The hard c-means algorithm for clustering Indonesian plantation commodity based on metabolites composition

Rustam$^1$, A Y Gunawan$^1$ and M T A P Kresnowati$^2$

$^1$Industrial and Financial Mathematics Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Indonesia
$^2$Food and Biomass Processing Technology Research Group, Faculty of Industrial Technology, Institut Teknologi Bandung, Indonesia

E-mail: rustam.math@s.itb.ac.id

Abstract. Indonesia is the second largest biodiversity country in the world. Indonesia has a variety of top agricultural and plantation commodities, whereas each commodity has premium products which are normally related to its specific character/flavour and are generally associated with the producing origins. Specific character/flavour is represented by composition of metabolites each origin. To find out the specific character/flavour in various Indonesian plantation commodities, clustering is needed. In this paper, we perform clustering on an Indonesian plantation commodity based on their metabolites composition. Metabolite compositions of samples of some origins of the commodity are clustered using the Hard C-Means algorithm with the Xie-Beni index as the cluster validity. Our present results confirm that each origin has a unique characteristic and belongs to a separate cluster.

1. Introduction

Indonesia produces various agricultural and plantation commodities. The quality and taste of each commodity is not uniform, and there are several superior commodities with specific tastes. For example, Java cloves bud is known having a prominent flavor of woody, dusty, and acid, whereas Bali cloves bud and its derivative product have a dominant flavor of sweet, spicy, and floral flavor [1]. Arabian coffee from Gayo has black tea, jasmine, and fruity flavor, Arabian coffee from Java Preanger has roasted sweet potatoe, honey, and fruity flavor, whereas Arabian coffee from Kintamani has sweet and fruity flavor [2]. To find out the specific flavour found in various Indonesian plantation commodities, clustering is needed.

Clustering is a process that classifies a set of observation data into several groups [3]. Clustering can also be interpreted as a process of grouping data into separate clusters, so that data in one cluster have the maximum level of similarity, while between two clusters the data have minimum similarity [4]. Generally, similarity in one cluster is measured using the concept of distance [5]. One the distance concept is commonly used to measure geometry proximity distance between the data is Euclidean distance. In clustering, two main approaches are used, namely partitional clustering and hierarchical clustering [6]. In this study, we apply partitional clustering because we have data set come from twelve regions that will be determined the best number of its partitions. Some applications of clustering have been widely used such as speech recognition in engineering [7], cancer cell classification in the medical world [8], and clustering of companies based on the share price in economics [9].
An Indonesian plantation commodity has the specific flavor [1]. The specific flavour is represented by composition of metabolites. To find out the specific flavour in various Indonesian plantation commodities, we perform clustering based on metabolites data set. Metabolite is chemical compounds involved in metabolism processes [10]. To extract useful information from these metabolites data set, multivariate analysis is needed. The Principle Component Analysis (PCA) is one of multivariate analysis that is widely used in clustering metabolites data set. However, not all of the metabolites data collected can be decomposed by PCA into a few principal components, especially if the differences between clusters are minor and obscured by other covariates [11]. Therefore, in this study we propose the Hard C-Means algorithm for clustering metabolites data set.

The composition of this paper is organized as follows. In Section 2, the Hard C-Means algorithm is explained. In Section 3, data, transformation of data, and dimension reduction of data are explained, and that consists on two data set i.e. with out data dimensions reduction (model 1) and with data dimensions reduction (model 2). In Section 4, clustering is performed to find out the best number of clusters of the data set. Finally, in Section 5 conclusions of this paper are written and Section 6 is references.

2. The Hard C-Means Algorithm

In this study, we propose the Hard C-Means algorithm because of its simple principle, ease of programming and performance in large data sets [12]. The Euclidean distance concept is one of the distance concepts that is commonly used to measure geometry proximity between the data [13]. In the Hard C-Means algorithm, clusters are considered as a hard subset which means that a data belongs to one of the existing clusters and does not belongs to two or more clusters at the same time. Consider a data set \( X = \{x_1, \ldots, x_N\} \) which contains \( N \) observations. Each observation is a \( n \)-dimensional row vector, \( x_j = \{x_{k1}, \ldots, x_{kn}\} \in \mathbb{R}^n \). We write the degree of membership the data \( j \)-th in the cluster \( i \)-th as \( u_{ij} \), then we form a partition matrix \( U = [u_{ij}]_{c \times N} \), where \( c \) is number of clusters. This partition matrix fulfills the following properties

- \( u_{ij} \in \{0,1\}, 1 \leq i \leq c, 1 \leq j \leq N \),
- \( \sum_{i=1}^{c} u_{ij} = 1, 1 \leq j \leq N \),
- \( 0 < \sum_{j=1}^{N} u_{ij} < N, 1 \leq i \leq c \).

