Herbal Plant Analysis Based on Leaf Features using K-Means Clustering

Sofianita Mutalib1,3, Nur Hanani Hasbullah1, Shuzlina Abdul-Rahman1,3, Mohd Razif Shamsuddin1,3, Ariff Md Ab Malik2,3

1Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor.
2Faculty of Business and Management, Universiti Teknologi MARA Cawangan Selangor, 42300 Puncak Alam, Selangor.
3Research Initiative Group Intelligent Systems, Universiti Teknologi MARA, 40450 Shah Alam, Selangor.
Email: sofianita@uitm.edu.my

Abstract: Plants are essential in the Earth, as it supplies the oxygen needed by human beings and animals, and becomes the source of foods and medical treatments. Many medicinal plants can treat diseases and it is also called herbal plants. Traditionally, these plants are processed and transformed as traditional medicines to cure any diseases. Nowadays, there are still practices that use medicinal plants. However, it is quite challenging to find herbal plants and these herbal plants come with different features such as size, shape, and colour. Therefore, this paper presents a machine learning approach, namely clustering, to classify the herbal plant species through images. We focused on six herbal plants in Malaysia which are Peacock Fern, Misai Adam, Mempisang, Tapak Sulaiman, Pandan Serapat and Kacip Fatimah. These species were collected from Taman Negara Pahang, Kuala Keniam, Malaysia. The k-means algorithm was employed by experimenting with several numbers of clusters in the range of two, three, four and five. The features extracted were colour and shape and this was performed using Python libraries. This study would benefit the researchers or botanists to identify the plant’s name based on the features. In order to give more interactive elements, a dashboard was developed in which the herbal plants are categorized in the group that has similar characteristics.

Keywords: Clustering, k-means, herbal plants, colour, shapes, Taman Negara Pahang

1. Introduction

Plants play an important role as a source for a human being such as oxygen to breathe, food, medicine, and shelter. Plants exist everywhere on the earth; it brings us a good life to live healthily and in good condition. Plants also give the animal a safe place to stay and hide from their enemies. Apart from this, the plants are used for various purposes, such as for medical purposes, ingredients in food, and agriculture industry. However, identifying plant species can be very difficult even though we see the plants every day. It is because people who do not have any knowledge regarding plants find it difficult to recognize the plant species [1]. Only those who are experts, can identify it correctly [2] due to inconsistent size or colours. So, it is difficult for people to identify herbal plants if they have less knowledge about the herbs. Though, plant species can be identified automatically and there are many attempts by researchers in developing a system that can recognize the plant species. Meanwhile plants
have many characteristics, such as flowers, leaves, seed, branch, and etc with specific features, for example, leaf shape, colour, vein structure, and texture. Basically, for the identification purpose, many methods or techniques of machine learning were explored. Hence, this study aims to identify the herbs plants based on leaf features using a clustering algorithm, namely k-means which is an unsupervised method. The data will be collected from Taman Negara Pahang.

Herbs are plants with savoury or aromatic properties that are used for ingredients in food, medical purpose, or fragrance. It is very well known because the herb plant acts as traditional medicine and is used by people to cure the disease. For example, Indonesian people typically use herbs to cure all kinds of ailments as medicine. It is considered as not expensive and has no unpleasant effects on people [3] [1]. One of the Sustainable Development Goals (SDG) and national heritage in Malaysia in Taman Negara, located in Pahang. The area of Taman Negara is around several hectares, and is enriched by flora and fauna from special tropical species that could be found only in this golden land rainforest. Plant species as reported by [4] [2] represent 116 genera and 44 families. Significantly the National Park is recognized for its high number of plant species [5] [3]. The next section 2 presents related studies while in Section 3 the Methodology is presented. Section 4 presents the result and finally the conclusion is drawn in Section 5.

2. Features and Unsupervised Learning

2.1. Leaf Features

According to [6], leaves are the most important part in plants compared to other parts like flowers, stems etc. The leaves hold a lot of information, and it has a reliable feature. The features that can be extracted from the leaves are shape, margin, colour or texture, which are often considered as special to identify the plants [7]. Based on other research articles, most of the research did feature extraction based on the shape and colour of the leaves. Each leaf has different shape features because it has different geometry of the leaves [8], [9]. The length and breadth can be calculated through the ratio of the leaves. Among important steps to classify the leaf image are 1) image acquisition, 2) image processing including conversion of gray-scale images, conversion of binary images, image segmentation and smoothing, 3) extract multiple feature vector matrices and lastly 4) clustering or classification [10,11]. [9] use Digital Morphological Features and Geometrical Features to extract the features from the leaf such as margin. Meanwhile, other research has analyzed the performance of the classification of leaf based on semi-supervised fuzzy clustering [7] with multiple features such as texture, shape, margin and combination features for leaf classification analysis from collected images at Wollaton Park, Nottingham United Kingdom. In the end, the result gives similar images in the same cluster/class while the different images will be categorized into different clusters/classes based on the measure or certain distance.

