ITS
Institut
Teknologi
Sepuluh Nopember

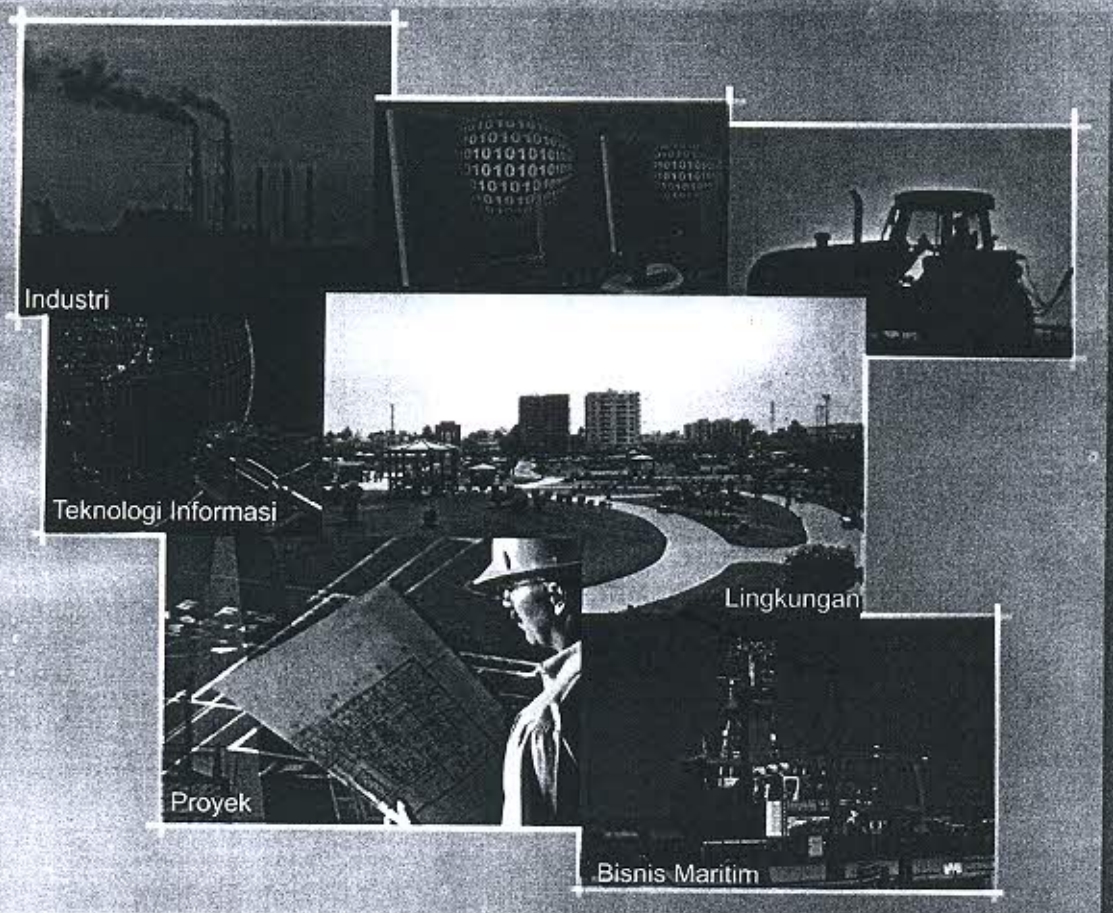
PROGRAM STUDI
MAGISTER MANAJEMEN TEKNOLOGI
PROGRAM PASCASARJANA

PROSIDING

SEMINAR NASIONAL
MANAJEMEN TEKNOLOGI XIV

EFFECTIVE RESOURCE MANAGEMENT

Surabaya, 23 Juli 2011



ISBN : 978-602-97491-3-7

Prosiding Seminar Nasional Manajemen Teknologi XIV
Program Studi MMT-ITS, Surabaya 23 Juli 2011

KLASTERISASI DATA IRIS MENGGUNAKAN METODE BERBASIS *ARTIFICIAL BEE COLONY* DAN *K-HARMONIC MEANS*

I Made Widiartha, Agus Zainal Arifin, Anny Yuniarti
Jurusan Teknik Informatika, Fakultas Teknologi Informasi, ITS
email : imadewidiartha@yahoo.com

ABSTRAK

Pengelompokan data ke dalam beberapa klaster dapat dilakukan dengan berbagai metode, salah satunya menggunakan K-Means Clustering (KM). KM memiliki kelemahan yaitu dari sisi sensitifitas hasil klaster pada inisialisasi awal titik pusat klaster dan adanya kemungkinan hasil klaster merupakan lokal optimal.

