Combination of k-means with naïve bayes classifier in the process of image classification

M R Wayahdi*, Tulus, M S Lydia

Department of Information Technology, Faculty of Computer Science and Information Technology, University of Sumatera Utara, Medan 20155, Indonesia

*muhammadrhifkywayahdi@gmail.com

Abstract. In this paper, the data which has been tested is the image of bananas extracted based on the colour characteristics and statistics, namely red, green, blue, hue, saturation, value, average, variance, skewness, kurtosis, and energy. Image data is extracted as much as 60 data for training data, and images for test data as many as 105 data. From the 105 image data has been tested, the number of correct test data is 89 data and the number of incorrect test data is 16 data. The total percentage obtained from the test is 85%. The results obtained are a collection of images based on the level of maturity that is mature, medium, or raw.

1. Introduction

The image can be defined as a two-dimensional function symbolized by $f(x, y)$. The intensity values of x, y, and discrete number are referred to as digital images. The field of image processing refers to processing using a digital computer, and then the images can be classified to obtain an information or new knowledge. Process of this image classification refers to artificial intelligence methods that focus on machine learning. Many other methods in machine learning are used for classification processes including K-Means and Naïve Bayes Classifier.

K-Means clustering is a grouping algorithm based on optimization of criteria function [1] which is widely used and studied in data mining [2]. While the Naive Bayes Classifier method is a group of popular statistical techniques for email filtering [3]. This method is also a very scalable probabilistic classifier. This method is quite popular for text categorization with frequency data as a feature [4]. In this paper the author will analyze the performance of the K-Means method and the Naïve Bayes Classifier in image classification. As well as combining K-Means and Naïve Bayes Classifiers to improve accuracy in the image classification process. The image used is fruit image to be classified as maturity based on colour characteristics and statistics.

2. Related Research

Xiao *et al.* use a fuzzy approach to predict fruit maturity levels based on colour features to help decision makers for farmers in maximizing profits [5]. Mulyani*et al.* in his research on the maturity classification of fuji apples with fuzzy logic using colour attributes, namely by converting RGB images to grayscale [6].

Mohd R S *et al.* in his research in detecting hotspots, proved K-Means Clustering was successful as an infrared image segmentation. By separating infrared thermal images into several layers, unrelated information on images can be removed to produce a faster and more efficient system response [7]. Ayech M W and Ziou D explained that the results obtained

from his research with the K-Means method for terahertz image segmentation can reach targets with a low sample size, this method is very useful for large-scale data collection [8]. Whereas Wang X *et al.* in his research shows the performance of the K-Means method is useful for parameter selection and grouping, but for large data sets it is less efficient. This algorithm can be improved by providing initial parameter estimates [1].

Adi A O and Celibi E in his research on the classification of 20 well-known news groups containing 20,000 documents with the Naïve Bayes Classifier method concluded that this method performed well in the classification process even though there were still deficiencies [9]. Granik M and Mesyura V use Naïve Bayes Classifier to detect false news, with this method the classification accuracy achieved is around 74% at the time of the test set which is a feasible result for a relatively simple model [3]. Some of the above underlie the conduct of this research.

3. Proposed Method

3.1. K-Means Clustering

K-Means clustering is an unsupervised algorithm that is used to form different clusters of data sets so that they can be grouped together. A cluster is a collection of similar (homogeneous) data objects in one cluster and diverse (heterogeneous) data on objects in another cluster [10]. The K-Means method is most appropriate for grouping problems. This algorithm organizes datasets in a way that can be managed well and more easily by determining the appropriate cluster center. Clusters that are declared can produce different output so that it needs a smart way to determine the cluster center [11].

K-means is a clustering algorithm based on optimizing the criteria function. If the sample data is presented as aggregate $X=\{x_1, x_2, ...,x_n\}$, x_i is a d-dimensional vector, and suppose the number of clusters is k, the initial K-means center is $C_i(0)$ [1]. The similarity measurement adopts Euclidean distance, as for α and β.

$$D = [\alpha - \beta] = \sqrt{(\alpha - \beta)^T(\alpha - \beta)} \tag{1}$$

Grouping criteria adopt the number of squared errors.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} [x - C_i]^2 \tag{2}$$

The steps in the K-Means clustering as follows: [1]

1) Initialization of parameters: specify cluster k and center, initial K-Means $C_i(0)$ are specified as random data points, where $j=1,2,\ldots,k$.
2) Repeat revision: allocates each x_i data from the data set $X=\{x_1, x_2, ..., x_n\}$ to class $C_p(l)$ when:

$$\left[x_i - C_p(l)\right] < [x_i - C_q(l)] \tag{3}$$

Where l for iterations. $p,q = 1,2,\ldots,n$, $p \neq q$, $I = 1,2,\ldots,n$.

