ANALISIS PERBANDINGAN ALGORITMA K-*MEANS CLUSTERING* DAN *EXPECTATION-MAXIMATION* (EM) UNTUK KLASIFIKASI BUTIR BERAS

A COMPARATIVE ANALYSIS OF K-MEANS CLUSTERING ALGORITHM AND EXPECTATION-MAXIMATION (EM) FOR CLASSIFICATION OF GRAIN RICE

Lina Septiana[1], Nani Djohan[2]

Fakultas Teknik dan Ilmu Komputer Program Studi Teknik Elektro
Universitas Kristen Krida Wacana – Jakarta
[1]linaseptiana@ukrida.ac.id, [2]nani.djohan@ukrida.ac.id

Abstrak

Penelitian ini mengamati perbandingan klasifikasi butir beras menggunakan algoritma K-*means clustering* dan *Expectation-Maximation*. Jenis beras yang digunakan dalam penelitian ini adalah butir beras Taiwan dan butir beras Thailand. Hal ini dilakukan menggunakan pengenalan pola *unsupervised* dengan mengidentifikasi panjang dan lebar dari kedua jenis butir beras tersebut, yaitu beras Taiwan CNS *grade* 1 dan beras Thailand Thai Hom Mali *grade* A. Hasil analisis menunjukkan kedua metode tersebut memberikan hasil yang hampir sama dalam pengelompokan kedua jenis beras tersebut.

Kata kunci: pengelompokan K-*means*, pengenalan pola *unsupervised*

Abstract

This study observed the comparison of rice grain classification using K-means clustering algorithm and Expectation-Maximation Algorithm. This research uses Taiwanese rice grain and Thai rice grain. The study was performed using the unsupervised pattern recognition by identifying the length and width of two different rice grains, which are CNS Taiwan rice Grade 1 (Taiwanese rice) and Thai Hom Mali Rice Grade A (Thai rice). The analysis result shows these two methods have a similar result in clustering the two kinds of rice grains.

***Key words**: k-means clustering, unsupervised pattern recognition.*

Tanggal Terima Naskah : 26 Februari 2015
Tanggal Persetujuan Naskah : 16 Maret 2015

1. PENDAHULUAN

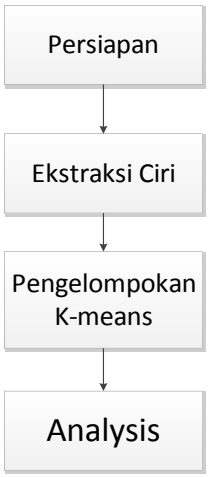
Klasifikasi *unsupervised*, yang disebut juga pengelompokan, sudah menjadi persoalan klasik dalam bidang pengenalan pola. Pengelompokan juga telah digunakan untuk berbagai tugas dalam analisis citra penginderaan jauh, seperti *pre-processing*, segmentasi, ekstraksi ciri, pengurangan dimensi, visualisasi data, dan klasifikasi akhir. Salah satu yang paling banyak digunakan dari algoritma pengelompokan adalah

pengelompoan k-*means*. K-*means* adalah metode analisis kelompok yang bertujuan untuk partisi n pengamatan ke dalam kelompok *k* di mana setiap observasi milik cluster dengan *mean* terdekat.

Penelitian ini mengkaji penggunaan algoritma pengelompokan k-*means* dan *Expectation Maximation* untuk mengelompokkan butir beras Taiwan jenis CNS *grade* A dan butir beras Thailand jenis Tai Hom Mali *grade* A. Adapun penelitian ini bertujuan untuk mengetahui unjuk kerja kedua algoritma *clustering* tersebut dalam mengelompokkan kedua jenis beras unggulan Taiwan dan Thailand.

2. MATERIAL DAN METODE

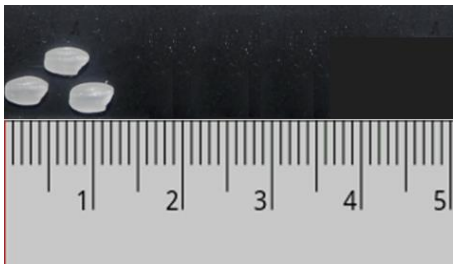
Langkah-langkah dari metode yang digunakan dalam penelitian ini ditunjukkan pada diagram alir berikut.



