Cocoa Beans Data Grouping With Fuzzy C-Means Clustering Method

Yasi Dani, Anneke Iswani, Yurika Permanasari

Abstract: Fermentation of cocoa beans plays an important role in determining the quality of the cocoa beans. Recently a new technique has been developed to test the taste of cocoa beans, namely the Metabolic Fingerprinting Technique. This technique observes the profile comparison of metabolite compounds to be applied in determining the similarities and differences between two or more samples. Since the experimental data may contain many uncertainty factors, such as measurement error, changes in lab environment, and so forth, the fuzzy clustering technique would give the appropriate answer. As a result, a set of metabolic fermentation data from several cocoa beans will be grouped by the fuzzy c-means clustering algorithm (FCM). FCM is a grouping technique where the existence of each data point in a cluster is determined by the degree of membership. This algorithm is based on the minimization of objective functions illustrated by finding the shortest distance between the center of the group and each data point weighted by the degree of membership. Furthermore, the most optimal cluster will be determined by using the Xie-Beni index.

Index Terms: Cocoa beans, cluster, metabolite compounds, fermentation, metabolic fingerprinting technique, Fuzzy c-Means clustering, Xie-Beni index.

1 INTRODUCTION

COCOA is one of the commodities that is quite widely used in the industrial world. Cocoa beans can be processed into various products, because cocoa beans contain distinctive flavors and colors that are very popular and much in demand. One of the downstream products from cocoa beans is cocoa powder which can then be processed into several new products of high economic value. The users of processed cocoa products are the food and beverage industry that is increasingly growing due to population growth and social welfare [1]. The taste and aroma of chocolate will be better in cocoa beans that have been a good fermentation, it can be analyzed by chemical methods. Currently, a new technique for testing the taste of food is being developed, namely the Metabolic Fingerprinting technique. This is a technique that observes the comparison of profiles of metabolites to be applied in determining the similarities and differences between two or more samples. Profile metabolites explain that such diversity is influenced by genotype and environmental factors. This metabolite compound is a compound that allows metabolism [2]. Measurement of various metabolite compounds can be done using a combination of chromatography and spectroscopy techniques, namely Gas Chromatography Mass Spectrometry and Liquid Chromatography Mass Spectrometry. In gas or liquid chromatography techniques the various metabolites in cocoa beans are separated while in the mass spectrometry technique the metabolites are identified based on their molecular weight. In this study, cocoa beans were processed with various treatments, namely beans not fermented or fermented either in the field or in the laboratory [3].

2 LITERATURE REVIEW

2.1 Fermentation

Fermentation is a process that utilizes the ability of microbes to produce primary and secondary metabolites in a controlled environment [4]. Fermented products can be classified into 4 types:

1. biomass products
2. enzyme products
3. metabolite products
4. transformation products

In bioprocess fermentation plays an important role since it is the key (main process) for the production of biologically based materials. Ingredients produced by fermentation are the results of microbial cell metabolites, such as antibiotics, organic acids, aldehydes, alcohols, and so on. In addition to the results of these metabolites, fermentation can also be applied to produce microbial cell biomass such as baker’s yeast which is used in making bread. To produce each of the fermentation products above, different fermentation conditions are required and the types of microbes that vary also have their characteristics. Therefore, environmental conditions, substrates (media), as well as appropriate treatments are needed so that the resulting product is optimal.