3) Update center cluster center: new cluster center on l+1 calculation,

$$C_j(l+1) = \frac{1}{N_j} \sum_{x_i \in C_j(l)} x_i \quad (4)$$

Where N_j is the amount of data in the cluster j.

4) Stop the iteration if $C_i(l+1) = C_i(l)$ or $| C_i(l+1)\ Ci(l) | < \varepsilon$, if not repeat to step 2.

3.2. Naïve Bayes Classifier

In machine learning, Naïve Bayes Classifier is a simple probabilistic family based on the application of the Bayes theorem with the assumption of strong independence (naïve) among these features [12] [13]. Naïve Bayes Classifier is a simple technique for constructing model classifiers that assigns class labels to examples of problems, represented as feature value vectors, where class labels are taken from several limited sets [3].

In general, Naïve Bayes Classifiers perform well when compared to other classifiers because of their simplicity, less computational complexity, small memory requirements, and good predictive accuracy [14] [15]. Calculation of Naïve Bayes Classifier is stated with:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (5)$$

Where, posterior probability, $P(C|X)$, is calculated using Class Prior Probability $P(C)$, Predictor Prior Probability $P(X)$ and Likelihood $P(X|C)$. In Naïve Bayes Classifier, the assumption that the probability of each attribute with respect to a class is independent of all other attribute values is very strong and simplifies the calculation of probability. This assumption is referred to as Conditional Independence[14][16][17].

The proposed method considers input datasets with attribute values as numerical and Gaussian distributions. For the Gaussian distribution the mean (μ) and standard deviation (σ) need to be calculated using the formula:

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} \quad (6)$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n} \quad (7)$$

While the Gaussian distribution function is calculated by the formula:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi^e}}^{\frac{-(X-\mu)^2}{2(\sigma^2)}} \quad (8)$$

In this case it can be explained that:

μ = The calculated results μ each attribute of training data based on the class that has been done before.

σ = The calculated results σ each attribute of training data based on the class that has been done before.

π = 3.14

e = 2.718282

X = The attribute value in the test data used.

3.3. Image Processing

Modern digital technology has made it possible to manipulate multi-dimensional signals with systems ranging from simple digital circuits to sophisticated parallel computers [18]. Texture is an important feature that identifies objects in any image [19]. Colour is also an attribute that plays a role in identifying certain objects, colour processing including extraction of information about the spectral properties of the object's surface and looking for the best similarities from a series of known descriptions [6]. In this study, seven image extraction techniques will be used, as follows [20]:

a. RGB, RGB colour space comes from red, green, and blue colours that are combined and produce a wider colour. Colour parameters are obtained by normalizing each RGB colour component.

$$r = \frac{R}{(R+G+B)} \; g = \frac{G}{(R+G+B)} \; b = \frac{B}{(R+G+B)} \tag{9}$$

b. HSV, HSV colour space is perceptual from hue which represents different colour tones such as yellow, red, green, expressed in degrees 0-360, saturation representing the depth of colour, such as dark red, bright red, and values that represent dark colours.

$$V = \max(r, g, b) \tag{10}$$

$$S = \begin{cases} 0, if\ V = 0 \\ 1 - \dfrac{\min(r,g,b)}{V}, if\ V > 0 \end{cases} \tag{11}$$

$$H = \begin{cases} 0, if\ S = 0 \\ \dfrac{60 * (g-b)}{S * V}, if\ V > r \\ 60 * \left[2 + \dfrac{b-r}{S*V}\right], if\ V = g \\ 60 * \left[4 + \dfrac{r-g}{S*V}\right], if\ V > b \end{cases} \tag{12}$$

$$H = H + 360\ if\ H < 0 \tag{13}$$

c. Mean, reflects the gray average value of an image.

$$\mu = \sum_{i=0}^{L-1} i.H(i) \tag{14}$$

d. Variance, reflects the gray value of the image in a numerical discrete distribution, the variance is the measure of the width of the histogram, the difference between the average and the gray level.

$$\sigma^2 = \sum_{i=0}^{L-1} (i-\mu)^2.H(i) \tag{15}$$

e. Skewness, reflects the distribution of the degree of histogram asymmetry, the greater the bias the more asymmetrical histogram distribution, but the more symmetrical.