2.2. Unsupervised Learning

Machine Learning requires adaptive mechanisms that enable computers to learn by experience, example and by analogy [12]. It also can make a computer change or adapt their acts so that all these actions will get more precise according to [13]. Among techniques that fall under machine learning are artificial neural networks, support vector machine and genetic algorithms [14] which represent available methods such as reinforcement learning, supervised learning, evolutionary learning and unsupervised learning. Unsupervised learning is known as “learning without teacher” [15] which means the data has no labels and the process of finding the same objects apply the similarity distance functions or known as clustering. Clustering will find groups of objects in other groups which are similar to each other or different from objects [16]. Hierarchical clustering algorithms repetitively find nested clusters beginning with a data point which successively merges the most connected cluster data pairs [17]. Meanwhile, partitional clustering differs in comparison to hierarchical clustering by portioning data eight that is not imposed on hierarchical structure [17]. Based on [18], “k-means is the most popular and simplest
unsupervised algorithm in machine learning.” The process starts in defining the number of cluster \( k \) in the first step [19]. Next, is to shuffle the data set to initialize centroids and select \( k \) data points for the centroids without changing [20]. The distance of each data point to the centroids will be calculated to form the clusters. The last step is to keep iterating till centroids do not change [19]. Other than that, k-means can very easily cluster large data points [17,15] and k-means was successfully applied in various research [21].

3. Research Methods

In research methods, there are several phases which are data collection, image pre-processing, feature extraction, clustering application, clustering evaluation and analysis.

3.1. Data Collection

In this stage, the images of herbal leaves were collected from Kuala Keniam, Taman Negara Pahang. This trip to Taman Negara was held for five days in conjunction with the Second Taman Negara Scientific Expedition program by Universiti Teknologi MARA. This trip was conducted from 4th September until 8th September 2020. The images were captured using a high-quality digital camera which are Sony a6500 and smartphones (iPhone 8/ Huawei P30). There are six species of herbal plants that contain 390 images of leaves. Table 1 shows the image of each species of herbal plants.

| Image | Herbal Plant’s name | Peacock Fern | Kacip Fatimah @ Labinia pumila Benth | Tapak Sulaiman @ Elephantopus scaber L | Misai Adam @ Actinostachys digitata | Mempisang @ Stelechocarpus cauliflorus | Pandan Serapat @ mapania cuspidata var. cuspidata |
|-------|---------------------|--------------|-------------------------------------|----------------------------------------|-------------------------------------|----------------------------------------|-------------------------------------------------|
| Total |                     | 7            | 65                                  | 77                                     | 81                                  | 104                                    | 50                                              |

3.2. Image Pre-Processing, Colour Extraction dan Comparison Dataset Construction

Image pre-processing is the most important phase to clean all the unwanted data and to transform the images into more readable format by the clustering algorithm and feature extraction. The processes involved shown in Figure 1 are image resizing, image/data augmentation, grayscale conversion from RGB, smoothing the image using Gaussian Blur and finally binary conversion from grayscale (for shape extraction). This study extracted two main features, which are colour features and shape features. The color features extracted are RGB Mean, which are the values of average colour from the image and RGB values and then the shape of the leaf with the coverage area of rectangularity. Next, five main datasets were constructed with the set of extracted features; 1) Red, Green, Blue (RGB) Mean, 2) Rectangularity, 3) RGB Mean and Rectangularity, 4) RGB pixel values and lastly 5) RGB pixel values.
3.3. Feature extraction for Shape

The value of the leaf shape was extracted in contour features (the area of the leaf). Contour features applied to find the difference of contours like area, perimeter bounding box, etc. The result of drawing the bounding box in the area of the object shown in Figure 2. From the bounding box, the area for rectangularity was calculated. The formula is:

$\text{Rectangularity} = \frac{\text{width} \times \text{length}}{\text{Leaf Area}} \quad \text{Equation (1)}$

Leaf area is the values that were obtained from the contour area before. Next, after gaining the results of rectangularity, the values were put into the CSV dataset as shown in Figure 3.

