Optimasi Fuzzy C-Means Clustering Untuk Data Besar dengan Pemrograman R

Budi Arif Dermawan[1], Taufik Djatna[2]

[1]Jl. Pangkal Perjuangan Kp. Bubulak Paracis Tanjungpura Karawang
[2]Jalan Raya Dramaga, Bogor, Jawa Barat 16680
Email: [1]budi.arif@staff.unsika.ac.id, [2]taufik.djatna@gmail.com

Abstrak. Penangkapan Abalone secara terus menerus untuk tujuan konsumsi dapat menyebabkan kepunahan dari spesies ini tanpa diiringi dengan pembudidayaan kembali. Maka dari itu dinilai penting untuk mengelompokan Abalone ke dalam kategori muda/benih, dewasa, dan indukan untuk tujuan konservasi. Analisis *cluster* diperlukan agar dapat mengelompokan data dengan baik. Analisis *cluster* merupakan sebuah alat yang bertujuan untuk memisahkan dataset kedalam subset menurut persamaan dan *dissimilarities* data. Penelitian ini menggunakan Relief untuk melakukan reduksi terhadap variabel dengan fungsi *attrEval* dan FCM yang bertujuan untuk mengelompokan data kedalam beberapa *cluster* dengan fungsi *cmeans*. Hasil dari penelitian ini menunjukkan *cluster* yang terbentuk dengan menggunakan algoritma Relief dan FCM menunjukkan hasil *cluster* yang lebih optimal dibandingkan hanya menggunakan algoritma K-Means. *Cluster* pada data Abalone dapat memberikan pengetahuan kepada nelayan pencari Abalone untuk memperhatikan keberlangsungan siklus kehidupan untuk spesies ini dengan tidak menangkap Abalone secara sembarangan.

Kata kunci: *abalone, cluster, fuzzy c-means ,reduksi, relief.*

1 Pendahuluan

Clustering merupakan sebuah metode pengelompokan suatu obyek kedalam sejumlah kelompok (*cluster*) yang sesuai. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu *cluster* dan meminimumkan kesamaan antar anggota *cluster* yang berbeda[5]. Analisis *cluster* berfungsi sebagai pemisah obyek kedalam beberapa kelompok yang memiliki perbedaan karakteristik antar kelompok. Terdapat beberapa metode yang digunakan untuk pengelompokan[8], diantaranya k-means, possibilistic c-means (PCM) dan fuzzy c-means (FCM).

K-Means merupakan teknik pengelompokan *hard partition* yang efisien dalam mengelompokan data besar, namun terbatas hanya pada data numerik[2]. PCM merupakan metode pengelompokan yang kuat terhadap *noise*[4], namun mengorbankan stabilitas algoritma dan terlalu sensitif terhadap inisialisasi

Diterima 16 November 2016, Direvisi 12 Desember 2016, Diterima untuk publikasi 16 Desember 2016

cluster[7]. Sedangkan Fuzzy C-Means merupakan suatu teknik pengelompokan yang mana keberadaan tiap titik data dalam suatu kelompok (*cluster*) ditentukan oleh derajat keanggotaan. Penempatan posisi data pada *cluster* dilakukan dengan perbaikan penentuan pusat *cluster* awal dan nilai keanggotaan secara berulang[6]. Algoritma Fuzzy C-Means merupakan salah satu teknik *clustering* yang populer karena efisien dan mudah diimplementasikan.

Penelitian ini akan melakukan analisis terhadap data yang besar yaitu data Abalone dengan 4.117 *instance* menggunakan algoritma Fuzzy C-Means yang telah dioptimasi dengan menggunakan seleksi atribut agar hasil *cluster* lebih optimal. Penelitian ini bertujuan untuk melakukan pengelompokan pada dataset Abalone menjadi tiga kelompok (muda/benih, dewasa, induk) dengan menggunakan algoritma relief dan FCM. Pengelompokan tersebut berguna untuk melindungi spesies Abalone dari kepunahan karena terus menerus dipanen tanpa ada tindakan pembibitan kembali. Data yang telah dikelompokan dapat menjadi dasar pengetahuan untuk menentukan kebijakan guna melakukan konservasi.