$$\mu_s = \frac{1}{\sigma^3} \sum_{i=0}^{L-1} (i-\mu)^3.H(i) \tag{16}$$

f. Kurtosis, reflects the general state of the gray level distribution approach to the mean and is more concentrated, but the more dispersion.

$$\mu_k = \frac{1}{\sigma^4}\sum_{i=0}^{L-1}(i-\mu)^4.H(i) \quad (17)$$

g. Energy, reflects the uniformity of the gray distribution, the gray level distribution is more homogeneous when the energy is greater, on the other hand the smaller.

$$\mu_N = \sum_{i=0}^{L-1} H(i)^2 \quad (18)$$

4. Discussion

In this paper, the image used is the image of a banana. The initial process is that the input image is resized to 100x100 pixels, then the image is extracted based on colour characteristics and statistics into 11 attributes, namely red, green, blue, hue, saturation, value, mean, variance, skewness, kurtosis, and energy. In this study using training images of 60 data and test images of 15 data. Each image is tested by using a combination of K-Means clustering and Naïve Bayes Classifier by using varying values of *k* or centroid, namely 3, 4, 5, 6, 7, 8, and 9. The total test images are 105 data.

The banana image tested will be grouped first using K-Means clustering with varying values of *k* or centroid so that a new group or class is formed. After the test image is entered into a group. In the process of grouping test images with K-Means clustering iterations are needed several times so that the image enters into the right group Then the image is classified by the Naïve Bayes Classifier method to determine the maturity level of the test image into the mature, medium or raw classes. The results of testing the image and the number of iterations can be seen in table 1.

Table 1. Testing Result

#	The number of iterations in the centroid (*k*)														T/F
	3	target	4	target	5	target	6	target	7	target	8	target	9	target	
1	4	T	9	T	5	T	5	T	6	T	8	T	4	T	7/0
2	4	T	9	T	5	T	5	T	6	T	8	T	6	T	7/0
3	4	T	9	T	5	T	5	T	6	T	8	T	4	T	7/0
4	4	T	9	T	5	T	5	T	6	T	8	T	4	T	7/0
5	4	T	9	T	5	T	5	T	6	T	8	T	4	T	7/0
6	4	T	8	T	5	T	5	T	4	T	10	T	4	T	7/0
7	3	T	9	T	5	T	5	T	6	T	8	T	4	T	7/0
8	4	F	6	T	5	F	5	F	6	F	6	F	4	F	1/6
9	4	T	8	T	4	T	5	T	6	T	8	T	4	T	7/0
10	4	T	7	T	6	T	6	T	5	T	8	T	4	T	7/0
11	3	T	3	T	10	T	5	F	3	T	4	F	4	F	4/3
12	4	T	5	T	10	T	5	T	4	T	8	T	8	T	7/0
13	4	F	7	F	5	F	5	F	6	T	8	F	4	F	1/6
14	4	T	8	T	6	T	5	T	6	T	6	F	4	T	6/1
15	4	T	9	T	6	T	5	T	6	T	4	T	4	T	7/0
												Correct Amount:			**89 data**
												Amount of False:			**16 data**
												Total Percentage :			**85%**

T = True F = False

In table 1 you can see the results of testing image data with a combination of K-Means and Naïve Bayes Classifier. Of the 105 image data tested, the number of correct test data is 89 data and the number of incorrect test data is 16 data. The total percentage obtained from the test is 85%. In table 1 it can also be seen that the largest number of iterations is 10 iterations, namely the 6th test data with the value k = 8 and the 11th and 12th test data with the value k = 5. In this case a large number of k does not always affect the number of iterations and target accuracy. In Figure 1 it can be seen a graph of the decrease and increase in the iteration value in testing the data.

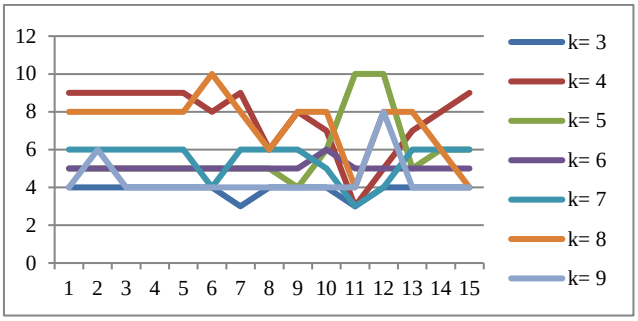

Figure 1. Iteration Result

5. Conclusion

In this paper, the tested image is the image of a banana that has been extracted, the combination of K-Means clustering with Naïve Bayes Classifier can recognize the target image of the test data properly. The results obtained are image classification based on the level of maturity, namely mature, medium, or raw. Of the 105 image data tested, 89 data can be recognized well with an accuracy rate of 85%. In addition to increasing accuracy, the combination of the K-Means method with Naïve Bayes Classifier can alleviate the performance of the Naïve Bayes Classifier because many datasets are first grouped according to the class closest to the test data.

