Science and Technology Indonesia
e-ISSN:2580-4391 *p*-ISSN:2580-4405
Vol. 5, No. 3, July 2020

Research Paper

Improved the Cans Waste Classification Rate of Naïve Bayes using Fuzzy Approach

Yulia Resti[1]*, Firmansyah Burlian[2], Irsyadi Yani[2] Des Alwine Zayanti[1] Indah Meiliana Sari[1]

[1]*Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya, Sumatera Selatan, Indonesia*
[2]*Jurusan Teknik Mesin, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya, Sumatera Selatan, Indonesia*
***Corresponding author**: yulia_resti@mipa.unsri.ac.id

Abstract
Cans is one type of inorganic waste that can take up to hundreds of years to be decomposed on the ground so that recycling is the right solution for managing cans waste. In the recycling industry, can classification systems are needed for the sorting system automation. This paper discusses the cans classification system based on the digital images using the Naive Bayes method, where the input variables are the pixel values of red, green, and blue (RGB) color, and the image of the can is captured by placing it on a conveyor belt which runs at a certain speed. The average accuracy rate of the k-fold cross-validation which is less satisfactory from the classification system obtained using the original Naive Bayes model is corrected using the fuzzy approach. This approach succeeded in improving the average accuracy of the can classification system which was originally from 50.26 % to 85.19 % or an increase of 34.93 %, where the standard deviation decreased from 14.01 % to only 6.29 %. A decrease in the standard deviation of 7.72 % also indicates that this model is better than the ONB model.

Keywords
classification system, fuzzy, improved Naïve Bayes

Received: 2 June 2020, Accepted: 19 July 2020
https://doi.org/10.26554/sti.2020.5.3.75-78

1. INTRODUCTION

Recently, the Naïve Bayes method is widely used to classify objects based on the digital images (Hsu et al. (2017); Mansour (2018); Park (2016)). The advantage of the Naïve Bayes classification method is that it has a simple algorithm and works well when the data has a higher dimensional space (Sequera et al. (2017); Kavila et al. (2016)). However, when the input variable is a continuous variable, the Gaussian distribution assumption used sometimes does not provide a satisfactory accuracy rate.

The can classification system can also be built based on digital images, where the input variables are the pixel values of the color of the can digital image. Resti et al. (2017);Resti et al. (2019a); Resti et al. (2019b) used the red, green, and blue (RGB) color models, while Resti et al. (2020) used the cyan, magenta, yellow, and black (CMYK) color models to represent the pixel values of the color of the can digital image, however the accuracy rate obtained is not satisfactory (Resti et al. (2017); Resti et al. (2019a) obtains an accuracy rate of less than 80%, whereas Resti et al. (2019b); Resti et al. (2020) obtains an accuracy rate of less than 50%). There is no specific definition of a minimum accuracy rate of a classification system, but obtaining a better accuracy rate makes the system built more accurate, efficient and useful. Generally the researches claim the minimum accuracy rate of a classification system for a satisfactory level at 85 % (Aronoff (1985); Foody (2008); Liu and An (2020)). Based on this fact, it is very important to develop and modify the classification system of cans waste from previous studies so that the classification system has a higher accuracy rate and at least achieves more than 85% accuracy.

One way to overcome this problem is to use a fuzzy approach (Rastogi et al. (2019); Soares and Moraes (2018); Aziz et al. (2016); Ferreira et al. (2015); Moraes (2015)). This paper discusses the classification of cans using a fuzzy approach to the Naïve Bayes model to obtain better classification results than the original Naïve Bayes. The validation technique used is k-fold cross validation, while the process of splitting up data and classification is done with the help of R software 4.01.

2. EXPERIMENTAL SECTION

2.1. Data

The data used in this study are data of pixel values of red (X_1), green (X_2), and blue (X_3) of 250 cans which are distributed into three types of cans; 29.6 % cans of type 1, 33.2 % cans of type 2, and 37.2 % cans of type 3. The pixels are obtained by capturing

cans placed on conveyor belts at a speed of 0.181 m/sec where the webcam was set at an angle of 900. The statistics summary of the pixel values of each colour are presented in Table 1.

Table 1. Statistics Summary of Input Variables

Statistics	Input variable		
	X_1	X_2	X_3
Minimum	129.2	134.7	123.8
1^{st} Quartile	139.1	140.7	139.1
Median	141.7	143.2	142.1
Mean	142.9	143.4	142.4
3^{rd} Quartile	145.8	146.2	145.1
Maximum	179.9	161.4	181.2

