http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 9, No.1 Desember 2020 ISSN : 2302-9706

EVALUATION OF THE K-NEAREST NEIGHBOR MODEL WITH K-FOLD CROSS VALIDATION ON IMAGE CLASSIFICATION

M. Rhifky Wayahdi [1], Dinur Syahputra [2], Subhan Hafiz Nanda Ginting [3]
Department of Information System, Faculty of Technology, Battuta University
[1]muhammadrhifkywayahdi@gmail.com, [2]dinsyahui12@gmail.com, [3]subhanhafiz16@gmail.com

Abstract

In this paper, the data used is the banana image which is extracted into the dataset into 4 attributes, namely red, green, blue, and the mean for the classification process. Image data is classified using the k-Nearest Neighbor method which will be optimized the model with the k-Fold Cross Validation algorithm. Evaluation of the k-NN model with the k-FCV algorithm can improve accuracy and can build better machine learning models in the image classification process. The default K-NN obtained an accuracy rate of 57%, while the results of the model evaluation with the k-FCV algorithm, on fold 3 obtained an accuracy rate of 68%. The percentage yield with the new model increased by 11% which indicates that the machine learning model that was built was quite optimal.

Keyword: Machine learning, k-NN, k-FCV, Image Processing, Classification.

1. Introduction

Image is defined as a 2-dimensional function with the symbol f (x, y) where x and y are spatial coordinates, and the amplitude f in each coordinate pair (x, y) is called the intensity or gray level of the image. Intensity values x, y, and discrete sums are called digital images. In the field of image processing, digital images are processed using a digital computer for classification or other purposes to obtain new information and knowledge [1].

Many algorithms can be used for the classification process, especially algorithms in the scope of artificial intelligence with a focus on the field of machine learning. Machine learning has played a role in advancing data analysis and artificial intelligence [2]. Machine learning is an algorithm or technique that can be used to design systems or models that can learn without being explicitly programmed [3]. Machine learning continues to evolve in computation, logical algorithmic patterns, and complex data structure designs [4].

One algorithm or method that can be used for classification is k-Nearest Neighbor. K-Nearest Neighbor (k-NN) is a classic classification method [5]. K-NN is the most suitable classification method for simplicity, adaptability, performance [6], easy to apply, and popular [7]. The k-NN method can not only be used for classification, but also for prediction or regression [8][9].

Jaafar *et al.* in his research on the classification of hand-based biometric image databases which are fingerprint and finger vein databases, utilizing k-NN as a classification process and optimizing k-NN to get a better percentage [10]. Whereas Li & Zhang in their research on music personalized recommendation, the k-NN algorithm is used in collaborative filtering and takes advantage of the advantages of k-NN to be modified to make it more effective [11].

By default, the k-NN method does not have a special algorithm in determining training data and testing data as parameters for building machine learning models, so that at the production stage it is often the cause of the model being built not working properly when it finds new data. The solution to this problem is to use a validation set, one of which is k-Fold Cross Validation. K-Fold Cross Validation (k-FCV) is a statistic that can be used to select a model to better predict the predictive model test error [12], as well as estimate generalization performance [13]. Caon *et al.* in his research on experiments on acoustic model explained that k-FCV is the best technique that can be used in every case [14].

Several things above underlie this research. The author will optimize or evaluate the model of the k-Nearest Neighbor method by using the k-Fold Cross Validation algorithm in the image classification process. The image used in this study is the image of the fruit which will be classified according to its maturity level based on color characteristics and statistics. This research was conducted to obtain the level of accuracy and the best machine learning model in the classification process.

2. Research Methodology

In the methodology or stages of this research, several processes will be described and explained. The data used are banana images which are transformed or extracted into a dataset for the classification process. The image is classified using the k-NN method which will be optimized for the model using the k-FCV algorithm. The research flow diagram can be seen in Figure 1.