K-Harmonic Means (KHM) merupakan metode klasterisasi data yang menyempurnakan KM. Meskipun metode KHM dapat mengurangi masalah inisialisasi, namun dalam KHM masih terdapat kemungkinan terjadinya masalah lokal optimal. Salah satu cara untuk mengatasi permasalahan lokal optimal ini adalah dengan memanfaatkan suatu metode yang memiliki solusi global ke dalam metode KHM.

Metode Artificial Bee Colony (ABC) merupakan suatu metode swarm yang berbasis pada perilaku mencari makan (*foraging behavior*) dari koloni lebah madu yang telah terbukti memiliki solusi global. Dalam penelitian ini diusulkan sebuah metode baru untuk klasterisasi data yang berbasis pada metode ABC dan KHM (ABC-KHM). Kinerja metode ABC-KHM ini telah dibandingkan dengan metode ABC dan KHM dengan menggunakan dataset iris. Dari hasil penelitian didapatkan hasil dimana metode ABC-KHM ini telah berhasil mengoptimalkan posisi titik pusat klaster yang mengarahkan hasil klaster menuju suatu solusi global.

***Kata kunci**: K-Means Clustering, K-Harmonic Clustering, Artificial Bee Colony, ABC-KHM.*

PENDAHULUAN

Klasterisasi data (*clustering*) adalah sebuah proses untuk mengelompokkan data kedalam beberapa klaster/kelompok sehingga data dalam satu klaster memiliki tingkat kemiripan yang maksimum dan data antar klaster memiliki kemiripan yang minimum [7]. Salah satu metode yang dapat digunakan dalam melakukan klasterisasi data adalah K-Means Clustering (KM). Metode KM ini memiliki kelemahan yaitu hasil klaster sangat sensitif dengan inisialisasi titik pusat awal dan sangat mudah terjebak pada lokal optimal [8].

Untuk mengatasi masalah yang terjadi pada inisialisasi titik pusat klaster, Zhang, Hsu, dan Dayal [10] mengusulkan sebuah metode baru yang diberi nama K-Harmonic Means (KHM) yang kemudian dimodifikasi oleh Hammerly dan Elkan [2]. Meskipun KHM dapat mengurangi masalah inisialisasi, namun dalam KHM masih terdapat kemungkinan terjadinya masalah lokal optimal [8]. Untuk mengatasi permasalahan lokal optimal ini maka diperlukan suatu metode yang memiliki kemampuan untuk menghindari kemungkinan adanya konvergensi terhadap lokal optimal.

Artificial Bee Colony (ABC) merupakan suatu metode yang mengadopsi perilaku mencari makan (foraging behavior) dari koloni lebah madu. Pada metode ini terdapat tiga jenis lebah yang dipakai yaitu .lebah pekerja/*employed bee*, lebah penunggu/*onlooker bee*, dan lebah pengintai/*scout* [4]. Metode ABC telah terbukti memiliki kemampuan untuk menangani permasalahan lokal optimal dan memiliki kualitas yang lebih baik atau setara jika dibandingkan dengan metode sejenis lainnya seperti Algoritma Genetika, Particel Swarm Optimization, Differential Evolution, dan Evolution Stategies [6].

Dalam penelitian ini diusulkan sebuah metode baru untuk klasterisasi data yang berbasis pada metode ABC dan KHM (ABC-KHM). Perilaku lebah pada metode ABC digunakan untuk membantu KHM untuk dapat keluar dari lokal optimal sehingga metode ABC-KHM ini diharapkan mampu mengoptimalkan posisi titik pusat klaster yang mengarah pada solusi global optimal. Metode yang diusulkan ini diterapkan pada dataset iris.

K-HARMONIC MEANS CLUSTERING

KHM merupakan suatu metode klasterisasi data dimana klaster-klaster dibentuk dengan peyempurnaan secara iteratif berdasarkan letak titik pusat dari masing-masing klaster. Pada KHM, nilai fungsi tujuan dihasilkan dengan mencari total rata-rata harmonik dari seluruh titik data untuk jarak antara masing-masing titik data ke seluruh titik pusat klaster yang ada [9].

