah dataset kategorikal. X dan Y adalah dua obyek kategorikal yang dideskripsikan oleh $[x_1, x_2, x_3, \dots, x_m]$ dan $[y_1, y_2, y_3, \dots, y_m]$. Maka ukuran ketidaksamaan antara X dan Y dapat didefinisikan sebagai total ketidaksamaan pada atribut kategori yang berkorespondensi pada dua obyek. Semakin kecil jumlah ketidaksamaan, semakin mirip dua obyek tersebut. Ketidaksamaan tersebut dapat di definisikan sebagai berikut :

$$d(X, Y) = \sum_{j=1}^m \delta(x_j, y_j) \quad (1)$$

Dimana

$$\delta(x_j, y_j) = \begin{cases} 0 & x_j = y_j \\ 1 & x_j \neq y_j \end{cases}$$

$d(X, Y)$ menggambarkan ketidaksamaan untuk tiap atribut.

Fungsi obyektif dari Fuzzy K-Modes adalah mencari W dan Z untuk meminimalisasi

$$F(W, Z) = \sum_{l=1}^k \sum_{i=1}^n w_{li}^\alpha d(Z_l, X_i) \quad (2)$$

Dimana $\alpha > 1$ adalah komponen pembobotan ($\alpha = 1$ di gunakan untuk hard K-Modes). $W = (w_{li})$ merupakan matrik keanggotaan fuzzy $k \times n$. $Z = \{z_1, z_2, z_3, \dots, z_k\}$ merupakan himpunan pusat kluster dan z_i merupakan pusat kluster ke i dengan atribut kategorikal $A_1, A_2, A_3, \dots, A_m$. Untuk melakukan update titik pusat kluster di berikan estimasi dari W , Huang (Huang dan Ng,1999) menyediakan teorema berikut :

Teorema 1 : Nilai $F_c(W, Z)$ dalam persamaan 2 di minimalisasi jika dan hanya jika $z_{ij} = a_{jr} \in DOM(A_j)$ dimana

$$r = \arg \max_{1 \leq t \leq n_j} \sum_{i, x_{ij}=a_{jr}} w_{li}^{\alpha} \quad (3)$$

Dimana

$$\sum_{i, x_{ij}=a_{jr}} w_{li}^{\alpha} \geq \sum_{i, x_{ij}=a_{jt}} w_{li}^{\alpha}, 1 \leq t \leq n_j$$

untuk $1 \leq j \leq d$ dan $1 \leq t \leq k$

Untuk update matrik keanggotaan fuzzy, di berikan estimasi dari Z. Huang (Huang dan Ng, 1999) memberikan teorima berikut :

Teorema 2 : Jika $Z = \{z_1, z_2, \dots, z_k\}$ di tetapkan. Maka matrik keanggotaan fuzzy W yang menimalisasi nilai $F_c(W, Z)$ yang di definisikan di persamaan 2 dijabarkan berikut :

$$w_{li} = \begin{cases} 1 & \text{jika } x_i = z_l; \\ 0 & \text{jika } x_i \neq z_l, h \neq l; \\ \frac{1}{\sum_{h=1}^k \left[\frac{d(x_i, z_h)}{d(x_i, z_l)} \right]^{\frac{1}{\alpha-1}}} & \text{jika yang lainnya, } 1 \leq i \leq k, l \leq i \leq n. \end{cases} \quad (4)$$

Berdasar dari dua teorema di atas, algoritma Fuzzy K-modes dapat diimplementasikan secara rekursif. Dengan asumsi bahwa r adalah jumlah iterasi maksimal, berikut adalah langkah langkah Metode Fuzzy K-Modes (Gan, Yang dan Wu, 2009):

1. Inisialisasi : pilih inisial pusat klaster (Z_0) awal secara random.
2. Tentukan W_0 dimana fungsi biaya $F(W_0, Z_0)$ di minimalisasi.
3. For $t = 1$ to r do
4. Tentukan Z_1 di mana fungsi biaya $F(W_0, Z_1)$ di minimalisasi;
5. if $F(W_0, Z_1) = F(W_0, Z_0)$ then
6. Berhenti;
7. Else
8. Tentukan tentukan W_1 diaman fungsi biaya di minimalisasi
9. if $F(W_1, Z_1) = F(W_0, Z_1)$ then
10. Berhenti;
11. Else
12. $W_0 \leftarrow W_1$
13. End if
14. End if
15. End for

