Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Exploring the Application of K-means Machine Learning Algorithm in Fruit Classification

Weidan Chen

Information Technology Faculty, Monash University, VIC 3168, Melbourne, Australia

*Corresponding author: wrh_41@163.com

Abstract. Fruits are very popular in people's lives because they contain vitamins and dietary fiber, making them an important source of human diet. According to fruit production statistics, global fruit production reaches millions of metric tons, so it is necessary for people to establish an advanced fruit recognition system. However, traditional methods, including physical inspect, are inefficient, labor-intensive, and have a high probability of errors. Based on these facts, developing more efficient classification algorithms is worth exploring, as they can help people classify fruits quickly and effectively. This paper describes an experiment using K-means machine learning algorithm in Python, aiming to classify different types of fruits using image datasets containing various types of fruit images. The K-means algorithm will cluster the fruit image data input from three apple images and three banana images samples as an efficient clustering algorithm. The conclusion that the success rate for apple is 33.3% and the same rate for banana is 66.7% can get from the experiment. Nevertheless, this does not mean K-means will classify banana better than apple absolutely, since many other factors like the fruit color ,image pixel, the unique feature of randomly assigning labels to clusters in the k-means algorithm can also lead to the success or failure of the experiments.

Keywords: K-means; Machine Learning; Fruit Classification.

1. Introduction

The emergence of digital imaging technology and the advancement of machine learning have opened up new avenues for the automation of complex tasks in agriculture. Accurate and efficient fruit classification is crucial in optimizing the post-harvest processing process .

Traditionally, fruit classification relies on manual inspection, which is time-consuming, subjective, and prone to inaccuracies and inconsistencies. With the increasing demand for precision agriculture and sustainable food production systems, there is an urgent and necessary need for automated solutions that can quickly and accurately classify fruits.

Nowadays there has been an interest in utilizing machine learning methods to construct models for fruit classification,The advantages of machine learning algorithms in fruit classification are becoming increasingly apparent. In 2023, three scholars in Indonesia called Ahmad, Muhamad and Riyan had done research on classify maturity level of bananas based on color and texture characteristics of the banana image [1]. Gray Level Co-Occurrence Matrix, K-Means Clustering, K-Nearest Neighbor are the methods used in research. Another method, CNN had been used by three authors includes Mauricio Rodriguez to classify the maturity level of the fruit like strawberries and green-apple [2].Three Thai scholars, Denchai Panarit and Surasak, have also conducted similar studies on mangoes, which also predict the maturity of mangoes [3]. Their research methods includes support vector machines (SVM), random forests (RF), and k-nearest neighbors.SVM is based on supervised learning algorithm [4].

Decision trees, statistical-based algorithms such as Bayesian and Regression, distance-based algorithms, and artificial neural networks are also commonly used classification techniques.[5]

Except these methods,K-means algorithm is the most basic and popular unsupervised machine learning clustering algorithm[6]. It is also an effective machine learning method can be applied in fruit classification since its significant advantage ,like relatively simple and computationally efficient ,

especially for large datasets. K-means can handle high-dimensional feature spaces and a large number of samples without requiring significant computational resources.

The primary aim of this article is to investigate and develop an unsupervised fruit classification system using the K-means clustering algorithm implemented in Python by the given dataset. K-means clustering operates by dividing the dataset into K clusters, where each observation belongs to the cluster with the nearest mean (centroid). In the context of fruit classification, this can be transformed into grouping fruit images based on shared colors, sizes, textures, or other visually distinguishable features extracted from image data. In this article the color and size has been selected.

This article explores an unsupervised machine learning method, especially the K-means clustering algorithm, for the implementation and effectiveness of classifying fruits based on their visual characteristics such as color or shape when applied in a Python programming environment, the performance of cluster will also be evaluated and the the limitation of K-means algorithm will be discussed. Furthermore, the comparison of classification accuracy and robustness of the K-means method with baseline or alternative unsupervised methods will also be explained.

2. Data and methodology

2.1. Datasource

The data source used in this project is from Kaggle, called “Fruits and Vegetables Image Recognition Dataset”.The images for this dataset were scraped from Bing Image Search on 03/11/2020.

The dataset contains the image items from 9 specific categories of fruits: apple,banana,grapes, orange, pear, lemon, pineapple, watermelon and mango.It contains three folders: train(100 images each subfolder for each category of fruits),test(10 images each subfolder for each category of fruits), and validation(10 images each for each subfolder for each category of fruits).