2 Penelitian Sebelumnya

Fuzzy partition didasarkan pada gagasan dari keanggotaan parsial dari masing-masing pola dalam sebuah klaster tertentu. Hal tersebut memberikan fleksibilitas untuk menyatakan bahwa titik data memiliki lebih dari satu klaster pada waktu yang sama dan derajat keanggotaanya jauh lebih halus dari model data. Derajat keanggotaan dapat mengungkapkan ambiguitas titik data yang memiliki sebuah klaster[1].

Fuzzy C-Means Clustering merupakan metode pengelompokkan yang paling banyak digunakan[11] dan paling terkenal[9]. Dalam hal ambiguitas, FCM dikenal memiliki karakteristik yang sangat kuat serta dapat menyimpan informasi lebih banyak dibandingkan metode Hard C-Means[10]. Namun algoritma FCM dalam pencarian klaster yang optimasl didasarkan pada fungsi obyektif, sehingga mudah terjebak pada kondisi dimana nilai yang dihasilkan bukan nilai terendah dari himpunan solusi atau disebut *local minimum*[3].

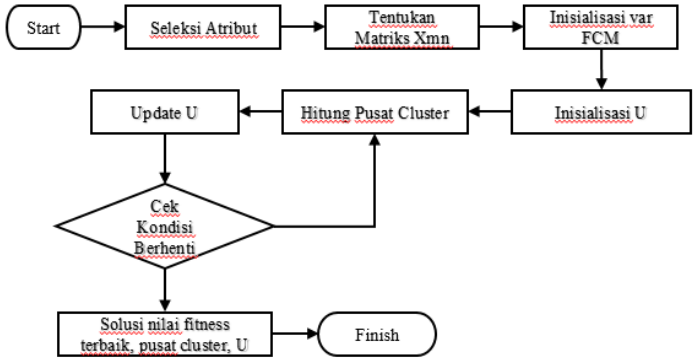
3 Metodologi Penelitian

Penelitian ini menggunakan dataset Abalone. Dataset tersebut kemudian dikelompokan menggunakan algoritma Fuzzy C-Means yang sebelumnya dilakukan proses seleksi atribut agar proses komputasi lebih optimal.

3.1 Dataset

Data yang digunakan merupakan dataset Abalone yang didapat dari situs UCI Machine Learning. Dataset terdiri dari 4.117 *instance* dan 9 variabel. Jumlah variabel tersebut kemudian direduksi untuk mendapatkan waktu komputasi yang lebih optimal. Terdapat beberapa metode untuk reduksi data, diantaranya PCA dan Relief, namun penelitian ini menggunakan metode Relief yang lebih sederhana. Variabel yang tersisa diambil dari 3 variabel dengan nilai tertinggi. Abalone merupakan spesies yang berada di laut yang masih terbilang sulit untuk dibudidayakan. Jika tidak ada perhatian khusus pada spesies ini, maka bukan tidak mungkin status spesies ini akan berubah menjadi "terancam punah".

3.2 Metode



Gambar 1. Skema alur kerja

Skema alur kerja yang digunakan pada penelitian ini menggunakan algoritma Fuzzy C-Means yang dioptimasi dengan menambahkan proses seleksi atribut untuk mereduksi variabel sesuai dengan gambar 1.

3.2.1 Seleksi Atribut

Metode yang digunakan untuk melakukan seleksi atribut pada penelitian ini yaitu menggunakan algoritma Relief yang dijelaskan pada tabel 1. Berikut ini merupakan algoritma dari Relief:

Tabel 1. algoritma relief

(1) **Initialization** $D=\{(x_n,y_n)\}_{n=1}^{N}$, set $w_i=0.1 \le i \le I$, number of iteration T;
(2) **for** t=1:T
(3) Randomly select a pattern x from D;
(4) Find the nearest hit NH[7] and miss NM[7] of x;
(5) **for** i = 1:I

(6) Compute :

$$w_i=w_i+\left|x^{(i)}-NM^{(i)}(x)x\right|-\left|x^{(i)}-NH^{(i)}(x)\right|; \quad (1)$$

(7) **end**

(8) **end**

3.2.2 Pengelompokan Data

Algoritma yang digunakan untuk mengelompokan data pada penelitian ini yaitu Fuzzy C-Means yang dijelaskan pada tabel 2. Berikut ini merupakan algoritma dari Fuzzy C-Means:

Tabel 2. algoritma fuzzy c-means

Step 1. Fix $C\in(2,N)$,$(m>0)$ and $(\varepsilon>0)$.