Adapun langkah-langkah Metode KHM adalah sebagai berikut [8] :

1. Inisialisasi posisi titik pusat klaster awal secara random
2. Hitung nilai fungsi tujuan dengan persamaan 1, dimana p adalah input parameter. Nilai p biasanya ≥ 2.

$$KHM(X,C) = \sum_{i=1}^{N} \frac{K}{\sum_{l=1}^{K} \frac{1}{\| x_i - c_l \|^p}} \tag{1}$$

3. Untuk setiap data xi, hitung nilai keanggotaan $m(c_l|x_i)$ untuk setiap titik pusat klaster c_l berdasarkan persamaan :

$$m(c_l \mid x_i) = \frac{\| x_i - c_l \|^{-p-2}}{\sum_{l=1}^{k} \| x_i - c_l \|^{-p-2}} \tag{2}$$

4. Untuk setiap data x_i, hitung nilai bobot $w(x_i)$ berdasarkan persamaan

$$w(x_i) = \frac{\sum_{l=1}^{K} \| x_i - c_l \|^{-p-2}}{\left(\sum_{l=1}^{K} \| x_i - c_l \|^{-p} \right)^2} \tag{3}$$

5. Untuk setiap titik pusat c_j, ulang kembali perhitungan untuk posisi titik pusat klaster dari semua data berdasarkan nilai keanggotaan dan bobot yang dimiliki tiap data.

$$c_l = \frac{\sum_{i=1}^{N} m(c_l \mid x_i).w(x_i).x_i}{\sum_{i=1}^{N} m(c_l \mid x_i).w(x_i)} \tag{4}$$

6. Ulangi langkah 2 sampai 5 sampai mendapatkan nilai fungsi tujuan yang tidak terdapat perubahan yang signifikan.

ISBN : 978-602-97491-3-7

7. Tetapkan keanggotaan data x_i pada suatu klaster dengan titik pusat klaster c_j sesuai dengan nilai keanggotaan x_i terhadap c_j.

ARTIFICIAL BEE COLONY

Metode Artificial Bee Colony (ABC) merupakan sebuah metode yang diperkenalkan oleh Karaboga pada tahun 2005. Koloni lebah tiruan terdiri dari tiga kelompok yaitu lebah pekerja (*employed bee*), lebah penunggu (*onlooker*) dan lebah *scout* (penjelajah). Metode ABC ini dapat digambarkan seperti pada Gambar 1.

Langkah pertama pada metode ABC ini adalah pengiriman lebah pekerja (yang berstatus scout) pada daerah pencarian untuk menghasilkan populasi awal yang didistribusikan secara random. Setelah inisialisasi, penentuan populasi dari posisi solusi berikutnya melalui siklus yang berulang, C = 1, 2, ... , MCN. Setelah semua lebah pekerja menyelesaikan proses pencarian, dilakukan penghitungan nilai fitness dari solusi yang dihasilkan (nilai nektar) dan lebah pekerja membagi informasi nektar dan informasi tentang posisi mereka dengan lebah penunggu di *dancing area*. Nilai fitness dapat dicari dengan menggunakan persamaan 5.

$$fit_i = \frac{1}{1+f_i} \tag{5}$$

Variabel f_i merupakan nilai *cost function* dari solusi i. Lebah penunggu mengevaluasi informasi yang diambil dari semua lebah pekerja dan memilih sumber makanan dengan probabilitas yang sesuai jumlah nektarnya. Seperti kasus lebah pekerja, lebah penunggu juga menghasilkan modifikasi pada posisi sumber makanan (solusi) dalam memorinya dan memeriksa jumlah nektar dari kandidat sumber makanan (solusi) yang baru. Jika nilai nektar lebih tinggi dari sebelumnya, lebah akan mengingat posisi yang baru tersebut dan melupakan posisi yang lama.

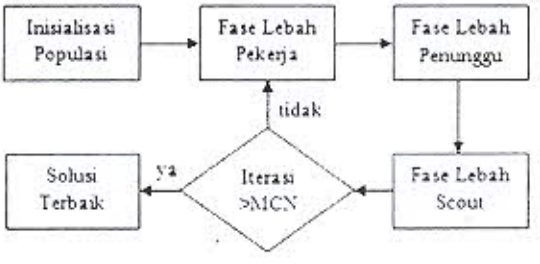
Lebah penunggu memilih sumber makanan berdasarkan pada nilai probabilitas p_i dengan menggunakan metode *roulette wheel selection* [5]. Nilai p_i ini dihitung melalui persamaan 6.

$$p_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \tag{6}$$

Dalam menghasilkan kandidat posisi makanan baru, ABC menggunakan persamaan 7.