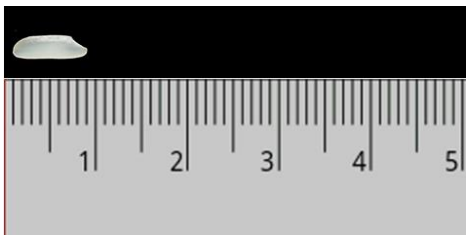
Gambar 1. Diagram alir metode percobaan

2.1. Persiapan

Disiapkan *sample* butir beras Taiwan jenis CNS *grade* 1 dan *sample* butir beras Thailand jenis Thai Hom Mali *grade* A masing-masing sebanyak 100 butir.



Gambar 2. Contoh butiran beras Taiwan jenis CNG *grade* 1



Gambar 3. Contoh butir beras Thailand jenis Thai Hom Mali *grade* 1

Setelah itu, disiapkan dua ratus butir sampel beras untuk masing-masing varian. Pertama-tama seratus butir *sample* beras Thailand jenis Thai Hom Mali *grade* 1 dimasukkan ke dalam botol dan dilanjutkan dengan memasukkan seratus butir *sample* beras Taiwan CNS *grade* A.

Gambar 4. Materi percobaan

Langkah kedua, dua ratus butir sampel beras dalam botol tersebut kemudian dikocok secara vertikal sebanyak duapuluh kali hingga kedua jenis beras tersebut tercampur dengan baik. Langkah ketiga, sebanyak seratus empat puluh butir beras diambil secara acak untuk selanjutnya dilakukan ekstraksi ciri.

2.2. Ekstraksi Ciri

Dalam penelitian ini digunakan ekstraksi ciri morfologi untuk mengamati karakteristik khusus dari masing-masing butiran beras. Berikut ini adalah ciri-ciri morfologi yang diekstrak/diukur dari setiap butiran beras:

a. Panjang, yaitu panjang maksimum dari butiran beras
b. Lebar, yaitu lebar maksimum dari butiran beras

2.3. Algoritma K-*means Clustering*

Pengelompokan K-*means* merupakan jenis pengelompokan yang sering digunakan dalam pengelompokan *unsupervised.* Tahapan dari implementasi algoritma melakukan K-*Means clustering* adalah sebagai berikut: [1]

1. Menentukan K buah titik *centroid* secara acak
2. Mengelompokkan data sehingga terbentuk *cluster* sebanyak K buah *cluster* dengan titik *centroid* dari setiap *cluster* merupakan titik *centroid* yang telah dipilih sebelumnya
3. Memperbaharui nilai titik *centroid*
4. Mengulangi langkah 2 dan 3 sampai nilai dari titik *centroid* cenderung tetap

Proses pengelompokan data ke dalam suatu *cluster* dilakukan dengan menghitung jarak terdekat antara data dengan titik *centroid*. Dalam hal ini digunakan jarak Minkowski, sebagai berikut [2]:

$$d(x_i, x_j) = (|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + ... + |x_{ip} - x_{jp}|^g)^{1/g} \quad .....(1)$$

dimana:
g = 1, untuk menghitung jarak Manhattan
g = 2, untuk menghitung jarak Euclidean
g = ∞, untuk menghitung jarak Chebychev
x_i , x_j adalah dua buah data yang akan dihitung jaraknya
p = dimensi dari sebuah data

Pembaharuan suatu titik *centroid* dapat dilakukan dengan rumus berikut: [2]

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad .....(2)$$

dimana:
μ_k = titik *centroid* dari *cluster* ke-K
N_k = banyaknya data pada *cluster* ke-K
x_q = data ke-q pada *cluster* ke-K

2.4. Algoritma *Estimated Maximation*

Algoritma *Estimated Maximation* (EM) merupakan suatu metode berulang untuk menemukan *mixture of Gaussian* yang bisa memodelkan *data set* [3]. EM adalah salah satu algoritma yang berdasarkan model (*model-based clustering*), dimana pendekatannya adalah menggunakan model sebagai acuan dalam pengelompokan kemudian mengoptimalkan kesamaan antara data dengan model [3].