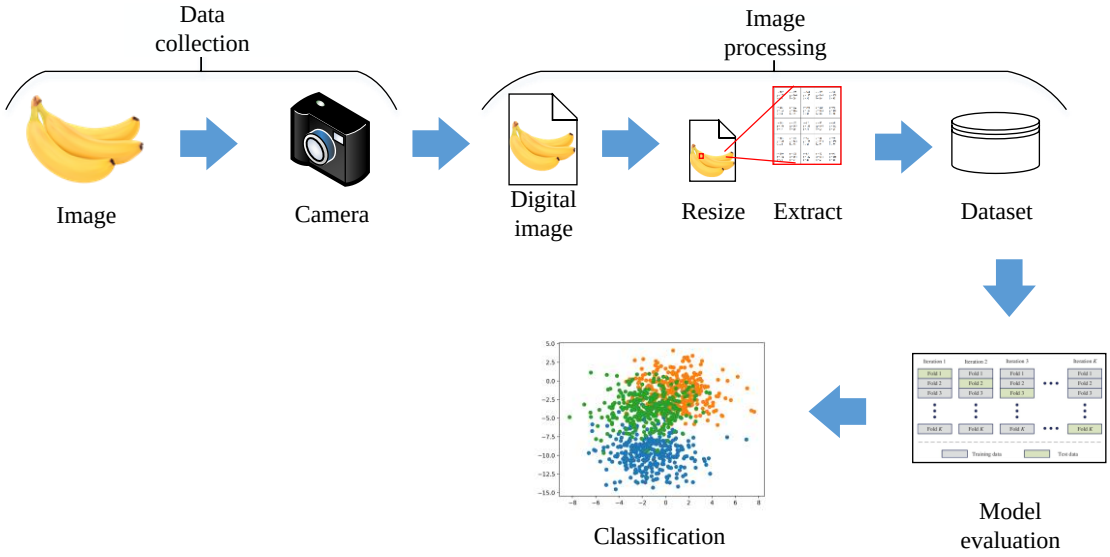

Figure 1. Research Flow Diagrams

In Figure 1, an outline of the flow of the research methodology can be seen. The following is a detailed explanation of the proposed research.

2.1. Data collection

The data in this study are image files with portable network graphics (png) format obtained using a digital camera as many as 75 data (records) with 3 targets (labels), namely mature, medium, and raw.

2.2. Image processing

Image processing is divided into 2 processes, namely resizing and extraction, here are the details:

a) resize the image

The original image will be resized to a size of 100 x 100 pixels to make the classification process easier because of the uniform image size. Following is the process of resizing the image in pseudocode form.

```
read(imgH, imgW);                    { ex. imgH = 354 px, imgW = 502 px }
scaleH=100;                          { height scale 100 px }
scaleW=100;                          { width scale 100 px }
resizeH=(scaleH / imgH) * imgH;      { result (height 100 px) }
resizeW=(scaleW / imgW) * imgW;      { result (width100 px) }
```

b) image extraction

After the image is resized, then the image is extracted into 4 attributes, namely red, green, blue, and the mean. The following is the image extraction process in pseudocode form.

```
read(imgR, imgG, imgB);                 { read r, g, b from image }
r=imgR / (imgR + imgG + imgB);          { red normalization }
g=imgG / (imgR + imgG + imgB);          { green normalization }
b=imgB / (imgR + imgG + imgB);          { blue normalization }
write(r,g,b);                           { normalization results of r, g, and b }

grayscale(image);                       { average (r, g, b / 3) }
n=10000;                                { image size 100x100 px }
for i=0 to 255 do
    read(grayscale[i]);                 { read grayscale }
    num[i]=sum(grayscale[i]);           { total grayscale }
    histogram[i]=num[i] / n;            { histogram }
end for;
for i=0 to 255 do
    mean=sum(grayscale[i]*histogram[i]); { mean }
end for;
write(mean);                            { show the mean value }
```

2.3. Model evaluation

The evaluation of the machine learning model used is the k-FCV algorithm. In cross validation, the dataset is divided by k fold. Where in iteration, each fold is used once as test data and the remaining fold is used as training data, the process is repeated until all data is evaluated. An overview of the k-FCV process as an evaluation model can be seen in Figure 2.