This data source is an excellent source for this project because it is not only easy to download, but also provides specific visual feature information for fruit classification.Image data are very effective in conveying certain information and contributing to certain evaluation attributes, such as shape, color, and so on [7]. Python has a rich machine learning library ecosystem (such as NumPy, Pandas, OpenCV, and scikit learn), which makes it an ideal choice for implementing a fruit classification system using K-means.

2.2. Data Preprocessing

Before implementing the K-means algorithm in python , some steps of data preprocessing must be done.Preprocessing the fruit image data is crucial for several reasons:

(1) Standardization and normalization: The pixel values of the original image may vary greatly due to factors such as lighting conditions, camera settings, and color configuration files. Preprocessing helps standardize these values, ensuring pixel intensity is within a consistent range.

(2) Image size adjustment: Images in the dataset may have different resolutions or aspect ratios. Adjusting the image size to a fixed size ensures that all input data has the same dimension, making it compatible with machine learning models.

In the data preprocessing process,a function called load_and_preprocess_images was used, there are two arguments in this function, the first one is image_paths, the second one is img_size. The argument image_paths is list of file paths pointing to the fruit images to be processed,it is a variable that holds a list of strings representing the file paths to the fruit images that will be processed and classified.

The fruits images loaded are from dataset, within a folder called fruit_dataset. The number in image_paths is not fixed, it can be 4 images, 5 images or more. The second argument is img_size,which is an optional tuple representing the desired output size (height, width) for the resized

images. The default value is (150, 150).The cv2.resize(img, img_size) function adjusts the loaded image to the specified size (img_size). Beside that ,in the load_and_preprocess_images function, the images are also converted most commonly used RGB format by cv2.cvtColor(img, cv2.COLOR_BGR2RGB) function. Then the final function flatten() flatten the image .The flattened image pixel values are divided by 255.0 to normalize them. The preprocessed image then appended to an empty array list “features”.

2.3. Models and algorithms

K-means is a unsupervised machine learning algorithm used for clustering, starts by first defines the number of clusters. Each cluster is identified using its centroid, which is the average value of all data points in the cluster. Then each data object is assigned to its nearest centroid while minimizing the distance between objects.

After data preprocessing,K-means clustering algorithms are performed to loaded images, the objective is to The main package used in Python includes pandas and sklearn. The number of clusters (n_clusters) are defined by the number of the unique fruit types For example if two apple images and 1 banana image are loaded, the number of clusters will be two. If one apple image, one banana image and one grape image are loaded the number of clusters will be three. After running the K-means algorithm, each data point is assigned a cluster label, which is an integer representing the cluster to which that data point belongs. These labels typically range from 0 to (n_clusters - 1).So the number of different types of fruit images determines the number of clusters, which in turn determines the labels. Assuming the number of clusters is 3, the cluster labels will be integers 0, 1, or 2. The proper number of n_clusters need to be set because the performance of K-means algorithm will also be affected by n_clusters [8].

The final step for K-means algorithm is to map cluster labels to fruit class labels in a list called fruit class labels which contain 9 types of fruits. At this step,each fruit image loaded can be classified as one fruit type in fruit class labels list. The outcome predict fruit class list index is the labels in cluster labels list. If the input fruit image (cluster label) matches the output fruit class label, it can be assumed that the experiment is successful, otherwise it is considered a failure. From this, it can be seen that the K-means algorithm has certain limitations in fruit classification. In a study using the k-means algorithm for apple image segmentation, k-means once mistakenly identified the leaf pixels as apples [9].

3. Result and Discussion

In experiment,3 different apple images and 3 different banana imaged are loaded, Fig. 1 shows three input apple sample images, which are from fruit_dataset folder.Three apple images represents three different type of apples: one single apple , three apples combination, and a bitten apple.

Fig.2 shows three input banana sample images, also represents three kinds of bananas: one single banana, a bunch of bananas made up of four connected bananas and a cartoon banana image. Expected output is for apples to be classfied as apple, bananas are classified as banana. However,this would not always be the case.