Terdapat fungsi *likelihood L (θ; x, z),* dimana *θ* adalah parameter *vector*, *x* adalah sekumpulan data yang diamati, dan *z* adalah *unobserved data*. *Maximum likelihood estimate* (MLE) ditentukan oleh *marginal likelihood* dari data yeng diobservasi [3].

Algoritma EM mencari *maximum likelihood estimate* (MLE) dari *marginal likelihood* dengan melakukan langkah berikut secara berulang: [3]

1. Langkah *Expectation* (Langkah E)
   Menghitung nilai ekspektasi dari fungsi *log likelihood*

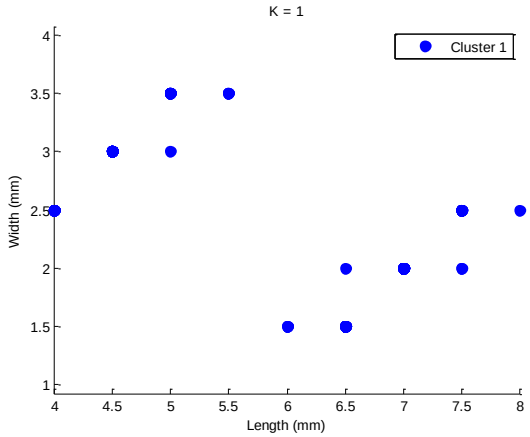
$$Q(\theta|\theta^{(t)}) = \mathrm{E}\left[\log L(\theta; x, Z) \middle| x, \theta^{(t)}\right] \quad .....(3)$$

2. Langkah *Maximation* (Langkah M)
   Menemukan parameter *maximizes*

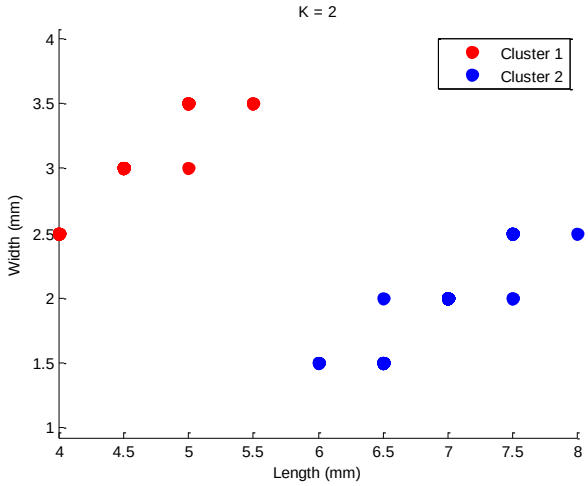
$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}) \quad .....(4)$$

3. HASIL

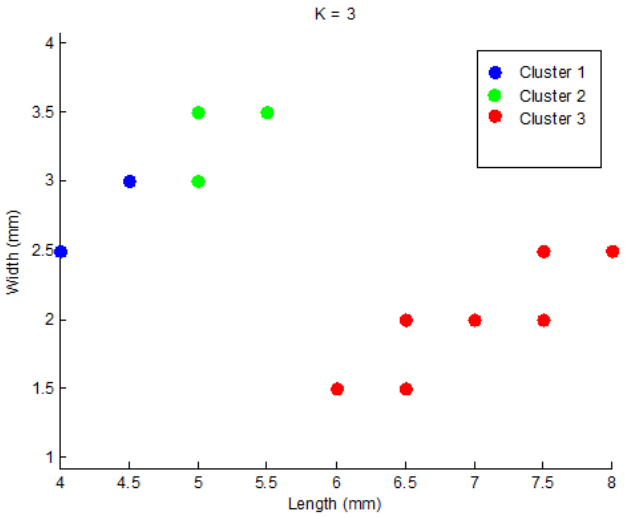
Berikut adalah hasil-hasil percobaan yang dilakukan.