2.2 Metabolites

Metabolites are organic compounds that are produced and involved in metabolism such as food, water and medicine. Metabolism involves the breakdown of complex organic matter, which is then used as a source of energy in the body. The body uses this energy for growth, reproduction and overall body maintenance. Metabolites are present at all stages of the metabolic process. They help in the breakdown of vitamins and amino acids, as well as in the construction of complex molecules. There are two basic forms of metabolites of microorganisms called primary and secondary metabolites. The primary metabolite is one that is formed during the primary growth phase of microorganisms, while the secondary metabolite is one that is formed towards the end of the primary growth phase of microorganisms, often towards or stationary phase of growth [8].
2.2.1 Primary Metabolites
Primary metabolite is a metabolite or molecule which is the final product or intermediate product in the metabolic process of living things, whose function is essential for the survival of the organism, and is formed intracellularly. Examples are protein, fat, carbohydrates, and DNA. In general, primary metabolites are not over-produced. In most microorganisms, excessive production of metabolites can inhibit growth, and sometimes can kill these microorganisms. The metabolic process to form primary metabolites is called primary metabolism. Microorganisms produce primary metabolites such as ethanol and secondary metabolites such as antibiotics. Primary metabolites are produced at the same time as new cell formation, and the production curve follows a population growth curve in parallel. Secondary metabolites of microorganisms are not produced until the microorganism cells complete the logarithmic growth phase, known as the trophophase phase and enter the stationary phase. The next period, when most secondary metabolites are produced, is referred to as idrophase. Secondary metabolites of microorganisms can be the conversion of primary metabolites of microorganisms.

2.2.1 Secondary metabolites
Secondary metabolites are produced by microorganisms after the active growth phase has stopped. These substances are usually not needed for metabolism or maintenance of important cell purposes. Although not needed for growth, secondary metabolites can also function as emergency nutrients for survival. The function of secondary metabolites for the producing microorganisms themselves is largely unclear. Secondary metabolites are made and stored extracellularly. Secondary metabolites are not produced during rapid cell growth (logarithmic phase) but are usually synthesized at the end of the cell growth cycle, i.e., in the stationary phase when the cell population remains constant because the number of cells growing is equal to the number of dead cells. In this phase microorganism cells are more resistant to extreme conditions, for example, hotter or cooler temperatures, radiation, chemicals, and metabolites that they produce themselves (antibiotics).

2.3 Metabolic Fingerprinting
Metabolic fingerprinting is based on modern science which is an effective way to analyze bio-samples with complex chemical information [7]. Metabolic fingerprinting is a technique used to determine the similarities and differences between two or more samples by using a comparison of the profiles of metabolite compounds. Measurement of various metabolite compounds can be done, among them using a combination of either gas or liquid chromatography techniques, with mass spectroscopy (LC-MS or GC-MS) techniques [9].

2.3.1 Gas Chromatography-Mass Spectrometry (GC-MS)
GC-MS is a method of separating organic compounds using two compound analysis methods namely gas chromatography (GC) to analyze the quantity of compounds quantitatively and mass spectrometry (MS) to analyze the molecular structure of analytes. The objects for GC-MS-based metabolomics are volatile compounds or derivates whose stationary phase for GC columns for metabolomic analysis are those coated with phenyl 5 and siloxane 95%. The advantage of this method is the excellent resolution and easy separation between components, the required sample is less than 10 µl. This method is very good for researching pharmacological aspects related to terpenoid components and some phenyl propanoids, for example aromatherapy, herbal components related to nerve drugs.

2.3.2 Liquid Chromatography-Mass Spectrometry (LC-MS)
LC-MS is an analytical chemistry technique that combines the ability of physical separation from liquid chromatography with the ability of mass spectrometer analysis. LC-MS is a technique that is widely used for various applications that have sensitivity and specificity very high. The LC-MS method has a wider ring level than GC-MS and covers almost all classes of secondary metabolites. For the analysis of micromolecular metabolites an inverse system is required, namely the nonpolar stationary phase. The general mobile phase is the polar mobile phase: water, acetonitrile, methanol, acidification with formic acid and phosphoric acid to increase separation. Information obtained such as GC-MS, peaks with a certain area along with Rt and BM fragmentation information. The advantage of the LC-MS method is that the sample size is very small and is sufficiently dissolved in certain organic solvents.