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \tag{7}$$

Nilai $k \in \{1, 2, ..., SN\}$ dengan $j \in \{1, 2, .., D\}$ adalah indeks yang dipilih secara random. Meskipun k ditentukan secara random, namun k



Gambar 1. Metode ABC

ISBN : 978-602-97491-3-7

harus berbeda dari i. ϕ_{ij} adalah sebuah bilangan random antara [-1,1], yang mengontrol produksi posisi sumber makanan tetangga di sekitar x_{ij}

Sumber makanan yang ditinggalkan oleh lebah pekerja, digantikan dengan sumber makanan baru oleh lebah scout. Dalam metode ABC, jika sebuah posisi sumber makanan tidak dapat ditingkatkan lebih lanjut melalui sejumlah siklus (*cycle*) yang telah ditetapkan, yang disebut dengan limit, maka sumber makanan tersebut diasumsikan untuk ditinggalkan. Misal sumber makanan yang ditinggalkan adalah x_i dan $j \in \{1, 2, ..., D\}$, maka lebah scout akan mencari sumber makanan baru untuk diganti dengan x_i. Operasi ini dilakukan dengan menggunakan persarnaan 8.

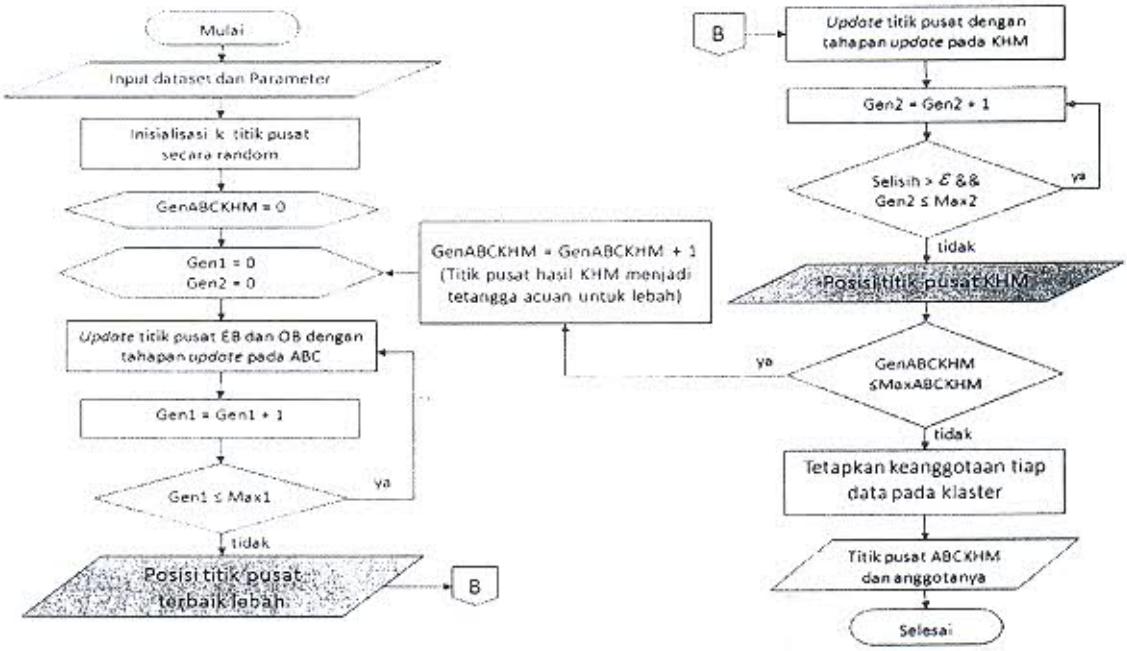
$$x_i^j = x_{\min}^j + \text{rand}[0,1](x_{\max}^j - x_{\min}^j) \quad (8)$$

Setelah masing-masing kandidat posisi sumber makanan v_{ij} diproduksi dan dievaluasi oleh lebah artificial, nilai fitnessnya dibandingkan dengan x_{ij}. Jika sumber makanan baru mempunyai nektar yang sama atau lebih baik daripada sumber yang lama, maka sumber yang lama tersebut akan digantikan dengan yang baru dalam memori, jika tidak maka yang lama dipertahankan.

METODE USULAN

Metode ABC-KHM

Metode yang diusulkan dalam melakukan proses klasterisasi data dalam penelitian ini adalah Metode ABC-KHM. Metode ini dihasilkan melalui hibridasi antara metode ABC dan metode KHM. Dalam metode ABC-KHM, hasil klaster diperoleh dengan memanfaatkan hubungan timbal balik antara kedua metode yaitu ABC dan KHM. Posisi titik pusat yang dihasilkan pada fase lebah akan dioptimalkan dengan fase *update* yang terdapat pada metode KHM. Hasil titik pusat yang diperoleh dari fase KHM, akan dimanfaatkan oleh fase lebah sebagai tetangga acuan dari lebah pekerja untuk melakukan ekplorasi dalam ruang pencarian.