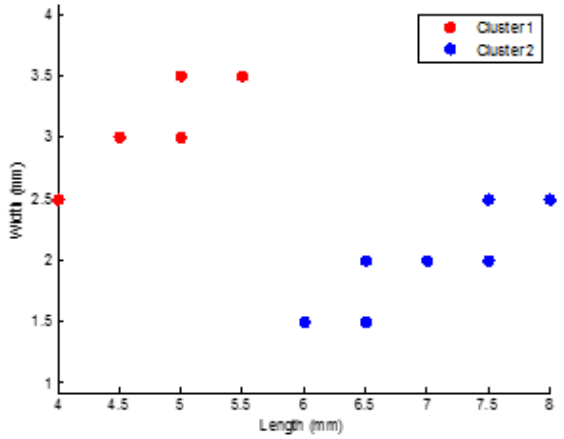
Gambar 5. Hasil pengelompokan K-*means clustering* dengan K =1



Gambar 6. Hasil pengelompokan K-*means clustering* dengan K =2



Gambar 7. Hasil pengelompokan K-*means clustering* dengan K =3

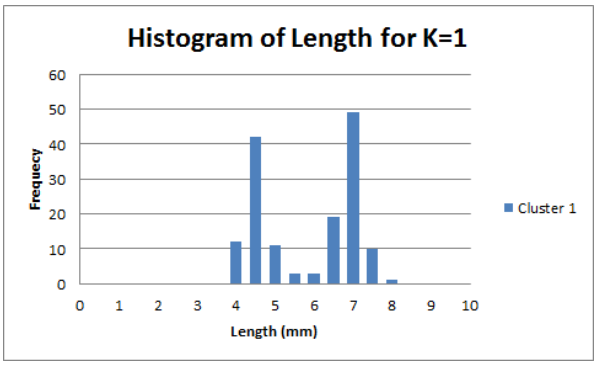


Gambar 8. Hasil Pengelompokan menggunakan algoritma EM

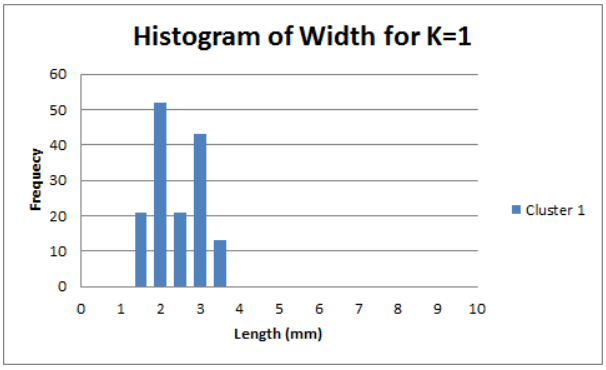
4. PEMBAHASAN

Hasil percobaan menunjukkan klasifikasi dari sebanyak 150 *sample* bulir beras campur antara beras Thailand dan beras Taiwan menggunakan algoritma K-*means clustering* dan algoritma EM. Klasifikasi tersebut didasarkan pada perbedaan ciri morfologi, yaitu panjang dan lebar dari masing-masing *sample* bulir beras.

Gambar 5 adalah hasil pengelompokan menggunakan algoritma K-*means clustering* dengan nilai K=1. Data histogram dari gambar 5 tersebut ditunjukkan pada Gambar 9 untuk fitur panjang dan Gambar 10 untuk fitur lebar.

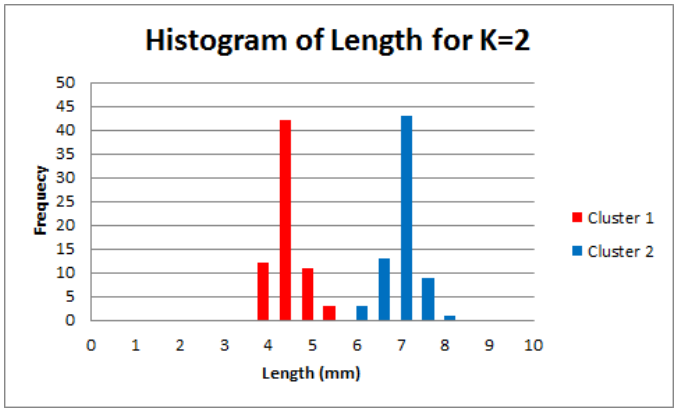


Gambar 9. Histogram fitur panjang hasil pengelompokan menggunakan algoritma K-*means* dengan k=1.