2.2 Artificial Bee Colony

Metode ABC pertama kali dikenalkan oleh Karaboga (karaboga, 2005). Terinspirasi dari perilaku alami lebah madu dalam proses mendapatkan sumber makanan dengan mengetahui kualitas dan lokasi lebah madu.

Dalam model metode ABC ini, Ada tiga jenis artificial bee dalam koloni lebah yang

digunakan yaitu : lebah pekerja (*employed bee*), lebah penunggu (*onlooker*), dan lebah scout. Setengah jumlah koloni terdiri atas lebah pekerja dan setengahnya adalah lebah penunggu. Jumlah lebah pekerja sama dengan jumlah sumber makanan di sekitar sarang karena dalam algoritma ini diasumsikan adanya satu lebah pekerja untuk satu sumber makanan. Lebah pekerja yang sudah meninggalkan sumber makanannya berubah menjadi lebah scout.

Sama seperti algoritma berbasis intelligent swarm yang lain, metode ABC adalah sebuah proses iteratif. Pendekatan dimulai dengan sebuah populasi solusi (atau sumber makanan) yang di hasilkan secara acak, diikuti langkah-langkah berikutnya yang diulang sampai kriteria penghentian terpenuhi. Langkah-langkah utama dari metode ABC diberikan di bawah ini.

1. Inisialisasi Populasi
2. Tempatkan lebah pekerja ke sumber makanan
3. Tempatkan lebah penunggu pada sumber makanan berdasarkan jumlah nektarnya
4. Kirim lebah scout ke daerah pencarian untuk menemukan sumber makanan baru
5. Mengingat sumber makanan terbaik yang telah ditemukan
6. Jika persyaratan penghentian tidak terpenuhi, balik ke langkah 2.

3. Skenario Uji Coba

3.1 dataset

Dataset yang digunakan dalam penelitian ini terdiri dari dataset Soybean Disease, Breast Cancer, dan Congressional Voting Record (Voting). Dataset tersebut diunduh dari UCI Machine Learning Repository (Frank and Asuncion, 2010). Semua dataset tersebut memiliki atribut kategorikal. Deskripsi mengenai jumlah atribut, kelas, dan data dalam penelitian ini dapat dilihat di tabel 1.

3.2 Parameter Uji coba

Uji coba dilakukan untuk membandingkan pengaruh penggunaan ABC untuk optimasi Fuzzy K-Modes. Oleh karena itu akan dibandingkan metode dasar dan metode yang diusulkan dalam beberapa parameter uji yaitu : *Objective Function*, *Precision*, *Recall*, *F-Measure*, *Accuracy*, dan waktu komputasi. *Precision* adalah tingkat relevansi data yang di hasilkan sistem. *Precision* dihasilkan dari

Tabel 1 Deskripsi dataset

Dataset	Jumlah atribut	Jumlah kelas	Jumlah Data
Soybean	35	4	47
Breast Cancer	9	2	286
Voting	16	2	435

Tabel 2 rata rata hasil uji coba dengan dataset soybean

Eksponen Fuzzy	Ukuran	Objective Function		F-Measure		Accuracy		Waktu	
		FKMO	ABC-FKMO	FKMO	ABC-FKMO	FKMO	ABC-FKMO	FKMO	ABC-FKMO
1.1	Rata	238,62	214,16	0,940	0,990	0,894	0,981	0,19	35,55
	stdev	34,72	0,29	0,063	0,003	0,112	0,007	0,15	4,36
1.2	Rata	216,63	210,35	0,976	1,000	0,955	1,000	0,15	30,89
	stdev	13,69	0,00	0,044	0,000	0,079	0,000	0,15	5,49
1.3	Rata	211,43	201,26	0,966	0,997	0,938	0,994	0,12	35,49
	stdev	24,58	0,13	0,053	0,005	0,095	0,010	0,14	4,52

pembagian antara tingkat data hasil sistem yang relevan dengan seluruh hasil yang didapat. *Recall* adalah tingkat data yang dihasilkan sistem data yang relevan. *F-Measure* merupakan pengukuran yang menggabungkan precision dan recall. *Accuracy* adalah tingkat keakuratan data yang di hasilkan sistem. Secara umum, parameter uji tersebut dapat dihitung dengan menggunakan rumus berikut :