3.4. Application of Clustering Algorithm for Unsupervised Learning

The datasets are ready for the clustering by using k-means algorithm after the image processing is done. Besides, two distance functions were used to get the different results between those clusters which are a) Euclidean and b) Manhattan distance in Equation 2. Then, the output for clustering is compared and evaluated in determining leaf or which plant species are categorized in the same group. The number of clusters was set into $k = 2, 3, 4$ and 5 and analysis was done to compare the clusters formed.

$$d_{(x,y)} = \sum_{i=1}^{n} (x_i - y_i)^2 \quad \text{a) Euclidean}$$

$$d_{(x,y)} = \sum_{i=1}^{n} |x_i - y_i| \quad \text{b) Manhattan} \quad \text{Equation (2)}$$
3.5. Clustering Evaluation and Analysis

In this phase, the clustering algorithm was evaluated based on the sum of squared errors for each, Euclidean distance and Manhattan distance. In addition, the elbow method will be used to find the optimal number of clusters. When the elbow effect shows any conflict in the result, the Silhouette method is applied to find the optimal number of clusters. The silhouette method computes the coefficients of each point that measure how much a point relative to other clusters is close to its cluster. Supplement analysis is provided in stack bar chart graphs which generate the results through visualization.

4. Results and Findings

4.1. Image Pre-Processing Results

The preprocessing is to reduce the noise and make the data readable by the algorithm. The captured data might be in different sizes and need to be resized or resized. Data is also augmented for peacock fern because it is originally only seven images, so after preprocessing there are 18 samples while other plants remain the same number as in Table 1.

4.2. Comparison dataset for the elbow effect

Five datasets, as in Table 2 were tested using several numbers of clusters with different distance functions. The study applies two different distances which are Euclidean distance and Manhattan distance in the k-means algorithm for each dataset.

| Dataset | Features | Values \((x_1, \ldots, x_n)\) | Evaluation |
|---------|----------|----------------------------|------------|
| 1       | RGB Mean | The average of RGB \((x_1)\) | Elbow      |
| 2       | Rectangularity | The rectangle area value \((x_1)\) | Elbow and silhouette |
| 3       | RGB Mean Vs Rectangularity | The average of RGB \((x_1)\) and the rectangle area value \((x_2)\) | Elbow and silhouette |
| 4       | RGB Values | RGB values for every pixel in an image with \(999119\) values \((x_1, \ldots, x_{999119})\) | Elbow |
| 5       | RGB Values Vs Rectangularity | RGB values for every pixel in an image with \(999119\) values \((x_1, \ldots, x_{999119})\) and the rectangle area value \((x_{999120})\) | Elbow |

Table 3 shows the result of sum squared error for different numbers of clusters. The comparison was done between these distances in finding the optimal number of cluster \(k\), and also by performing visualization of the clustering. The evaluation was done using Elbow effect, as shown in Figure 4 and Silhouette method (for any ambiguous elbow effect result). For dataset RGB values and RGB Values Vs Rectangularity, no elbow effect was shown on the graph.
Table 3. Result for RGB Mean datasets

| Dataset          | Distance | k = 2  | k = 3  | k = 4  | k = 5  | Optimal cluster based on Elbow effect | Silhouette method? |
|------------------|----------|--------|--------|--------|--------|--------------------------------------|-------------------|
| RGB Mean         | Euclidean| 604.21 | 543.53 | 541.67 | 460.4  | k = 3                                |                   |
| RGB Mean         | Manhattan| 637.75 | 576.87 | 573.3  | 491.49 | k = 3 or k = 4                       |                   |
| Rectangularity   | Euclidean| 622.93 | 518.16 | 469.1  | 409.41 | k = 3                                | Yes               |
| Rectangularity   | Manhattan| 637.31 | 532.78 | 483.55 | 424.51 | k = 3 or k = 4                       |                   |
| RGB Mean and     | Euclidean| 607.95 | 547.3  | 545.39 | 464.00 | k = 3 or k = 4                       | Yes               |
| Rectangularity   | Manhattan| 657.8  | 596.89 | 593.28 | 511.14 | k = 3 or k = 4                       |                   |

Figure 4. Sample of elbow effect for RGB Mean using different distances based on sum of squared error

