Comparative Analysis of Hierarchical Clustering with Improve Feature for Herbs Leaves

Febri Liantoni, Nurcahya Pradana Taufik Prakisya, Yusfia Hafid Aristyagama, Puspanda Hatta

1,2,3,4Department of Computer and Informatics Education, Sebelas Maret University, Indonesia
1febri.liantoni@gmail.com, 2nurcahya.pradana@gmail.com, 3yusfia.hafidz@gmail.com
4hatta.puspanda@gmail.com

Abstract. Plants have different leaf morphological characteristics starting from the leaf surface, leaf bone, leaf tips, leaf edges, leaf base, and even leaf color. The study of plants has been progressing rapidly. One of them is the branch of science in plant morphology. Morphological characteristics of leaves are one of the properties possessed by leaves and can be seen clearly. This morphological study learns about the structure of the plant body, especially about the edge leaf shape. This study will be done based on leaf shape. Leaf shape can be used as a reference in the classification process. The classification process requires a good data extraction feature, so it is necessary to improve the feature process at the pre-processing level. Combining the median filter and image erosion is used to improve the feature process. As for feature extraction, the seven moment invariant method is used. The classification method used for grouping is the Hierarchical Clustering. This method consists of four kinds. There are complete link clustering, single link clustering, average link clustering, and centroid linked clustering. Based on the results of experiments conducted, the use of Average Linkage Clustering method obtained clustering accuracy of 87%. It proves that the average method successfully optimizes the result by calculating any distance between points of the cluster. While the smallest accuracy results using complete linkage clustering of 81%.

1. Introduction

Botanical study has been making progress very rapidly. The study that previously were only branches of plants now have become a field of science that stands alone, one of them is Plant Morphology. Plant morphology which learns about the shape and composition of the plant body has been developing rapidly until it is separated into morphology and anatomy plants. One of them was a discussion of plant morphology based on leaf margin. The Plants are the most important part of life on earth. Plants are helpful as the largest oxygen supply for breathing. The other is food, fuel, medicine, cosmetics and more. The process of grouping plants can be done by identifying the leaf shape image of the plant itself. How to take a leaf image of the plant, it can be conducted in step recognition of leaf pattern by recognizing the leaf structural characteristics such as the shape and texture of the leaves (1), (2).

In the process of planting the plants, it can be done by identifying the leaf shape image of the plant. In this way, leaf pattern recognition steps can be recognized by recognizing the structural characteristics.
of leaves such as the shape and texture of a leaf. The method for processing the input image with the utilization of digital image processing techniques is done to analyze the leaf structural characteristics. Actually, shape, color and texture features are common features involved in several applications, such as in (3) and (4). However, some researchers used part of those features only. Invariant moments proposed by Hue (5) are very popular in image processing to recognize objects (6), including leaves of plants (1), (7).

Plant morphology is the study of the shape and composition of the plant body. It is separated into external morphology or morphology only and the inner morphology or plant anatomy. One of them is taken from plant morphology based on leaf margin. According to Gembong, the concept of various kinds of leaf form distribution is as follows (8). Integer (flat edge), Divisus (uneven edge), Leaflets with free shapes (not affecting general shape), Serratus (edge serrated or sawed), Dentatus (toothed edge), Creatus (toothed edge, sharp sinus and blunted angulus), Leaflets with non-free shapes (affecting general shape), Lobatus (paved and crooked grooves), Fissus / vidus (peripheral fracture and fingers), Paritus / diversifolia (edge of pinnate and dance). The data used in the form of a leaf image with a white background.

One technique that can be used to classify data is clustering. The clustering in data mining is the grouping of a number of data or objects into the cluster. So that, each object in the cluster will contain data that are similar and different from the objects in the other cluster. There are two clustering methods that we know, namely hierarchical clustering and partitioning. The hierarchical clustering method itself consists of complete linkage clustering, single linkage clustering, average linkage clustering and centroid linkage clustering.

Digital image processing techniques performed at the stage of image preprocessing to obtain the shape of the edge and structural features of each leaf. The method used in the extraction of this feature is the introduction of digital morphological features (9). The feature extraction process is the Moment Invariant method. After that, performed feature extraction from the leaf image to obtain leaf structural information which then was used as data grouping.