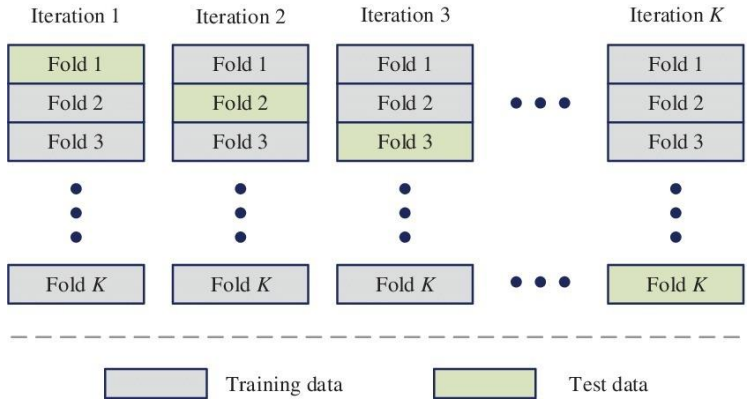

Figure 2. k-Fold Cross Validation Diagrams

2.4. Classification

Image classification using the k-NN method with attribute values obtained from image resizing and extraction. The following is the image classification process using the k-NN method in the form of pseudocode.

```
read(k);                                { k value (ex. k=5 }
read(dataset);                          { read image dataset }
for i=1 to dataset do                   { Euclidean }
    d[i]=sqrt[sum(r-rDataset[i])^2 +(g-gDataset[i])^2+(b-bDataset[i])^2+(mean-meanDataset[i])^2];
    write(d[i]);                        { euclidean results }
end for;
for i=1 to dataset do
    write(asc(d[i]),k);                 { show euclidean as asc number of k }
end for;
result(knn);                            { k-NN result }
```

3. Results and Discussion

In this study, the image used is the image of a banana. The input image is resized to be 100 x 100 pixels, and extracted 4 attributes, namely red, green, blue, and the mean. The amount of data in the image dataset that has been processed is 75 data which is divided into 60 training data and 15 test data randomly. Each data was tested using the default k-NN method with varying k values, namely 3, 5, 7, and 9. Total tests with the default k-NN were 60 trials. The results of the classification using the default k-NN method can be seen in Table 1.

Table 1. Test Results with k-NN Default

Image to-	Classification Results				T/F
	k=3	*k*=5	*k*=7	*k*=9	
1	*True*	*True*	*True*	*True*	4/0
2	*False*	*False*	*False*	*False*	0/4
3	*False*	*False*	*False*	*False*	0/4
4	*True*	*True*	*True*	*True*	4/0
5	*False*	*False*	*False*	*False*	0/4
6	*False*	*False*	*False*	*False*	0/4
7	*False*	*False*	*False*	*False*	0/4
8	*False*	*False*	*True*	*True*	2/2
9	*False*	*False*	*False*	*False*	0/4
10	*True*	*True*	*True*	*True*	4/0
11	*True*	*True*	*True*	*True*	4/0
12	*True*	*True*	*True*	*True*	4/0
13	*True*	*True*	*True*	*True*	4/0
14	*True*	*True*	*True*	*True*	4/0
15	*True*	*True*	*True*	*True*	4/0
			Correct amount:		**34 data**
			Wrong amount:		**26 data**
			Percentages:		**57%**

In Table 1, you can see the results of the image classification test using the default k-NN method. With 60 tests, the correct number in the classification process is 34 data and the remaining 26 data results in a wrong classification. So as to produce a total percentage of 57%. After testing the data with k-NN is complete, the next step is to evaluate the k-NN method model with the k-FCV algorithm. Evaluation of the k-NN model with k-FCV also with varying k values, namely 3, 5, 7, and 9. As well as variations in the k-fold value of 5 fold. So that the total test with k-NN optimization is 300 experiments. The results of the classification using the k-NN optimization method can be seen in Table

Table 2. Test Results with k-NN Optimization

k-Fold	Classification Results				T/F	Percentages
	k=3	*k*=5	*k*=7	*k*=9		
1	4/15	9/15	8/15	8/15	29/60	48%
2	7/15	8/15	7/15	7/15	29/60	48%
3	9/15	10/15	11/15	11/15	41/60	68%
4	7/15	8/15	8/15	10/15	33/60	55%
5	9/15	10/15	10/15	10/15	39/60	65%

In Table 2, you can see the results of the image classification test with k-NN optimization. With the k-fold variation of 1 to 5, it produces a better percentage than the percentage generated with the default k-NN, that is, with the highest percentage in the 3rd fold, which is 68% with the number of classifications that are true as many as 41 data from a total of 60 data test. The visualization of the comparison between the default k-NN method and the k-NN optimization method can be seen in Figure 3.