2.4 Fuzzy Partition
Fuzzy partition [11] is the result of generalization of hard partition which the degree of membership has a real degree at intervals [0,1]. This fuzzy partition fulfills the following properties:

\[ \mu_{ik} \in [0,1], \quad 1 \leq i \leq c, \quad 1 \leq k \leq N \]  
(1)

\[ \sum_{i=1}^{c} \mu_{ik} = 1, \quad 1 \leq k \leq N \]  
(2)

\[ 0 < \sum_{i=1}^{c} \mu_{ik} < N, \quad 1 \leq i \leq c \]  
(3)

Definition 1. Given universe X. A fuzzy set A in universe X is defined

\[ A = \{ (x, \mu_A(x)) \mid x \in X \} \]  
(4)

where \( \mu_A : X \rightarrow [0,1] \) is the degree of membership of the fuzzy set \( A \).

The space of all possible fuzzy partition matrices for X called hard partition space is defined as

\[ M_x = \left\{ U \in R^{c \times n} \mid \mu_{ik} \in [0,1], \forall i, k ; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k ; 0 < \sum_{i=1}^{c} \mu_{ik} < N, \forall i \right\} \]  
(5)

2.4 c-MEANS FUNCTION
Most of the fuzzy clustering algorithms are based on the minimization of c-means functions formulated as:

\[ J(Z; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D_{ik}^2 \]  
(6)

with:

\[ U = [\mu_{ik}] \in M_{xc} \] is a fuzzy partition of the data matrix Z

\[ V = [v_1, v_2, v_3, ..., v_c] \] is a vector from the center of the group A = an identity matrix of size N x N since in this case the Euclid norm is used

\[ D_{ik}^2 \] is the norm of the
square of inner product from the i-th center to the k-th data \( m \in [1, \infty) \) is a fuzzy parameter

The value of the function \( J \) above can be considered as a measure of the total variance \( \zeta_i \) of \( \nu_i \). The minimization of the c-means \( J \) function is a nonlinear optimization problem that can be solved by using various methods including simulation annealing, genetic algorithm and Picard iteration. In this study the method to be used is Picard iteration. Picard iteration using the first derivative of the conditions for stationary points on \( J \).

Stationary points for function \( J \) with the \( \sum_{i=1}^{c} \mu_{ik} = 1 \) constraint can be determined using the Lagrange multiplier factor:

\[
J_{\mu}(Z; U, V, \lambda) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D_{ik}^2 + \sum_{k=1}^{N} \lambda \left[ \sum_{i=1}^{c} \mu_{ik} - 1 \right] \tag{7}
\]

by determining the derivative of \( J \) with respect to \( U, V, \) and \( \lambda \) equal to 0.

\[ \text{Theorem 1. For } D_{ik}^2 > 0 \forall i, k \text{ and } m > 0, \quad (U, V) \in M \times R \quad \text{global minimum of } J \text{ only if} \]

\[ \mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \frac{D_{ik}}{D_{jk}} \right)^{m-1}} \quad 1 \leq i \leq c; \quad 1 \leq k \leq N \tag{8} \]

\[ v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m \cdot z_k}{\sum_{k=1}^{N} \mu_{ik}^m} \quad 1 \leq i \leq c \tag{9} \]

2.5 Group Validity Index

Validity index is a measure used to determine the optimal number of groups. In this study using the Xie-Beni index, recommends this index for grouping because it has high accuracy and reliability to be used as a criterion in determining the optimum number of groups. The optimum criteria for many groups are shown in the minimum Xie-Beni index value. For the fuzzy c-means clustering algorithm, the Xie-Beni validity index is as follows [10]:

\[ \chi(Z; U, V) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ik})^m \| z_i - v_i \|^2}{\min_{i, j} \left( \frac{1}{2} \| v_i - v_j \|^2 \right)} \tag{10} \]

This index can be defined as the ratio of total group variance to the distance between group centers. This optimal group means that each data in the group has a maximum of similarity characteristics between each data in a group and the maximum difference between each group.