Step 2. Give initials randomly $\mu_{ij}^{(0)}\sim U(0,1)$ and let t=1.

Step 3. Compute cluster centers (v_j) by using equation (2).

$$v_j=\frac{\sum_{i=1}^{N}\mu_{ij}^{m}x_i}{\sum_{i=1}^{N}\mu_{ij}^{m}},(j=1,2,\ldots,c) \quad (2)$$

Step 4. Update μ_{ij} with v_j by using equation (3).

$$\mu_{ij}=\left(\sum_{k=1}^{c}\left(\frac{\|x_i-v_j\|}{\|x_i-v_k\|}\right)^{\frac{2}{m-1}}\right)^{-1}, \quad (3)$$

$$(i=1,2,\ldots,N;j=1,2,\ldots,C)$$

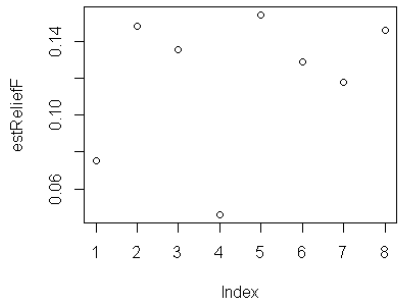
Step 5. Compute $\left\|\mu^{(t)}-\mu^{t-1}\right\|$

If $\left\|\mu^{(t)}-\mu^{t-1}\right\|<\varepsilon$, Stop

Else t=t+1 and return to step 3.

4 Hasil dan Pembahasan

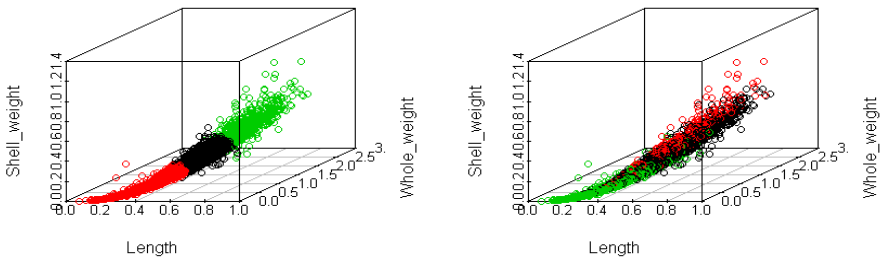
Proses awal yang dilakukan pada penelitian ini yaitu mengunduh data dari situs UCI Machine Learning melalui aplikasi R. Kemudian setelah dataset didapatkan, dilakukan langkah seleksi atribut yang bertujuan untuk mereduksi variabel yang tidak berpengaruh nyata terhadap kelas label.



Gambar 2. Hasil seleksi atribut

Gambar 2 menunjukan pemetaan masing-masing variabel terhadap nilai yang dihasilkan oleh algoritma Relief. Terlihat didalam grafik, bahwa 3 index dengan nilai terbaik yaitu index 2, 5 dan 8. Ketiga index itu yang akan digunakan sebagai variabel untuk melakukan proses selanjutnya. Ketiga variabel itu adalah *length*, *whole weight* dan *shell weight* yang mempunyai pengaruh besar terhadap penentuan kelas label.

Setelah melalui tahap seleksi atribut, tahap selanjutnya adalah mengelompokan spesies Abalone kedalam 3 kelompok (*cluster*). Pengelompokan ini bertujuan untuk mengikis rantai pengambilan Abalone tanpa memperdulikan pelestariannya. Pengelompokan dilakukan dengan membandingkan penggunaan FCM pada aplikasi R dengan paket kmeans dan cmeans.