![Elbow method - Euclidean vs Manhattan](image)

a) Euclidean Distance

b) Manhattan Distance

Generally, the distance will be increasing when more variables are presented in the dataset. When dimensions are increasing the errors are increasing too. The elbow effect for RGB Mean of pixel value using Euclidean and Manhattan distance produces a clear effect when $k = 3$, so the optimal cluster number is 3 for the dataset. Meanwhile, the rectangularity value shows an ambiguous effect between $k = 3$ and $k = 4$ for both distance functions. However, $k=4$ also can be used since the point is not very far with cluster 3. So, the silhouette method was applied in this dataset. The combination of RGB mean value and rectangularity also shows an ambiguous effect. Manhattan distance gives clearer elbow effect for high dimensionality data (RGB values) with the most optimal number of clusters is 3. Last dataset for RGB Values and rectangularity, the clearer effect was for Manhattan distance. The elbow method for Euclidean distance is difficult to choose because the graph is a straight line. Therefore, it cannot determine which one is the optimal number for Euclidean distance. Based on the results of the elbow method for Euclidean, observation shows the optimal point is at cluster = 3 for the datasets with colors features and for the Manhattan distance, the optimal point is whether cluster = 3 or 4. So, further evaluation was done using the Silhouette method for rectangularity as shape feature in determining the optimal number of clusters.
4.3. Comparison dataset for the elbow effect using Silhouette Method

Two datasets were run for the Silhouette method due to confusion of the elbow method optimal cluster selection. The dataset 2 and 3, produce an ambiguous choice between cluster = 3 and cluster = 4. The dataset 5, also had the shape feature with high dimensionality, but due to the limited computational power, the method cannot be run. It is difficult to choose which is a better choice for the number of clusters. So, another solution which is using the Silhouette method was plotted and, cluster = 3 can be considered as the optimal number of clusters as the elbow effect shown in the graph.

4.4. Bar Chart Analysis

Bar chart is a visualization tool that represents the count of samples for each predetermined range for numeric variables or for each group for categorical variables. For this task, the chosen number of clusters is 3 for both Euclidean and Manhattan distance in each dataset. According to the stack bar chart in Table 4, there is a difference of distribution of samples when applying Euclidean and Manhattan distance. It is clearly shown that Mempisang, Tapak Sulaiman and Kacip Fatimah stay in the same groups for dataset 1 – RGB mean values. While the other plants some of it is placed in every cluster 1, 2 and 3. Meanwhile, rectangularity (dataset 2), shows consistent outcome for Mempisang in cluster 3, Misai Adam in cluster 2 and Tapak Sulaiman in cluster 1. For Kacip Fatimah and Peacock fern samples are distributed in all three clusters.

Table 4. Sample of bar chart for dataset 1 with 3 clusters for Euclidean and Manhattan

| Dataset 1 | k=3: Euclidean | k=3: Manhattan |
|-----------|----------------|----------------|
| RGB Mean  | Category of plant by cluster | Category of plant by cluster |

The most consistent samples in clustering are Mempisang, Tapak Sulaiman, Kacip Fatimah and Misai Adam. All these species are grouping in the same group. It can be assumed that these plants have similar features which are colour and shape features. More pages in the dashboard were provided to see the details of clustering results in scatterplot and also for the elbow effect.

5. Conclusion

This paper discusses the application of k-means algorithms for analysis of herbal plant species which is to find the similarity between different leaves based on features, which are colour and shape. The k-means algorithm is simple to apply, and it also can handle the large dataset as shown in dataset 4 and 5. Besides, the process is very easy and would require users to specify the number of $k$. In determining the optimal number of clusters, the elbow effect was studied based on the sum squared of errors for Euclidean and Manhattan, and Silhouette method for any ambiguous results. The overall results give $k = 3$ as the optimal number of clusters, for most of the datasets. We could see Euclidean distance might
be suitable for a smaller dataset, meanwhile Manhattan would be more appropriate for a high dimensional dataset. So, this paper also discussed the distribution analysis using a bar chart for each plant leaf. Some of the plants have stayed in the same group with other plants, this is due to their similarity of the features. Obviously, Mempisang is consistent in the same cluster through colour and shape features, meanwhile Tapak Sulaiman leaves nearly give similar behaviour. Meanwhile, Kacip Fatimah leaves are clustered together by using the RGB mean values and also rectangularity values (shape features). Misai Adam leaves are in the same cluster when applying rectangularity and RGB values in separate analysis. The outcome of the study is to help the researchers or botanists that are unable to recognize the plant's species because of their different size or shape and sometimes due to changeable colour. Future research is to improve the feature extraction method and add more features. Any of the limitations and problems will be improved in the field of image processing for future work.