Various types of plants with various forms of leaf edge will be quite difficult to do the plants grouping based on the morphology of leaves. It would be easier if there is an automated system that can categorize the plant according to its morphology. Some examples of research on crop forming, generally using artificial neural network method (1), (10). Artificial neural network methods are widely used because these methods are known to be substantially faster. However, the determination of the number of hidden layers used to affect the results and required parameters of large epochs so that requires higher computation. In 2012, Priya conducted a plant recognition study using the Support Vector Machine (SVM) and k-nearest neighbor (k-NN) (9). The results of the k-NN method are strongly influenced by the presence or absence of irrelevant features. The SVM method is a linear classifier and theoretically only developed for the problem of two classes. In previous research the classification of herbal leaves using Naïve Bayes Classifier method and K Nearest Neighbor. The study compared the two methods with the accuracy of Naïve Bayes Classifier method higher than K Nearest Neighbor (11).

In this research will conduct the comparative analysis of Hierarchical Clustering methods. Prior to the grouping stage, we first perform the image pretest and extraction stage of the leaf edge feature to obtain the appropriate input value for the leaf species classification phase based on the leaf image.

2. Proposed Approach

In this research, there are several steps that are done, there are pre-processed, feature extraction and phase classification of herbal leaf image with Hierarchical Clustering method. The example image used as shown in Figure 1.
A. Preprocessing

Preprocessing is done with the aim to process input data, so that it can be used for the feature extraction process as well as clustering. There are several methods used for preprocessing: image conversion in the RGB color space to the grayscale color space, median filter, binary image, and erosion process. The first stage is to change the image to grayscale. This process is done to convert an image pixel domain to an 8-bit grayscale \((12)\). For the conversion is used as shown in Equation 1.

\[
gray = \frac{red \times 299 + green \times 587 + blue \times 114}{1000}
\]  

(1)

After the image becomes grayscale, then it made the formation of the binary image by using thresholding. A pixel with a larger threshold value will be considered a background, whereas a pixel less than a threshold will be considered a leaf object.

After that edge detection of the image is done by examining the discontinuity on the intensity value. The discontinuities can be detected by using first-order derivatives and second-order derivatives. First-order derivatives of the 2D image gradient function with the function equation \(f(x, y)\) are shown in Equation 2.

\[
\nabla f = [Gx \; Gy] = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}
\]  

(2)

Second-order derivatives are calculated using Laplacian of function \(f(x, y)\) as shown in Equation 3.

\[
\nabla^2 f(x, y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2}
\]  

(3)

The steps performed on the pre-processing are shown in Figure 2.
B. Feature Extraction

Feature extraction is a process for generating feature values in the form of feature vectors from binary leaf images that have been done edge detection. The feature vector is then used for the clustering stage. The feature used is seven-moment invariant which will generate seven values on the feature vector.

C. Moment Invariant Feature

The process of recognizing an object in an image after the segmentation process is often based on object positioning problems, object axis rotation, and scale changes of objects (11), (13), (14). The position of a shifted or rotating object or its size smaller or larger than can cause errors in the recognition or identification of an object.

In the use of the calculation of the value of 2 (two) dimensions of the sample Mx M sample moments of the continuous function \( f(x, y) \), \( (x, y = 0, \ldots, M-1) \) we find Equation 4.

\[
m_{pq} = \sum_{x=0}^{x=M-1} \sum_{y=0}^{y=M-1} (x)^p \cdot (y)^q \cdot f(x, y) \quad (4)
\]

Moments can represent an object in terms of area, position, orientation and other undefined parameters. By obtaining a certain amount of moment information, either the zero \( m_{00} \) and the unity \( m_{10} \) and \( m_{01} \) moments or the central moment, and the moment at the level of \( \geq 2 \) or the invariant moment of an object, the object can be identified even if it has a shift (translation), rotation (rotation) and scale changes.