To ensure that each data point is assigned exactly to one cluster, the second property of the Equation (1) is required. Together \( u_{ij} \in \{0,1\} \), it is possible that data may be assigned to one or more clusters, while there are several remaining clusters left empty. Because such a situation is undesirable, the third property of Equation (1) is required.

The objective function that will be minimized in the Hard C-Means algorithm is [14]

\[
Z = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij} \|x_j - v_i\|^2,
\]

where \( v_i \) is a vector of cluster center \( i \)-th, \( D_{ij}^2 = \|x_j - v_i\|^2 \) is Euclidean distance the data \( j \)-th to the cluster centre \( i \)-th. Basically to determine \( u_{ij} \) and \( v_i \) must be done simultaneously. However, in this study, the determination of \( u_{ij} \) and \( v_i \) is carried out numerically by the recursive method. In general, we have the choice to determine which one is the initial guess. In this study, we chose \( v_i \) to determine \( u_{ij} \). The Hard C-Means algorithm is given as follows [15].

**Step I** : Given data \( x_j \in \mathbb{R}^n, j = 1,2, \ldots N \), determine the number of clusters \((1 < c \leq N)\), specify iterate termination criteria \( \varepsilon > 0 \). Initiation the cluster centre \( v_i^{(0)} \). Cluster centers can be selected randomly from a collection of data set \( x_j \in \mathbb{R}^n, j = 1,2, \ldots N \).

Repeat for \( l = 1,2, \ldots :\)

**Step II** : Calculate the Euclid distance \( D_{ij}^{(l)} = \|x_j - v_i^{(l-1)}\|^2 \).

**Step III** : Each data point is assigned to its closest cluster center with membership degree
\[ u_{ij}^{(l)} = \begin{cases} 1, & \text{min}_{i=1}^{c} \| x_j - v_i^{(l-1)} \|^2, \quad 1 \leq j \leq N. \\ 0, & \text{otherwise} \end{cases} \]  

(3)

**Step IV:** Update the cluster centers with the formula as follows

\[ v_i^{(l)} = \frac{\sum_{j=1}^{c} u_{ij}^{(l)} x_j}{\sum_{j=1}^{c} u_{ij}^{(l)}}, \quad 1 \leq i \leq c. \]  

(4)

Repeat step II to step IV until \( \| v_i^{(l)} - v_i^{(l-1)} \| < \varepsilon \) is fullfied, \( \forall i \) \( (1 \leq i \leq c) \). In this study, we chose \( \varepsilon = 10^{-5} \).

In this study, we propose the Xie-Beni index as cluster validity because the Xie-Beni Index involves cluster centres \( (v_i) \) and data \( (x_j) \) in calculating its index value. It aims to measure the ratio of the total variation within clusters and the separation of clusters. The Xie-Beni index is given as follows

\[ XB = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij} \| x_j - v_i \|^2}{N \min_{i \neq j} \| v_i - v_j \|^2}. \]  

(5)

The smallest Xie-Beni index value shows the best number of clusters [16].

**3. Data and Data Dimension Reduction**

In this study, we will apply the Hard C-means algorithm for the metabolites data set on an Indonesian plantation commodity. The commodity in question is clove buds. Indonesian clove buds consist of four origins, namely Java, Bali, Manado, and Toli-Toli. Three regions of the aforementioned origin. Totally we have twelve regions. For the data from its region, it is conducted eight experiments, except one region of the Java has only six experiments. For each experiment, there are 47 metabolites recorded [17]. So the total size of the data set that owned in the matrix is 94 × 47. The rows of this matrix represent the sample of data and the columns represent the metabolites. We called the data arranged by that matrix as model 1.

The data in Kresnowati et al. [17] has large numerical range between \( 10^{-4} \) and 10. For this type of data, it is common to apply logarithmic transformation in order to have reliable numerical data. However, some data in Kresnowati et al. [17] have zero values. So that the logarithmic transformation cannot be applied directly. The metabolites data with zero values may mean that the metabolites are not present in the origin or the tools of the instrument used in the laboratory cannot detect the metabolites. We do not remove the metabolites data which zero values because they may have an important contribution to the origin as a biological marker in the origin. Therefore, we chose to replace zero values with \( 10^{-5} \).