2.2. Methods

This study uses k-folds cross validation technique by splitting data into k=10 fold to obtain the best classification model (Sharma et al., 2017). This data splitting process is carried out with the help of software R 4.0.1. The initial step after having 10 fold data is to randomly select one fold data as test data from the 10 folds data where 9 folds data that are not selected as test data are used as training data. Second, determine the prior probabilities, likelihood function parameters of each input variable, and the posterior probabilities of each type of can in each training data for obtained a classification model using original naïve Bayes (ONB) as in equation (1). Let X_1, X_2, X_3 be the input variables of pixel values of red, green, and blue successively, K_j be the j-th cans type, $j = 1, 2, 3$. Probability ofK_j, given X_1, X_2, X_3 according to the Bayes theorem is expressed as (Ferreira et al., 2015).

$$P(K_j|X_1, X_2, X_3) = \frac{P(X_1, X_2, X_3|K_j P(K_j)}{P(X_1, X_2, X_3)} = \frac{P(K_j)}{P(X_1, X_2, X_3} \prod_{d=1}^{3} P(X_d|K_j) \quad (1)$$

Third, classify each can of observations in the test data using the ONB model obtained in step 2 and create a confusion matrix as Table 2 to obtain an accurate level of classification as in equation (2), where T_jj be the percentage of cans coming from the j-th cans type predicted exactly to the j-th cans type, whereas F_jl is the percentage of the number of cans coming from the j-th cans type predicted to the l-th cans type.

Table 2. Confusion matrix

Cans Type Actual (j-th)	Cans Type Predicted (l-th) 1st	2nd	3rd
1^{st}	T_{11}	F_{12}	F_{13}
2^{nd}	F_{21}	T_{22}	F_{23}
3^{rd}	F_{31}	F_{32}	T_{33}

$$Accuracyrate = T_{11} + T_{22} + T_{33} \quad (2)$$

The next step is to use a fuzzy approach such as on each input variable of each type of can in each training data to determine the fuzzy probability as in equations (3) to obtain a classification model using this approach to the naïve Bayes (ONB) model.

$$P(X) = \int \varphi_T(x) f(x) dx \quad (3)$$

Where for each $d = 1, 2, 3, f(x)4$ be the probability density function of Gaussian distribution, $\varphi_T\ (x)$ be the membership functions of fuzzy set in this research are denoted in eq. (4) – (6). Let a be the element of the domain that has the greatest membership value and b be the element of the domain that has the smallest membership value, the membership functions for dark color is

$$\varphi_T(x) = \begin{cases} 1 & ; \quad x \le a \\ \frac{b-x}{b-a} & ; \quad a \le |x| \le b \\ 0 & ; \quad x \le b \end{cases} \quad (4)$$

Let α be the element of the domain which is the smallest value and also has the smallest membership value, b be the element of domain which is the median of data and has the greatest membership value, and c be the element of domain which is the greatest value but has the smallest membership value, the membership functions for moderate color is

$$\varphi_T(x) = \begin{cases} 0 & ; \quad x \le a, ataux \ge c \\ \frac{x-a}{b-a} & ; \quad a \le x \le b \\ \frac{c-x}{c-b} & ; \quad b \le x \le c \end{cases} \quad (5)$$

Let α be the element of domain that has the smallest membership value and b be the element of domain that has the greatest membership value, the membership functions for light color is

$$\varphi_T(x) = \begin{cases} 0 & ; \quad x \le a \\ \frac{x-a}{b-a} & ; \quad a \le x \le b \\ 1 & ; \quad x \ge b \end{cases} \quad (6)$$

Next is to classify each can of observations on the test data using the model obtained in previous step and make a confusion matrix to obtain the classification accuracy rate, and finally analyze the classification results.

3. RESULTS AND DISCUSSION

3.1. The Original Model of Naïve Bayes

In the ONB model, the parameters of the input variables that are assumed to be Gaussian distributions are estimated with maximum likelihood estimation. The estimated parameter results for the 10^{th} fold are presented in Table 3. The parameters for other folds are obtained in the same way.

Table 3. Parameter Gaussian Distributions for the 10^{th} Fold

Can type	Input variable					
	μ	σ	μ	σ	μ	σ
1^{st}	153.84	8.9	151.28	6.11	142.4	9.34
2^{nd}	145.93	4.81	148.81	4.12	147.76	2.39
3^{rd}	146.69	5.52	147.76	4.76	145.18	4.38

In this study, the 6^{th} fold data was chosen as the test data so that the other 9 fold data as training data. Implementation of the 10^{th} fold data as training data and the 6^{th} fold data as test data to equation (1) gives the classification results as presented in Table 4 with a classification accuracy rate of 75.00%. Only cans from the 2^{nd} type have all been classified correctly. The highest percentage of misclassification occurs in cans from the 3^{rd} type are classified as cans of the 2^{nd} type, which is 16.67%.