Acknowledgement
The authors would like to express the gratitude to Pejabat Timbalan Naib Canselor (Penyelidikan & Inovasi for sponsoring our teams in the 2nd Scientific Expedition Taman Negara Pahang, Institute for Biodiversity and Sustainable Development (IBSD) for the research grant 600-RMC/DANA 5/3/MITRA SDG (1/2019)-42 and Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support.

References
[1] R. Rizal I, Achmad H, Ajub A Z, Eskanesiari, Aditya I B, and Risma S 2017 Herb Leaves Recognition Using Combinations of Hu’s Moment Variants-Backpropagation Neural Network and 2-D Gabor Filter-Learning Vector Quantization (LVQ) Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017 2018-Janua:219–24.
[2] Surleek K and Prabhpreet K 2019 Plant Species Identification Based on Plant Leaf Using Computer Vision and Machine Learning Techniques J. of Multimedia Information System 6(2):49–60.
[3] Sabu A, Sreekumar K and Rahul R N 2018 2017 Recognition of Ayurvedic Medicinal Plants from Leaves: A Computer Vision Approach 4th International Conference on Image Information Processing, ICIIP 2017 2018-Janua:574–78
[4] Mohd Nazip S 2012 Tree Species Diversity and Forest Stand Structure of Pahang National Park, Malaysia. Biodiversity Enrichment in a Diverse World, 473-92
[5] Jambari A, Selvadurai S, Hazril Rafhan A H, Khairul Amirin M, Adham A-B, Alex L, Badrul A 2019 Quantifying species richness and composition of elusive rainforest mammals in Taman Negara National Park, Peninsular Malaysia. Global Ecology and Conservation, 8 e00607
[6] Ionel Bujorel P, Robert A, Claudia M E, and Andrei V 2017 Important Shape Features for Romanian Medicinal Herb Identification Based on Leaf Image. E-Health and Bioengineering Conference, EHB 2017 599–602.
[7] Li L, Jonathan M G, and He D 2015 Leaf Classification Using Multiple Feature Analysis Based on Semi-Supervised Clustering. J. of Intelligent and Fuzzy Systems 29(4):1465–77.
[8] Venkataraman D and Mangayarkarasi N 2017 Computer Vision Based Feature Extraction of Leaves for Identification of Medicinal Values of Plants. 2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016 7–11
[9] Arunpriya C, and Antony S T 2014 A New Leaf Analysis And Clustering For Tea Species Identification International J. Of Computer Engineering & Technology (IJCET) 2014: 88
[10] Mohd Razif S, Zaidah I and Azlinah M 2008 Slant Classification Using FuzzySIS, 2008 Third International Conference on Convergence and Hybrid Information Technology, 1, 1080
[11] Muhammad Naim M, Shuzlina Abdul-Rahman and Sofianita M 2020 Pre-processing methods for vehicle lane detection, ESTEEM Academic J. 16, 74
[12] Micheal N 2011 Artificial Intelligence: A Guide to Intelligent Systems (3rd Edition) (3rd ed.).
Pearson Education Canada.

[13] Marsland S 2014 Machine Learning: An Algorithmic Perspective, Second Edition, Chapman and Hall/CRC.

[14] Sofianita M, Mohd Hafiz A, Shuzlina Abdul-Rahman, Zalilah A A 2016 A brief study on paddy applications with image processing and proposed architecture, IEEE Conference on Systems, Process and Control (ICSPC) 124

[15] Trevor H, Robert T and Jerome F 2009 Springer Series in Statistics. The Elements of Statistical Learning 27(2):83–85.

[16] Anil K J 2010 Data Clustering: 50 Years beyond K-Means. Pattern Recognition Letters 31(8):651–66.

[17] Valliammal N and Geethalakshmi S N 2012 Plant Leaf Segmentation Using Non Linear K Means Clustering. International J. of Computer Science Issues 9(3):212–18.

[18] Medium 2020 Understanding K-Means Clustering In Machine Learning. Received June 21, 2020, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[19] Medium 2020 K-Means Clustering: Algorithm, Applications, Evaluation Methods, And Drawbacks. Received June 21, 2020, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[20] Noor M A, Yasmin M K, and Kannappan P 2018 Unsupervised Learning Method for Plant and Leaf Segmentation Proceedings - Applied Imagery Pattern Recognition Workshop 2017-Octob:0–3

[21] Sofianita M, Siti Nur Fadlun J, Shuzlina Abdul-Rahman and Azlinah M 2010 Soil classification: An application of self organising map and k-means. 10th International Conference on Intelligent Systems Design and Applications, 439