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$FMeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Metode dasar dan metode yang diusulkan diuji coba dengan ketiga dataset dengan parameter eksponen Fuzzy 1.1, 1.2, dan 1.3. Tiap metode dan tiap parameter di uji sebanyak 10 kali.

4. Hasil Uji Coba

Hasil uji coba dengan menggunakan dataset soybean disease dapat dilihat di Tabel 3, Gambaran perbandingan performa dengan dataset soybean dapat di lihat di grafik 1 dan 2.

Hasil uji coba dengan menggunakan dataset Breast cancer dapat dilihat di Tabel 3. Gambaran perbandingan performa dengan dataset Breast cancer dapat di lihat di grafik 3 dan 4.

Hasil uji coba dengan menggunakan dataset Voting dapat dilihat di Tabel 4. Gambaran perbandingan performa dengan dataset Breast cancer dapat di lihat di grafik 5 dan 6.

Grafik 1 Objective Function dataset Soybean

Grafik 2 Accuracy dan F-Measure dataset Soybean

Grafik 3 Objective Function dataset Breast Cancer

Tabel 3 rata rata hasil uji coba dengan dataset Breast Cancer

Eksponen Fuzzy	Ukuran	Objective Function		F-Measure		Accuracy		Waktu	
		FKMO	ABC-FKMO	FKMO	ABC-FKMO	FKMO	ABC-FKMO	FKMO	ABC-FKMO
1.1	Rata	1.084,93	1.056,48	0,651	0,723	0,553	0,608	0,26	23,97
	stdev	32,85	6,45	0,058	0,015	0,047	0,017	0,27	4,17
1.2	Rata	1.051,86	1.022,59	0,667	0,732	0,572	0,619	0,11	17,75
	stdev	39,65	0,00	0,081	0,000	0,075	0,000	0,19	6,27
1.3	Rata	980,46	973,97	0,671	0,732	0,572	0,619	0,15	17,61
	stdev	8,28	0,00	0,059	0,000	0,043	0,000	0,21	4,38

Tabel 4 rata rata hasil uji coba dengan dataset Voting

Eksponen Fuzzy	Ukuran	Objective Function		F-Measure		Accuracy		Waktu	
		FKMO	ABC-FKMO	FKMO	ABC-FKMO	FKMO	ABC-FKMO	FKMO	ABC-FKMO
1.1	Rata	1.442,28	1.439,33	0,843	0,840	0,865	0,867	0,11	155,64
	stdev	2,63	0,00	0,011	0,000	0,010	0,000	0,07	7,09
1.2	Rata	1419,77	1416,93	0,844	0,840	0,866	0,867	0,11	23,02
	stdev	2,19	0,00	0,011	0,000	0,011	0,000	0,06	3,23
1.3	Rata	1.386,79	1.384,21	0,851	0,840	0,871	0,867	0,17	13,95
	stdev	1,07	0,00	0,010	0,000	0,008	0,000	0,30	2,76

Grafik 4 Accuracy dan F-Measure dataset Breast Cancer

Grafik 5 Objective Function dataset voting

Grafik 6 Accuracy dan F-Measure dataset voting

4. Pembahasan Hasil

Hasil uji coba dengan menggunakan tiga dataset dan tiga parameter eksponen Fuzzy menunjukkan bahwa metode ABC-FKMO menghasilkan kinerja lebih baik di banding FKMO. Hal ini dapat di lihat dari hasil *Objective Function*, *Accuracy*, *F-Measure*. Dari segi *Objective Function*, ABC-FKMO lebih baik 2,73 %, dari segi *F-Measure* unggul 4,31 %, dan dari segi *Accuracy* unggul 5.16 % di banding FKMO.