References

[1] Wang X, Jiao Y, and Fei Shumin 2015*Estimation of Clusters Number and Initial Centers of K-means Algorithm using Watershed Method.*IEEE DCABES, pp 505-508.

[2] Qi J, Yu Y, Wang L, and Liu J 2016*K*-Means: An Effective and Efficient K-means Clustering Algorithm.* IEEE BDCloud, pp 242-249.

[3] Granik M and Mesyura V 2017 *Fake News Detection using Naïve Bayes Classifier.*IEEE UKRCON, pp 900-903.

[4] Chandrasekar P and Qian K 2016*The Impact of Data Preprocessing on the Performance of Naïve Bayes Classifier.* IEEE COMPSAC, pp 618-619.

[5] Xiao Q, Niu W, and Zhang H 2015*Prediction Fruit Maturity Stage Dynamically Based of Fuzzy Recognition and Colour Feature.* IEEE ICSESS,pp 944-948.

[6] Mulyani E D S, Susanto, and Poniman J 2017 *Classification of Maturity Level of Fuji Apple Fruit with Fuzzy Logic Method.* IEEE CITSM, pp 1-4.

[7] Mohd R S, Herman S H, and Sharif Z 2014 *Application of K-means Clustering in Hot Spot Detection for Thermal Infrared* Images. IEEE ISCAIE, pp 107-110.

[8] Ayech M W and Ziou D 2015*Ranked K-means Clustering for Terahertz Image Segmentation.*IEEE ICIP, pp 4391-4395.

[9] Adi A O and Celibi E 2014*Classification of 20 News Group with Naïve Bayes*

Classifier. IEEE SIU, pp 2150-2153.
[10] Kapil S, Chawla M, and Ansari M D 2016*On K-means Data Clustering Algorithm with Genetic Algorithm.*IEEE PDGC, pp 202-206.
[11] Bhadana A and Singh M 2017*Fusion of K-means Algorithm with Dunn's Index for Improved Clustering.*IEEE CSITSS, pp 273-277.
[12] Khan M R, Padhi S K, Sahu B N, and Sehera S 2015 *Non Stationary Signal Analysis and Classification using FTT Transform and Naïve Bayes Classifier*. IEEE PCITC, pp 1-6.
[13] Wang K and Shang W 2017 *Outcome Prediction of DOTA2 Based and Naïve Bayes Classifier.* IEEE ICIS, pp 591-593.
[14] Netti K and Radhika Y 2015 *A Novel Method for Minimizing Loss of Accuracy in Naïve Bayes Classifier*. IEEE ICCIC, pp 1-4.
[15] Wang Y, Wang J, Cheng W, Zhao Z, and Cao J 2014 *HPLC Method for the Simultaneous Quantification of the Major Organic Acids in Angeleno Plum Fruit.* IOP Conf. Series: Materials Science and Engineering, pp 1-6.
[16] Srisuan J and Hanskunatai A 2014 *The Ensemble of Naïve Bayes Classifier for Hotel Searching.* IEEE ICSEC, pp 168-173.
[17] Haleem H, Sharma P K, and Beg M M S 2014 *Novel Frequent Sequential Patterns Based Probabilistic Model for Effective Classification of Web Documents.* IEEE ICCCT, pp 361-371.
[18] Young I T, Gerbrands J J, and Vliet L J 2007*Fundamentals of Image Processing.* Delft University of Technology (handbook), pp 1-112.
[19] Kavitha J C and Suruliandi 2016 *Texture and Colour Feature Extraction for Classification of Melanoma using SVM.* IEEE ICCTIDE, pp 1-6.
[20] Yuan W, Hamit M, Kutluk A, Yan C, Li L, Chen J, Hu Y, and Fang F 2013 *Feature Extraction and Analysis on Xinjiang Uygur Medicine Image by using Colour Histogram.* IEEE ICMIPE, pp 259-264.