Gambar 3. Kmeans dan cmeans

Gambar 3 menampilkan perbandingan yang dihasilkan oleh proses *cluster* menggunakan paket kmeans dan cmeans. Proses pengelompokan dengan kmeans menghasilkan 3 kelompok dengan rincian 1.721 *instance* masuk kedalam kelompok benih, 1.662 kedalam kelompok dewasa dan 794 kedalam

kelompok induk. Sedangkan dengan cmeans menghasilkan 3 kelompok dengan rincian 2.280 *instane* masuk kedalam kelompok benih, 490 kedalam kelompok dewasa dan 1.407 kedalam kelompok induk. Gambar 3 menjelaskan perbandingan data yang *overlap* dapat diatasi dengan penggunaan FCM, tidak demikian dengan kmeans.

5 Kesimpulan

Pengelompokan data dengan menggunakan seleksi atribut dan algoritma FCM memberikan hasil yang lebih baik dibandingkan dengan K-Means karena memiliki ambiguitas titik data. Algoritma ini dapat menangani sebaran data yang mengalami *overlap* yang tidak bisa ditangani oleh K-Means. Dataset yang telah dikelompokan dapat dilakukan penambangan pengetahuan yang menjadi bekal untuk para nelayan pencari Abalon agar dapat menangkap serta memperhatikan sisi kelestarian dari spesies Abalone. Dengan adanya pengetahuan mengenai populasi Abalone, pemerintah setempat dapat membuat kebijakan untuk memanen dengan usia spesies tertentu guna menjaga siklus perkembangbiakan spesies ini.

Penelitian selanjutnya dapat melakukan penggunaan algoritma optimasi untuk mengoptimalkan fungsi obyektif dan menambang informasi dari pakar dibidang gizi untuk mengetahui informasi gizi dari masing-masing kelas usia dari Abalone serta dapat menangkap ciri dari Abalone dengan menggunakan teknologi *image processing*.

6 Referensi

[1] Azar, A.T., El-Said, S.A. & Hassanien, A.E., Fuzzy and hard clustering analysis for thyroid disease, Computer methods and programs in biomedicine, 111, pp. 1-16, 2013.

[2] Bai, L., Liang, J. & Dang, C., An Initialization Method to Simultaneously Find Initial Cluster Centers and the Number of Clusters for Clustering Categorical Data, Knowledge-Based System, 24, pp. 785 - 795, 2011.

[3] Dong, H., Dong, Y., Zhou, C., Yin, G. & Hou, W., A fuzzy clustering algorithm based on evolutionary programming, Expert Systems with Applications, 36, pp. 11792-11800, 2009.

[4] Hamasuna, Y., Endo, Y. & Miyamoto, S., On Tolerant Fuzzy C-Means Clustering and Tolerant Possibilistic Clustering, Soft Computing, 14, pp. 487 - 494, 2009.

[5] Han, J., Kamber, M. & Pei, J., Data mining: concepts and techniques, Elsevier, 2011.

[6] James, C.B., Robert, E. & William, F., FCM : The Fuzzy C-Means Clustering Algorithm, Computer & Geosciences, 10, pp. 191 - 203, 1984.

[7] Ji, Z., Sun, Q. & Xia, D., Computerized Medical Imaging and Graphics a Modified Possibilistic Fuzzy C-Means Clustering Algorithm for Bias Field Estimation and Segmentation of Brain MR Image, Computerized Medical Imaging and Graphics, 35, pp. 383 - 397, 2011.
[8] Oliveira, J.V.D. & Pedrycz, W., Advances in Fuzzy Clustering and its Applications, John Wiley \& Sons, Ltd, 2007.
[9] Wu, K.-L., Analysis of parameter selections for fuzzy c-means, Pattern Recognition, 45, pp. 407-415, 2012.
[10] Zhang, Y., Huang, D., Ji, M. & Xie, F., Image segmentation using PSO and PCM with Mahalanobis distance, Expert Systems with Applications, 38, pp. 9036-9040, 2011.
[11] Zhao, F., Jiao, L. & Liu, H., Kernel generalized fuzzy c-means clustering with spatial information for image segmentation, Digital Signal Processing, 23, pp. 184-199, 2013.