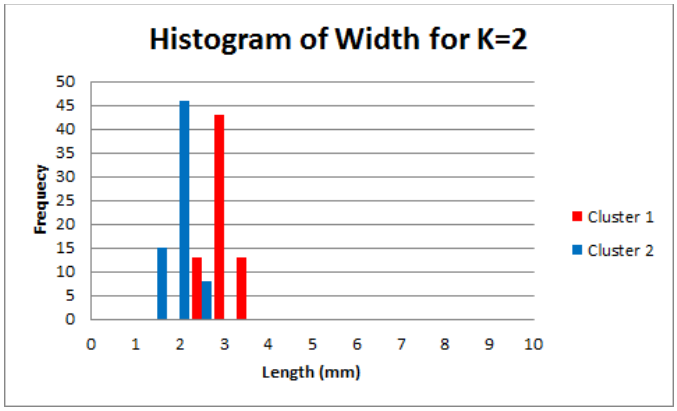


Gambar 10. Histogram fitur lebar hasil pengelompokan menggunakan algoritma K-*means* dengan k=1.

Gambar 6 adalah hasil pengelompokan menggunakan algoritma K-*means clustering* dengan nilai K=2. Data histogram dari Gambar 6 tersebut ditunjukkan pada Gambar 11 untuk fitur panjang dan Gambar 12 untuk fitur lebar.

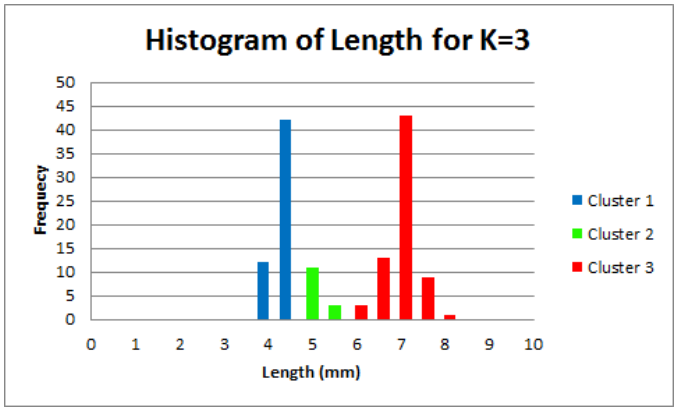


Gambar 11. Histogram fitur panjang hasil pengelompokan menggunakan algoritma K-*means* dengan k=2.

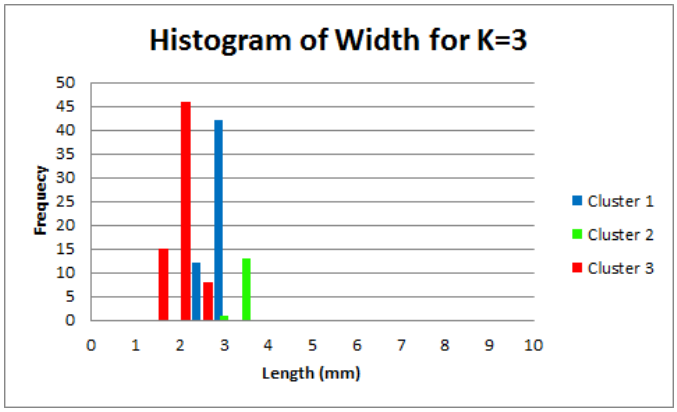


Gambar 12. Histogram fitur lebar hasil pengelompokan menggunakan algoritma K-*means* dengan k=2.

Gambar 7 adalah hasil pengelompokan menggunakan algoritma K-*means clustering* dengan nilai K=3. Data histogram dari Gambar 7 tersebut ditunjukkan pada Gambar 13 untuk fitur panjang dan Gambar 14 untuk fitur lebar.