3 RESEARCH METHOD

The first thing to start fuzzy c-means clustering is to represent the data that will process in \( Z \). The fuzzy c-means clustering algorithm is as follows:

4 RESULT AND DISCUSSION

4.1 Reduction of Outlier Data

This cocoa bean data was obtained from the results of Hana Nurfitriana’s thesis research from the Master program in Chemical Engineering ITB. The data consisted of six treatments for some dried cocoa beans, first, unfermented beans not roasted, second, fermented beans in the field were not roasted, third, seeds fermented in the laboratory were not roasted, fourth, beans were not fermented, roasted, fifth, fermented beans or seeds fermentation results in the roasted field, and sixth, the fermented beans in the laboratory are roasted. Each treatment given to the cocoa beans is carried out three times in data collection, which aims to reduce doubts or anticipate errors, for example in the first treatment three data collection is performed A1 (first data collection), A2 (second data collection) and A3 (third data collection) and so it is with other treatments. All of these treatments resulted in 960 chemical compound profile data with units per peak area. The data is formed into a matrix size of 960 × 18, where the matrix row states the number of peak compounds produced in the treatment and the matrix column states the number of treatments given by three times data collection. The data is formed into a matrix size of 960 × 18, where the line matrix stating the number of compounds that produce the highest peak at the time of the study in one hour and the column stating the number of treatments that are given with three repetitions of experiment, then the data matrix is stored in the form of Microsoft Excel. After all data is stored in Microsoft Excel, the grouping process is carried out using the fuzzy c-means clustering method using MATLAB R2012a software, which obtained the best or optimal group based on the Xie-Beni validity index for all the results of their chemical compound profiles, which is preceded by determining the many desired groups, namely grouping from two to six groups and then getting the Xie-Beni validity index and selecting the smallest value of all groupings.

Steps to reduce these data are:

1. If the values of each iteration do not have a significant difference, then it is enough to find an average of three iteration data for each treatment so that the average results obtained will be representative of the data for each treatment. For example, in the first treatment:
\[ A = \frac{A_1 + A_2 + A_3}{3} \quad (11) \]

2. If the values of each iteration have significant differences, the data reduction process is carried out as follows:
   - Determine the quartile of sorted data.
   - Reducing data which is outlier.

### 4.2 Cocoa Beans Grouping with Fuzzy c-Means Clustering Method

After all the data has been reduced, the grouping process is carried out using the method using MATLAB R2012a software, which is obtained the best or optimal group based on the Xie-Beni validity index for all the results of the chemical compound profile, which is preceded by determining the desired number of groups in which grouping from two to six the group then obtained the Xie-Beni validity index and selected the smallest value of all groupings. The following are the results of grouping using the fuzzy c-means clustering method as shown in Table 1.

#### TABLE 1

<table>
<thead>
<tr>
<th>Numbers of Groups</th>
<th>Group Partition</th>
<th>The Treatment consists of</th>
<th>Xie-Beni Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Groups</td>
<td>Group 1 A, C, D, F</td>
<td>0.0946</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 2 B, E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Groups</td>
<td>Group 1 A</td>
<td>0.5422</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 2 B, E</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 3 C, D, F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Groups</td>
<td>Group 1 E</td>
<td>0.064</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 2 A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 3 C, D, F</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 4 B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Groups</td>
<td>Group 1 A</td>
<td>0.0164</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 2 C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 3 E</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group 4 B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 4.3 Optimal Group

Table 1 shows that the five groups is optimal grouping, because the 5 groups have the smallest Xie-Beni validity value of 0.0164. The grouping of 5 groups (A, B, C, E, and DF) consists of the first group un-fermented un-roasted seeds (A), the second group of un-roasted field fermented beans (B), the third group of un-roasted laboratory fermented beans (C), the fourth group of roasted field fermented beans (E) and the fifth group of unfermented roasted seeds (D) with roasted laboratory fermented seeds (F).

### 5 CONCLUSIONS

In the application of cocoa beans data using the fuzzy c-means clustering method, the more numbers will be grouped, the more processes must be carried out to get the optimal group. Many optimal groupings in this method are 5 groups.

Then, the combination of this method with the application of the estimated mean value on the chemical compound profile values can produce a physical treatment group that is controlled by the chemical compound profile values.

### 6 ACKNOWLEDGMENT

The author would like to thank Dr. Agus Yodi Gunawan for useful discussions.

### 7 REFERENCES