From moment \( f(x, y) \) will be translated with the value \( (a, b) \) to obtain a new calculation like Equation 5.

\[
\mu_{pq} = \sum_{x} \sum_{y} (x + a)^p \cdot (y + b)^q \cdot f(x, y) \quad (5)
\]

From the main central moment is \( m_{pq} \) or \( \mu_{pq} \) computed through the substitution process to the value of \( a = -x \) and value \( b = -y \) it will get the calculation in Equation 6.

\[
\mu_{pq} = \sum_{x} \sum_{y} (x - \bar{x})^p \cdot (y - \bar{y})^q \cdot f(x, y) \quad (6)
\]

where \( \bar{x} = \frac{m_{10}}{m_{00}} \) and \( \bar{y} = \frac{m_{01}}{m_{00}} \)

When the normalization process then scaling value used in the calculation in Equation 7.

\[
\eta_{pq} = \mu_{pq} / \mu_{pq}^\gamma \quad (7)
\]

where \( \gamma = [(p + q) / 2] + 1 \)

Then from the process then got the value of seven-moments invariant with the calculation in Equation 8 (11), (14).

\[
H_{u1} = \eta_{20} + \eta_{02} \\
H_{u2} = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \\
H_{u3} = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
H_{u4} = (\eta_{30} + \eta_{12})^2 + (3\eta_{21} + \eta_{03})^2 \\
H_{u5} = (\eta_{30} + 3\eta_{12})(\eta_{30} + \eta_{12} + \eta_{30} + \eta_{12})^2 - (3\eta_{12} + \eta_{03})^2 + (3\eta_{21} + \eta_{03}) \\
H_{u6} = (\eta_{30} - \eta_{02})(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
H_{u7} = (3\eta_{21} - \eta_{30})(\eta_{30} + \eta_{12})(\eta_{30} + \eta_{12}^2 - 3(\eta_{12} + \eta_{03})^2) + (3\eta_{21} - \eta_{03})
\]
\[(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\] (8)

D. Feature Normalization

Normalization is a method to equalize the interval of all variable values (15). Normalization occurs when there are variables that have a very large value range, while other variables have a very small value range. Normalized values will have an interval between 0 and 1.

Normalization can be done by using formula as following Equation 9.

\[
\hat{X} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}
\] (9)

Where \(\hat{X}\) represents a new value of the feature, \(X\) represents an original value of the feature, \(X_{\text{min}}\) is the smallest value of the variable, and \(X_{\text{max}}\) is the largest value of that variable.

E. Hierarchical Clustering

Clustering is a method for grouping data that has similarities into a particular cluster. While, the data residing on different clusters will have a little in common (16). One clustering algorithm is Hierarchical Clustering. Hierarchical Clustering also called hierarchical cluster analysis or HCA is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types (17), (18).

The hierarchical clustering algorithm begins by assuming all data is a cluster. In the next iteration, the cluster will decrease iteratively and the closest spaced cluster will be combined until iteration finally stops when the number of clusters corresponds to the desired amount (18). Hierarchical Clustering consists of complete linked clustering, single link clustering, average link clustering, and centroid linked clustering (19). Complete linked is also known as farthest neighbor clustering. At the beginning of the process in the complete linkage, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster. At each step, the two clusters separated by the shortest distance are combined. Then, single linkage is based on grouping clusters in bottom-up fashion (agglomerative clustering), at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other (20). Next, average linkage, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. The last, Centroid Linkage Clustering is a clustering process based on the distance between its centroids. This method is good for minimizing the variance within the cluster because it involves the centroid at the time of clustering. Illustration of the process of Hierarchical Clustering method as shown in Figure 3.
3. Experimental Results and Discussions

This research used 10 types of leaves with the number of each species is 10 pieces. So, the total data used is 100 leaves images data. Testing is done by counting the amount of data that has been in the appropriate cluster. Testing will be done in the accuracy of each hierarchical clustering algorithm used.

From the results of system tests performed, the system has successfully done the process of preparing the system which has changed the image in the RGB color space into the grayscale.

![Figure 4. RGB to Grayscale](image1)

The grayscale image then processes the median filter with the aim of removing noise from the image. An example of image improvement results with a median filter is shown in Figure 5.

![Figure 5. Improvements with Median Filter](image2)
The process of binarization of grayscale images is converted to black and white images. The image of the binarization results is shown in Figure 6.

![Figure 6. The process of image binarization](image)

The next stage is the erosion process. This process aims to improve the image resulting from the binarization process. This process is done by removing the structure of the leaf so that a clear objective is obtained. The image of the results of improvements with the erosion process is shown in Figure 7.

![Figure 7. Improvements with Erosion Filter](image)

Then the inverse process is carried out to get the black and white image is shown in Figure 8.