![Figure 1. Plot of the metabolites data set in model 1](image-url)
Figure 1 shows a plot of the transformed metabolites data on model 1. When we consider all data to cluster, it may some experimental data in a region mix with other regions. To overcome this problem, we propose by doing data dimensions reduction. In this study, data dimension reduction is carried out in each region so that each region is represented by one data set metabolites with size in the matrix is $1 \times 47$. Because there are twelve regions, the size of the data set in matrix is $12 \times 47$. We called the data arranged by that matrix as model 2. We perform data dimension reduction using the median because of our data contains outliers. The median is less affected by outliers [18].

Figure 2. Plot of the metabolites data set in model 2

Figure 2 shows the metabolites data set after data dimensions reduction. Both of these models (Figure 1 and Figure 2) are clustered using the Hard C-Means algorithm. The best cluster result both of these models are explained in Section 4.

4. Cluster Results and Discussion
In this study, the number of data grouped based on the number of origin data sets. Therefore, for model 1, we evaluated twelve clusters, $(1 < c \leq N)$. For each cluster, the value of the Xie-Beni index is calculated using Equation (5). Similarly for model 2, the number of clusters evaluated is twelve clusters. For each cluster the value of the Xie-Beni index is calculated using Equation (5).

Figure 3. The Xie-Beni index values in model 1

The values of the Xie-Beni index shown in Figure 3 are the values for model 1. The smallest Xie-Beni index value obtained is 0.183 on four clusters. This means there are four the best number clusters for model 1. In Table 1 below, the J initial denotes the Java origin, the B initial denotes the Bali origin,
the initial M denotes the Manado origin, and the T initial denotes the Toli-Toli origin. The first number after the initial letter shows the region. The second number shows the experiment on the first number (the region). For example, M27 means the data sample comes from Manado, the second region, and it is carried out on the seventh experiments.

| Number of Clusters | Cluster Member | Xie-Beni |
|--------------------|----------------|----------|
| 1-st               | J11,J12,J13,J14,J15,J16, J21,J22,J23,J24,J25,J26,J27,J28, J31,J32,J33,J34,J35,J36,J37,J38 |          |
| 2-nd               | B11,B12,B13,B14,B15,B16,B17,B18, B21,B22,B23,B24,B25,B26,B27,B28, B31,B32,B33,B34,B35,B36,B37,B38 | 0.183 |
| 3-rd               | M11,M12,M13,M14,M15,M16,M17,M18, M21,M22,M23,M24,M25,M26,M27,M28, M31,M32,M33,M34,M35,M36,M37,M38,T22,T33 |          |
| 4-th               | T11,T12,T13,T14,T15,T16,T17,T18, T21 ,T23,T24,T25,T26,T27,T28, T31,T32,T34,T35,T36,T37,T38 |          |

Table 1 shows the best number of clusters in model 1, and its cluster members for each cluster. While the minimum value of the objective function (Equation 2) is shown in Table 2 and Figure 4.

### Table 2. Iteration in model 1

| Iteration | Value of Objective Function |
|-----------|-----------------------------|
| 1-st      | 987,779                     |
| 2-nd      | 711,228                     |
| 3-rd      | 652,791                     |
| 4-th      | 637,829                     |
| 5-th      | 633,663                     |
| 6-th      | 625,731                     |
| 7-th      | 625,731                     |

Table 2 and Figure 4 shows the minimum value of the objective function is achieved at the 7-th iteration. This guarantees that at the 7-th iteration the membership in each cluster has been optimum. If we pay attention to the cluster results obtained in Table 1, it can be seen that T22 and T33 are members of the third cluster which should be at the fourth cluster. Therefore we perform data dimensions reduction to overcome these errors. The best number of clusters for model 2 is shown in Figure 5.

![Figure 4. Plot the value of the objective function in model 1](image-url)
Figure 5. The Xie-Beni index values in model 2

Figure 5 shows for model 2 there are four the best number of clusters with the smallest Xie-Beni index value is 0.0958 (see Table 3). As explained earlier, the reduction of data dimensions is carried out for each region, so that each region is represented by one data set. For example, the first region of the Toli-Toli origin (T11, T12, T13, T14, T15, T16, T17, T18) is reduced to T1.

Table 3. The best number of clusters in model 2

| Number of Clusters | Cluster | Cluster Member | Xie-Beni |
|--------------------|---------|----------------|----------|
| 4                  | 1st     | J1, J2, J3     | 0.0958   |
|                    | 2nd     | B1, B2, B3     |          |
|                    | 3rd     | M1, M2, M3     |          |
|                    | 4th     | T1, T2, T3     |          |

Table 1 shows the best number of clusters in model 2, and its cluster members for each cluster. While the minimum value of the objective function (Equation 2) is shown below.