Namun demikian waktu komputasi yang dibutuhkan untuk menjalankan ABC-FKMO lebih besar lama dari FKMO.

5. Kesimpulan

Dari hasil uji coba dan analisa dapat disimpulkan bahwa :

1. Metode ABC-FKMO telah berhasil mengoptimalkan posisi titik pusat kluster dengan mengarahkan hasil kluster menuju solusi global optimal. Hal ini dibuktikan dengan hasil penelitian yang menunjukkan nilai *Objective Function* selalu lebih kecil dari FKMO dengan selisih rata rata 2,73 %.
2. Dalam hal pengukuran *F-Measure* dan *Accuracy*, metode ABC-FKMO juga menunjukkan keunggulan-nya dibandingkan dengan metode FKMO. Hal ini dibuktikan dengan hasil penelitian yang menunjukkan bahwa *F-Measure* ABC-FKMO lebih baik 4.31 % dan *Accuracy* ABC-FKMO lebih baik 5.16 %.
3. Dari sisi waktu komputasi, Metode ABC-FKMO membutuhkan waktu yang lebih lama dibandingkan dengan metode FKMO.

5. Daftar Pustaka

- Andritsos, P., Tsaparas, P., Miller, R.J., dan Sevcik, K.C. (2004), LIMBO: Scalable Clustering of Categorical Data, *Extending Database Technology - EDBT*, pp. 123-146
- Frank, A. dan Asuncion, A. (2010), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- Funderlic, R.E., Chu, M.T., Orlowski, N., Schlör, D., Blevins, J., Canas, D., (2004), Convergence and Other Aspects of The K-Modes Algorithm for clustering Categorical data, N.C. State University Departmen of Computer Science, March 25.
- Gan, G., Yang, Z., dan Wu, J., (2005), A Genetic k-Modes Algorithm for Clustering Categorical Data, *Lecture Notes in Artificial*

- intelligence* 3584, pp.195-202, Springer-Verlag Berlin Heidelberg.
- Gan, G., Ma, C., dan Wu, J., (2007), Data Clustering Theory, Algorithms, and Applications, Society for Industrial and Applied Mathematics, Philadelphia, PA.
- Gan, G., Wu, J., dan Yang, Z., (2009), A genetic fuzzy k-Modes algorithm for clustering categorical data, *Expert Systems with Applications*. 36:1615-1620.
- Guha, S., Rastogi, R, dan Shim, K., (2000), ROCK: A Robust Clustering Algorithm for Categorical Attributes, *Information Systems*, Vol 25, issue 5, pp 345-366.
- Han, J., dan Kamber, M., (2006), Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann Publishers.
- Huang, Z. (1998), Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining Knowledge Discovery*, 2, pp.283-304.
- Huang, Z. dan Ng, M.K. (1999), A fuzzy k-modes algorithm for clustering categorical data, *IEEE Transaction on fuzzy systems*, pp.446-452.
- Jain, A.K., Murty, M.N., dan Flynn, P.J., (1999), Data Clustering: A Review, *ACM Computing Surveys*, Vol. 31, No. 3, September
- Karaboga, D. (2005), *An Idea Based on Honey Bee Swarm for Numerical Optimization*, Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department.
- Karaboga, D. dan Basturk, B. (2007), “A Powerful and Efficient Algorithm for Numerical Function Optimization: artificial bee colony (ABC) algorithm”, *Journal of Global Optimization*, Vol. 39, hal. 459–471.
- Karaboga, D. dan Basturk, B. (2008), “On The Performance of Artificial Bee Colony (ABC) Algorithm”, *Applied Soft Computing*, Vol. 8, hal. 687–697.
- Karaboga, D. dan Akay, B. (2009), “A Comparative Study of Artificial Bee Colony Algorithm”, *Applied Mathematics and Computation*, Vol 214, hal. 108–132.
- Tan, P.N., Steinbach, M., Kumar, V., (2006), Introduction to data mining, 4th edition, Academic Press, San Diego.