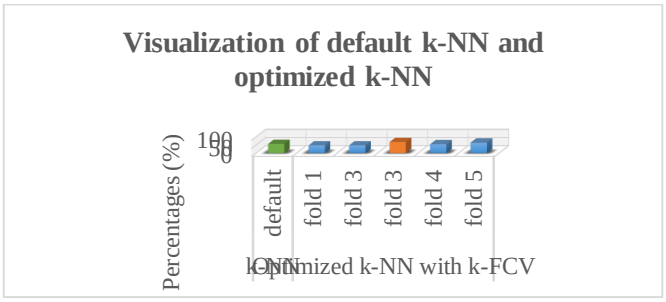

Figure 3. Visualization of default k-NN and optimized k-NN

From Figure 3 it can be seen that the evaluation of the k-NN model with the k-FCV algorithm can improve accuracy and get the best machine learning model in the image classification process. Where the default k-NN with random test data obtained an accuracy rate of 57%. Meanwhile, the results of the model evaluation using the k-FCV algorithm, on fold 3 get an accuracy rate of 68%.

4. Conclusions

Based on the results of the analysis and testing of the k-Nearest Neighbor method and the evaluation of the k-Nearest Neighbor method with the k-Fold Cross Validation algorithm in image classification, the results show that the k-Fold Cross Validation algorithm can improve accuracy and build a better model in the method. k-Nearest Neighbor. The accuracy rate obtained by the default k-Nearest Neighbor method is 57%, while the accuracy rate of the evaluation results of the k-Nearest Neighbor method with k-Fold Cross Validation is 68% on the 3rd fold. The percentage yield with the new model increased by 11%, so that the increasing accuracy of the image classification obtained is directly proportional to the good machine learning model being built.

References

[1] M.R. Wayahdi, Tulus, & M.S. Lydia, "Combination of k-Means with Naïve Bayes Classifier in the Process of Image Classification", *IOP Conference Series: Materials Science and Engineering, 3rd NICTE*, pp. 1-7, 2020.

[2] Z. Chen & B. Liu, "Lifelong Machine Learning (Second Edition)", *Handbook Morgan and Claypool Publisher*, pp. 1-187, 2018.

[3] M. Kang & N.J. Jameson, "Machine Learning: Fundamentals.", *Handbook Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and Internet of Things*, pp. 85-109, 2018.

[4] A. Nayak & K. Dutta, "Impacts of Machine Learning and Artificial Intelligence on Mankind", *International Conference on Intelligent Computing and Control (12C2)*, pp. 1-3, 2017.

[5] D. Pan, Z. Zhao, L. Zhan, & C. Tang, "Recursive Clustering k-Nearest Neighbors Algorithm and the Application in the Classification of Power Quality Disturbance, *IEEE Confrence on Energy Internet and Energy System Integration (EI2)*, pp. 1-5, 2017.

[6] V.L. Boiculese, G. Dimitru, & M. Moscalu, "Improving Recall of k-Nearest Neighbor Algorithm for Classes of Uneven Size", *the 4th IEEE International Conference on E-Healt and Bioengineering-EHB*, pp. 1-4, 2013.

[7] S. Du, & J. Li, "Parallel Processing of Improved k-NN Text Classification Algorithm Based on Hadoop", *IEEE 7th International Conference on Information, Communication, and Networks*, pp. 167-170, 2019.

[8] W. Lei & L. Zhaowei, "Research on the Humanlike Trajectories Control of Robots Based on the k-Nearest Neighbors", *IEEE Chinese Automation Congress (CAC)*, pp. 7746-7751, 2017.

[9] M. Kang, "Machine Learning: Anomaly Detection", *Handbook Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and Internet of Things*, pp. 131-162, 2018.

[10] H. Jaafar, N. Mukahar, & D.A. Ramli, "Methodology of Nearest Neighbor: Design and Comparison of Biometric Image Database", *IEEE Student Conference on Research and Development (SCOReD)*, pp. 1-6, 2016.

[11] G. Li & J. Zhang, "Music Personalized Recommendation System Based on Improved KNN Algorithm", *IEEE 3rd Advanced Information Technology, Electronic, and Automation Control Conference (IAEAC)*, pp. 777-781, 2018.

[12] P. Tamilarasi & U. Riani, "Diagnosis of Crime Rate Against Women using k-Fold Cross Validation through Machine Learning Algorithms", *IEEE 4th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1034-1038, 2020.

[13] B. Mu, T. Chen, & L. Ljung, "Asymptotic Properties of Hyperparameter Estimators by using Cross-Validation for Regularized System Identification", *IEEE Confrence on Decision and Control (CDC)*, pp. 644-649, 2018.

[14] D.R.S. Caon, A. Amehraye, J. Razik, G. Chollet, R.V. Andreao, & C. Mokbel, "Experiments on Acoustic Model Supervised Adaptation and Evaluation by k-Fold Cross Validation Techique", *IEEE International Journal*, pp. 1-4, 2010.