Table 4. Iteration in model 2

| Iteration | Values of Objective Function |
|-----------|------------------------------|
| 1         | 60,747                       |
| 2         | 46,108                       |
| 3         | 46,108                       |

Figure 6. Plot the values of objective function in model 2

Table 4 and Figure 6 show the minimum value of the objective function achieved at the 3-rd iteration. This guarantees that at the 3-rd iteration the membership in each cluster has been optimum. This result confirms that data dimensions reduction can efficient clustering time from 7 iterations become 3 iterations. Four the best number of clusters obtained on model 2 show that each region is properly mapped, in other words, each origin has a specific taste or unique character/flavour. Thus, by reducing the dimensions of the data set using the median, errors in the measurement of several metabolites in each region that may occur in experiments in the laboratory can be overcome.
5. Conclusion
In this study, we have clustered one of Indonesian plantation commodity based on the composition of their metabolites. One of the Indonesian plantation commodities in question is cloves bud. Clove buds consist of four origins, namely Java, Bali, Manado, and Toli-Toli. The data set of metabolites consist of two data sets i.e. with out data dimensions reduction (model 1) and with data dimensions reduction (model 2). By using the Hard C-Means algorithm with the Xie-Beni index as cluster validity, we found that model 1 there was four best clusters. While for model 2 also there was four best clusters. Our results confirmed that dimensional data reduction using the median could be overcome errors about some experimental data in a region mix with other regions. Finally, we concluded that each origin of one of Indonesian plantation commodity (clove buds) had a specific taste or unique character/flavour, different from each other.

6. References
[1] Broto L S 2014 Derivatisasi minyak cengkeh, dalam cengkeh: Sejarah, Budidaya dan Industri (Salatiga: Indesco Jakarta dan Biologi UKSW Salatiga) p.123-164.
[2] Coffeeland 2010 Jenis kopi Arabika terbaik dari berbagai daerah di Indonesia Online: [https://coffeeland.co.id/product-category/kopi-supply/](https://coffeeland.co.id/product-category/kopi-supply/)
[3] Kumari N, Sharma B and Gaur D 2012 Implementation of possibilistic fuzzy c-means clustering algorithm in Matlab International Journal of Scientific & Engineering Research 3 1.
[4] Tan P N, Steinbach M and Kumar V 2006 Introduction to Data Mining (Boston: Addison-Wesley Longman Publishing Co., Inc.).
[5] Soler J, Tencé F, Gaubert L and Buche C 2013 Data clustering and similarity The Twenty-Sixth International FLAIRS Conference.
[6] Agusta Y 2007 K-means-penerapan, permasalahan dan metode terkait Jurnal Sistem dan Informatika 3 47.
[7] Tran D and Wagner M 1999 Fuzzy hidden Markov models for speech and speaker recognition 18th International Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 99TH8397) pp 426-430 IEEE.
[8] Taher F and Sammouda R 2011 Lung cancer detection by using artificial neural network and fuzzy clustering methods IEEE GCC Conference and Exhibition (GCC) pp 295-298.
[9] Gumilar A C 2014 Penerapan fuzzy c-means clustering pada pengelompokan perusahaan berdasarkan nilai harga saham Tesis Program Studi Magister Matematika Institut Teknologi Bandung.
[10] Meinicke P, Lingner T, Kaever A, Feussner K, Göbel C, Feussner I and Morgenstern B 2008 Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps Algorithms for Molecular Biology 3 9.
[11] Li X, Lu X, Tian J, Gao P, Kong H and Xu G 2009 Application of fuzzy c-means clustering in data analysis of metabolomics Analytical Chemistry 81 4468.
[12] Zhi X B, Fan J L and Zhao F 2014 Robust local feature weighting hard c-means clustering algorithm Neurocomputing 134 20.
[13] Ross T J 2004 Fuzzy Logic with Engineering Applications (New York: Wiley).
[14] Bezdek J C 1981 Pattern Recognition with Fuzzy Objective Function Algorithms (New York: Plenum Press).
[15] De Oliveira J V and Pedrycz W 2007 Advances in Fuzzy Clustering and Its Applications (West Sussex: Wiley).
[16] Xie X L and Beni G 1999 A validity measure for fuzzy clustering IEEE Transactions on Pattern Analysis and Machine Intelligence 13 841.
[17] Kresnowati M T A P, Purwadi R, Zunita M, Sudarman R and Putri A O 2018 Metabolite profiling of four origins Indonesian clove buds using multivariate analysis in preparation.
[18] Manikandam S 2011 Measures of central tendency: Median and mode Journal of pharmacology and pharmacotherapeutics 2 214.