Gambar 2. Metode ABC-KHM

ISBN : 978-602-97491-3-7

Dalam implementasi metode ABC-KHM terdapat beberapa variabel yang digunakan untuk membatasi setiap fase yang ada. Parameter Max1 digunakan untuk membatasi jumlah iterasi pada tahapan pencarian titik pusat oleh para lebah. Hasil titik pusat dari tahapan ini akan menjadi titik pusat awal pada tahapan selanjutnya yaitu tahapan KHM. Tahapan iterasi pada KHM ini dibatasi oleh dua hal yaitu *threshold* selisih posisi titik pusat antar iterasi (ε) dan Max2. Fase lebah dan KHM ini akan dilakukan terus sampai iterasi melampaui batas iterasi maksimum yaitu MaxABCKHM.

Data

Dataset yang digunakan dalam penelitian ini adalah dataset Iris yang diambil dari *UCI Machine Learning Repository* (ftp://ftp.ics.uci.edu./pub/machine-learning-databases/). Dataset iris ini terdiri dari empat fitur, dan tiga kelas. Jumlah total data iris ini sebanyak 150 data. Dalam penelitian ini, 80% data akan digunakan sebagai data training dan sisanya digunakan sebagai data testing. Data training ini digunakan untuk melihat performa dari ketiga metode dalam melakukan klasterisasi data. Penilaian performa ini dilihat dari tiga sudut pandang yaitu nilai fungsi tujuan KHM(X,C), F-Measure, dan running time. Data testing hanya digunakan untuk melihat korelasi secara eksternal (kelas label) yaitu bagaimana hasil klasifikasi data testing dengan memanfaatkan hasil titik pusat klaster dengan menggunakan data training.

HASIL

Dalam melakukan uji coba pada penelitian ini, nilai parameter yang digunakan untuk metode ABC mengacu pada nilai parameter yang digunakan oleh Zhang. Parameter tersebut antara lain Limit bernilai 10 dan jumlah MCN yang bernilai 2000 [10]. Untuk metode ABC-KHM penentuan, Limit, Max1, Max2 dan MaxABCKHM ditentukan dengan melakukan uji coba nilai-nilai parameter ini. Dari hasil uji coba yang telah dilakukan, didapatkan bahwa hasil terbaik diperoleh dengan menggunakan Max1=20, Limit=3, Max2=10, dan MaxABCKHM=20.

Untuk mengetahui performa masing-masing metode maka pada penelitian ini digunakan tiga tolak ukur yaitu nilai fungsi tujuan KHM(X,C), F-measure, dan running time. Uji coba pada penelitian ini dilakukan melalui beberapa sekenario untuk menguji performa dari metode-metode yang ada. Sekenario ini dibuat dengan menggunakan fungsi tujuan yang berbeda-beda. Perbedaan fungsi tujuan ini terletak pada parameter p. Pada penelitian ini, terdapat dua buah sekenario nilai p yaitu $p = 2$, dan $p = 4$.

Dari sisi penilaian eksternal (kelas label) pada penelitian ini digunakan penilaian F-measure. Nilai F-measure didapat dari persamaan 10 [1].

$$F(i,j) = \frac{(b^2+1).(p(i,j).r(i,j))}{b^2.p(i,j)+r(i,j)} \quad (9)$$

$p(i,j) = n_{ij}/n_j$ dan $r(i,j) = n_{ij}/n_i$ dimana n_i adalah jumlah data dari kelas i yang diharapkan sebagai hasil query, n_j adalah jumlah data dari klaster j yang dihasilkan oleh query, dan n_{ij} adalah jumlah elemen dari kelas i yang masuk di klaster j. untuk mendapatkan pembobotan yang seimbang antara *precision* dan *recall* maka nilai $b = 1$ digunakan dalam menghitung nilai F-measure [3].

Untuk mendapatkan kesimpulan akhir hasil klasterisasi menggunakan metode-metode yang ada, maka uji coba klasterisasi dilakukan sebanyak 10 kali untuk tiap-tiap sekenario yang dibuat. Kesimpulan kinerja dari metode akan didapatkan melalui nilai rata-rata (mean) dan standar deviasi dari 10 percobaan tersebut.

ISBN : 978-602-97491-3-7