![Figure 8. Inverse](image)

After obtaining the edge detection of the leaf image and then extraction feature for the next process, the feature extraction stage has succeeded in producing a feature vector containing seven values from seven moments invariant. The result of the extraction process is then processed normalized at intervals between 0 and 1. The feature extraction stage has succeeded in producing a feature vector containing seven values from the invariant moment. Examples of values on feature extraction generated by the system are shown in Table 1.

| Data | Seven Moment Invariant |
|------|------------------------|
|      | Hu1   | Hu2   | Hu3   | Hu4   | Hu5   | Hu6   | Hu7   |
| 1    | 0.012631 | 0.063516 | 1.09E-06 | 0.002117 | -1.4E-06 | 0.034993 | 0.670392 |
| 2    | 0.027894 | 0.194553 | 1.31E-05 | 0.011321 | -5.9E-06 | 0.043663 | 0.678573 |
| 3    | 0.029353 | 0.051733 | -6.7E-08 | -0.0024 | -2E-06 | 0.042272 | 0.686155 |
| 4    | 0.00735 | 0.035482 | 3.28E-07 | -0.00053 | -4.7E-07 | 0.045965 | 0.679213 |
| 5    | 0.010419 | 0.183735 | 8.02E-06 | 0.017604 | -5.6E-07 | 0.050373 | 0.687206 |

After normalization process of data hence obtained data feature of herbal leaves. From this data will be a clustering process using hierarchical clustering algorithms. The method used is complete link clustering, single link clustering, average link clustering and centroid linked clustering. The four methods will be comparative analysis to determine the accuracy of the method.
The test results to measure the quality of clustering will be shown by showing the number of detected members correctly in a cluster. The following is the correct number of members of the ten clusters available for the hierarchical clustering algorithm. The clustering test results are shown in Table 2.

**Table 2. The result of Hierarchical Clustering**

| Cluster Group | Complete | Single | Average | Centroid |
|---------------|----------|--------|---------|----------|
| 1             | 9        | 10     | 10      | 10       |
| 2             | 8        | 9      | 10      | 10       |
| 3             | 4        | 5      | 5       | 6        |
| 4             | 9        | 8      | 9       | 8        |
| 5             | 9        | 8      | 9       | 9        |
| 6             | 7        | 6      | 7       | 7        |
| 7             | 9        | 9      | 10      | 9        |
| 8             | 8        | 8      | 8       | 8        |
| 9             | 10       | 10     | 10      | 9        |
| 10            | 8        | 9      | 9       | 9        |
| Total         | 81       | 82     | 87      | 85       |

Test results from the dataset of 100 data using Complete Linkage Clustering method 81 data that can be grouped perfectly. So it can be concluded that the accuracy of the clustering process is 81/100 = 81%. Results using Single Linkage Clustering method yielded 82 data that can be grouped perfectly. So it can be concluded that the accuracy of the clustering process is 82/100 = 82%. The result using the Average Linkage Clustering method yields 87 data that can be grouped perfectly. So it can be concluded that the accuracy of the clustering process is 87/100 = 87%. Results using Centroid Linkage Clustering method yield 85 data that can be grouped perfectly. So it can be concluded that the accuracy of the clustering process is 85/100 = 85%. The graph of the trial results as shown in Figure 9.

![Result of accuracy](image)

**Figure 9. Result of accuracy**

Based on the experimental results obtained the best accuracy of the Average Linkage Clustering algorithm with a value of 87%. This is because this method is the only clustering method that calculates any spacing between the points in determining the order of forming a cluster.

4. Conclusions
A method for leaf classification has been developed. From the results of research that have been done include the process of preprocessing, feature extraction and clustering can be drawn conclusions, among
The test results with the greatest accuracy of the Average Linkage Clustering algorithm has a value of 87%. This proves that the average method successfully optimizes the result by calculating any distance between points of the cluster. The test result with the smallest accuracy is Complete Linkage Clustering with 78% value.