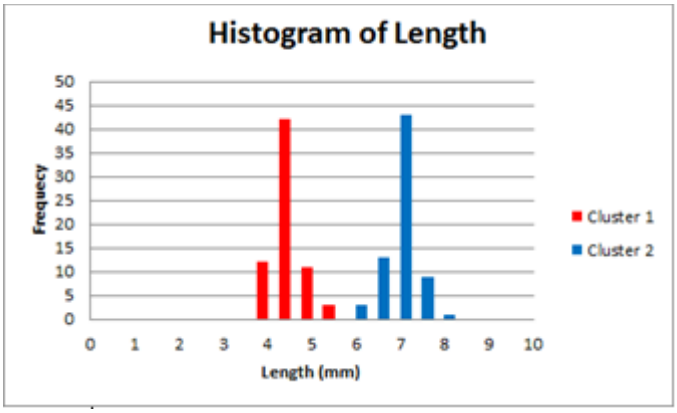


Gambar 13. Histogram fitur panjang hasil pengelompokan menggunakan algoritma K-*means* dengan k=3.

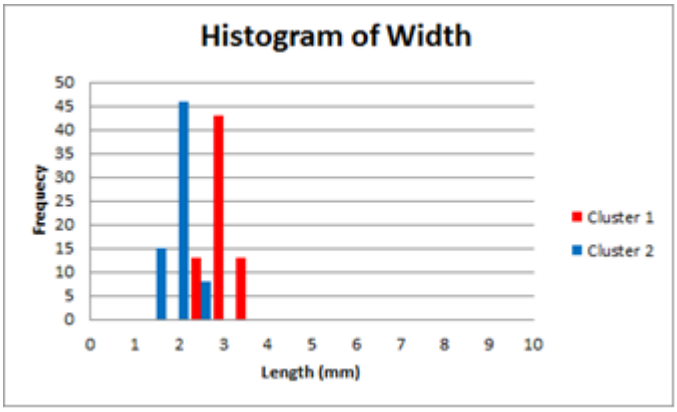


Gambar 14. Histogram fitur lebar hasil pengelompokan menggunakan algoritma K-*means* dengan k=3.

Gambar 8 merupakan hasil pengelompokan menggunakan algoritma EM. Data histogram dari Gambar 8 tersebut ditunjukkan pada Gambar 15 untuk fitur panjang dan Gambar 16 untuk fitur lebar.



Gambar 15. Histogram fitur panjang hasil pengelompokan menggunakan algoritma EM.



Gambar 16. Histogram fitur lebar hasil pengelompokan menggunakan algoritma EM

Analisis hasil histogram menunjukkan kemiripan data *cluster* antara hasil pengelompokan K-*means clustering* dengan nilai k=2 dengan hasil pengelompokan menggunakan algoritma EM. Dari percobaan ini juga terlihat kekurangan dari algoritma K-*means clustering* dalam pengelompokan jenis beras, yaitu dalam mencari nilai k yang optimal. Penggunaan algoritma EM dalam penelitian ini dapat digunakan sebagai data pembanding sehingga bisa diperoleh kesimpulan hasil *unsupervised clustering* dua jenis

bulir beras Taiwan dan Thailand berdasarkan karakteristik morfologi menghasilkan *cluster* optimal sebanyak dua kelompok, yaitu:

Cluster 1 : 4 (mm) ≤ Length ≤ 5.5 mm
Cluster 2 : 6 (mm) ≤ Length ≤ 8 mm

5. KESIMPULAN

Penentuan nilai K dalam algoritma K-*means clustering* sangat penting untuk mendapatkan hasil pengelompokkan yang optimal. Pengelompokan K-*means clustering* akan lebih baik jika disertai juga dengan metode pembanding untuk memperoleh hasil pengelompokkan yang optimal dalam hal ini menggunakan algortima EM. Algoritma EM dan algoritma K-*means clustering* merupakan algoritma yang baik untuk mengatasi masalah *unsupervised clustering* seperti dalam penelitian ini.

Selain dari metode pengelompokan itu sendiri, penentuan fitur-fitur yang digunakan untuk klasifikasi juga sangat penting untuk mendapatkan hasil pengelompokkan yang optimal.

DAFTAR PUSTAKA

[1]. P.-N. Tan, M. Steinbach, dan V. Kumar. 2005. Introduction to Data Mining, (First Edition). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

[2]. O. Maimon dan L. Rokach. 2005. Data Mining and Knowledge Discovery Handbook. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

[3]. Dempster, A.P., N.M. Laird, dan D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society, Series B* 39 (1): 1–38. JSTOR 2984875. MR 0501537.