Fig. 1 Apple Input Sample

Fig. 2 Banana Input Sample

Table 1 Input fruit category name and success rate

Input fruit category image	Classification success	Classification fail	Success rate
Apple(3 images)	1	2	33.30%
Banana (3 images)	2	1	66.70%

Table 1 is the data come from the experiment. It shows the input fruit image. In this case apple 3 images, banana 3 images. It can be any other fruit like grapes or oranges. It also indicates that the success rate of apple is only 33.3% whereas the success rate of bananas is 66.7%, which intuitively indicates that the K-means algorithm has a higher accuracy rate in identifying bananas and a lower accuracy rate for apples, but the factors affecting success rate are not solely determined by the type of fruit. Because the clustering results of the K-means algorithm are sensitive to the initial cluster center points, different initial center points(the same number with n_clusters) may lead to completely different clustering results [10]. Since the function cvtColor has been used before in data preprocessing,the input BGR (Blue-Green-Red) image has been converted to the RGB (Red-Green-Blue) color format, therefore the color of the fruit and image pixel can also be a factor affecting the success rate. Inaccuracies can caused by the low resolution of the image as well[11]. Moreover, another important factor is that the k-means labels is randomly assigned. For example, if the n_cluster is 3, the labels can be 0,1,2 or 0,2,1. When map cluster labels to fruit class labels ,this will lead to the higher probability of inconsistency between the input fruit image and output fruit category.

4. Conclusion

In this paper, the research of K-means machine learning algorithms by python in fruit classification is presented, two categories of fruits,apple and banana are used in the experiment.However, the source datasets contain not only these two types, it consists of many other types of fruit, like pear, grapes or pineapple. Thus the outcome of the experiment has certain limitations.Through the analysis of the experimental results, it was found that:

(1) The success rate of apple is 33.3% ,and the success rate of banana is 66.7% but it is not absolutely. Another factors like image pixel can also be contributed to the success or failure of the experiment.

(2) The only case in which the input fruit image will match the output classification fruit class label is when the cluster label matches the fruit class labels.

K-means algorithm has advantages for fruit classification, like being easy to implement and computationally efficient and being suitable for real-time or batch processing of fruit classification tasks.This paper demonstrated the effectiveness of K-means algorithm. However, it also has limitations like we must determine the expected number of fruit classes (n_clusters) beforehand, when the true number of classes is unknown or the data distribution is complex, this may be difficult. Convolutional neural networks (CNNs), or other deep learning architectures may also very suitable machine learning algorithms for fruit classification.

References

[1] Maskar, V., Chouhan, K., Bhandare, P., & Pawar, M. (2021, May). Clustering of Fruits Image Based on Color and Shape Using K-Means Algorithm. In Techno-Societal 2020: Proceedings of the 3rd International Conference on Advanced Technologies for Societal Applications—Volume 1 (pp. 639-650). Cham: Springer International Publishing.

[2] Yu, Y., Velastin, S. A., & Yin, F. (2020). Automatic grading of apples based on multi-features and weighted K-means clustering algorithm. Information Processing in Agriculture, 7(4), 556-565.

[3] Pham, V. H., & Lee, B. R. (2015). An image segmentation approach for fruit defect detection using k-means clustering and graph-based algorithm. Vietnam Journal of Computer Science, 2, 25-33

[4] Chithra, P. L., & Henila, M. (2017). DEFECT IDENTIFICATION IN THE FRUIT APPLE USING K-MEANS COLOR IMAGE SEGMENTATION ALGORITHM. International Journal of Advanced Research in Computer Science, 8(8).

[5] Kahfi, A. H., Hasan, M., & Hasanah, R. L. (2023). Classification of Banana Ripeness Based on Color and Texture Characteristics. Journal of Computer Networks, Architecture and High Performance Computing, 5(1), 10-17.

[6] Aherwadi, N. A. G. N. A. T. H., & Mittal, U. S. H. A. (2022). Fruit quality identification using image processing, machine learning, and deep learning: A review. Adv. Appl. Math. Sci, 21(5), 2645-2660.

[7] Zhang, C., Zou, K., & Pan, Y. (2020). A method of apple image segmentation based on color-texture fusion feature and machine learning. Agronomy, 10(7), 972.

[8] Worasawate, D., Sakunasinha, P., & Chiangga, S. (2022). Automatic classification of the ripeness stage of mango fruit using a machine learning approach. AgriEngineering, 4(1), 32-47.

[9] Patil, P. U., Lande, S. B., Nagalkar, V. J., Nikam, S. B., & Wakchaure, G. C. (2021). Grading and sorting technique of dragon fruits using machine learning algorithms. Journal of Agriculture and Food Research, 4, 100118

[10] Kutyrev, A., Kiktev, N. A., Kalivoshko, O., & Rakhmedov, R. S. (2022, November). Recognition and Classification Apple Fruits Based on a Convolutional Neural Network Model. In IT&I (pp. 90-101).

[11] Akpolat, O., & Ertürk, G. An Application of Data Mining Problem by Python: A Case Study on Fruit Classification.