5. Reference

1. S. Wu, F. Bao, E. Xu, Y.-X. Wang, Y.-F. Chang, and C.-L. Shiang, “A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network,” 2007, doi: 10.1109/ISSPIT.2007.4458016.
2. A. Kadira and L. Nugroho, “Leaf Classification Using Shape, Color, and Texture Features”, International Journal of Computer Trends and Technology,” Int. J. Comput. Trends Technol., 2011.
3. L. Fu et al., “Banana detection based on color and texture features in the natural environment,” Comput. Electron. Agric., vol. 167, p. 105057, Dec. 2019, doi: 10.1016/j.compag.2019.105057.
4. J. Yue, Z. Li, L. Liu, and Z. Fu, “Content-based image retrieval using color and texture fused features,” Math. Comput. Model., vol. 54, no. 3–4, pp. 1121–1127, Aug. 2011, doi: 10.1016/j.mcm.2010.11.044.
5. S. Theodoridis and K. Koutroumbas, “Pattern Recognition,” in Pattern Recognition (Fourth Edition), Boston: Academic Press, 2009.
6. Arun D Kulkarni, “Artificial Neural Networks for Image Understanding,” New York: Van Nostrand Reinhold, 1994. https://works.bepress.com/arun-kulkarni/28/ (accessed Jun. 29, 2020).
7. M. Ben Haj Rouma, J. Žunić, and M. C. Younis, “Moment invariants for multi-component shapes with applications to leaf classification,” Comput. Electron. Agric., vol. 142, pp. 326–337, 2017, doi: 10.1016/j.compag.2017.08.029.
8. G. Tjitrosoepomo, Morfologi Tanaman. Yogyakarta: Gadjah Mada University, 2005.
9. C. A. Priya, T. Balasaravanan, and A. S. Thanamani, “An efficient leaf recognition algorithm for plant classification using support vector machine,” in International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), 2012, pp. 428–432, doi: 10.1109/ICPRIME.2012.6208384.
10. Z. Husin, A. Shakaﬀ, A. Aziz, and R. Farook, “Embedded portable device for herb leaves recognition using image processing techniques and neural network algorithm,” Comput. Electron. Agric., vol. 89, pp. 18–29, 2012, doi: 10.1016/j.compag.2012.07.009.
11. F. Liantoni, R. Indra Perwira, S. Muharom, R. Agung Firmansyah, and A. Fahruzi, “Leaf classification with improved image feature based on the seven moment invariant,” J. Phys. Conf. Ser., vol. 1175, p. 012034, Mar. 2019, doi: 10.1088/1742-6596/1175/1/012034.
12. X. Bai, X. Li, Z. Fu, X. Lv, and L. Zhang, “A fuzzy clustering segmentation method based on neighborhood grayscale information for deﬁning cucumber leaf spot disease images,” Comput. Electron. Agric., vol. 136, pp. 157–165, Apr. 2017, doi: 10.1016/j.compag.2017.03.004.
13. Z. Wu et al., “Application of image retrieval based on convolutional neural networks and Hu invariant moment algorithm in computer telecommunications,” Comput. Commun., vol. 150, pp. 729–738, Jan. 2020, doi: 10.1016/j.comcom.2019.11.053.
14. C. Singh, S. K. Ranade, and K. Singh, “Invariant moments and transform-based unbiased nonlocal means for denoising of MR images,” Biomed. Signal Process. Control, vol. 30, pp. 13–24, Sep. 2016, doi: 10.1016/j.bspc.2016.05.007.
15. D. Singh and B. Singh, “Investigating the impact of data normalization on classiﬁcation performance,” Appl. Soft Comput., p. 105524, May 2019, doi: 10.1016/j.asoc.2019.105524.
16. K. Tian, J. Li, J. Zeng, A. Evans, and L. Zhang, “Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm,” Comput. Electron. Agric., vol. 165, p. 104962, Oct. 2019, doi: 10.1016/j.compag.2019.104962.
17. B. Everitt, S. Landau, M. Leese, and D. Stahl, Cluster Analysis. West Sussex, U.K: Wiley: Chichester, 2011.
18. L. Rokach and O. Maimon, “Clustering Methods,” in Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA, 2005, pp. 321–352.
19 G. Goujon, Chaoqun, and W. Jianhong, *Data Clustering Theory, Algorithms, and Applications*, Virginia, 2007.

20 F. Ros and S. Guillaume, “A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise,” *Expert Syst. Appl.*, vol. 128, pp. 96–108, Aug. 2019, doi: 10.1016/j.eswa.2019.03.031.