Table 4. Classification Result of ONB Model for the 10^{th} Fold

Original Naïve Bayes Model		% number of cans from type		
		1^{st}	2^{nd}	3^{rd}
% number of	1^{st}	20.83	0	0
cans classified	2^{nd}	0	33.33	16.67
into can type	3^{rd}	8.33	0	20.83
Accuracy rate		75.00 %		

The classification results of the 6^{th} fold as test data using 8 other data folds as training data for the ONB model are obtained in the same way. The classification accuracy rate of the ONB model using the k-fold cross validation technique given in Table 5 shows that the ONB model has an average accuracy rate of 50.26 % with a standard deviation of 14.01%. The accuracy of this model ONB can be improved significantly using the fuzzy approach as presented in section 3.2.

Table 5. Accuracy Rate of ONB Model

Training Fold	Testing Fold	Accuracy Rate of ONB
1		51.85%
2		55.56%
3		29.63%
4		32.00%
5	6	50.00%
7		62.50%
8		50.00%
9		45.83%
10		75.00%
average		50.26%
Standard deviation		14.01%

3.2. Improved Model of Naive Bayes using Fuzzy Approach

The fuzzy membership function of each input variable is obtained using equation (4) - (6) where the parameters are points with the same distance in the interval [min, max] of pixel values. In the 10^{th} fold data, the fuzzy membership function of variable X_1 which has a pixel value in the interval [137.59, 167.49] is expressed by,

$$\varphi_D(x_1) = \begin{cases} 1 & ; \quad x_1 \le 137.59 \\ \frac{147.56 - x_1}{9.97} & ; \quad 137.59 \le x_1 \le 147.56 \\ 0 & ; \quad x_1 \le 162.50 \end{cases} \tag{7}$$

$$\varphi_M(x_1) = \begin{cases} 0 & ; \quad x_1 \le 142.57, ataux_1 \ge 162.50 \\ \frac{x - 142.57}{9.97} & ; \quad 142.57 \le x_1 \le 152.54 \\ \frac{162.50 - x_1}{9.97} & ; \quad 152.54 \le x_1 \le 162.50 \end{cases} \tag{8}$$

$$\varphi_L(x_1) = \begin{cases} 0 & ; \quad x_1 \le 157.52 \\ \frac{x_1 - 157.52}{9.97} & ; \quad 157.52 \le x_1 \le 167.49 \\ 1 & ; \quad x_1 \ge 167.49 \end{cases} \tag{9}$$

The fuzzy membership function and the parameters for other variables and fold data are obtained in the same way.

The classification results of the 6^{th} fold as test data and the 10^{th}fold data as training data using Fuzzy approach (IONBF) are presented in Table 6 with an accuracy rate of 83.33 %. The percentage of misclassification in each type at 4.17 % where cans from the 1st and 2nd types are classified as cans of the 3rd type and cans from the 2nd as type. There is not one type of can which all of its members are classified correctly as a whole.

Table 6. Classification Result of IONBF Model for the 10th Fold

IONBF Model		% number of cans from type		
		1^{st}	2^{nd}	3^{rd}
% number of	1^{st}	25	0	0
cans classified	2^{nd}	0	29.17	4.17
into can type	3^{rd}	4.17	4.17	29.17
Accuracy rate		83.33%		

The classification results of the 6^{th} fold as test data using IONBF model where 8 other fold data as training are obtained in the same way. Accuracy Rate of IONBF model using k-fold cross validation technique given in Table 7 noted that the IONBF model has an average accuracy rate of 85.19 % with a standard deviation of 6.29 %. All accuracy rate of testing data for all training data in the IONBF model is higher than the ONB model, as well as the average accuracy overall. The improvement of the average accuracy rate from ONB model to IONBF of this model is 34.93%. This fact shows that the fuzzy approach on the ONB model can improve the classification accuracy rate.

Table 7. Accuracy Rate of IONBF Model

Training Fold	Testing Fold	Accuracy Rate of ONB
1		83.33%
2		79.17%
3		79.17%
4		79.17%
5	6	95.83%
7		83.33%
8		91.67%
9		91.67%
10		83.33%
average		85.19%
Standard deviation		6.29%

4. CONCLUSIONS

In this study, the accuracy of the model was obtained as the average level of accuracy of one test data that was randomly selected using a model of 9 data fold as training data. The average accuracy of the IONBF model using cross validation technique is 85.19% with a standard deviation of 6.29 %. This accuracy level is higher than the average accuracy of the ONB model which is only 50.26 % with a standard deviation of 14.01 %. The accuracy of the ONB model can be improved by the fuzzy approach. A decrease in the standard deviation of 7.72 % also indicates that this model is better than the ONB model.

5. ACKNOWLEDGEMENT

This research was supported by DIPA, University of Sriwijaya, No. SP DIPA-042.01.2.400953/2019, for the Competitive Research, No. 0015 /UN9/SK.LP2M.PT/2019.

REFERENCES

Aronoff, S. (1985). The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering and Remote Sensing*, **51**(1); 99–111

Aziz, R., C. Verma, and N. Srivastava (2016). A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data. *Genomics data*, **8**; 4–15

Ferreira, J. A., E. A. Soares, L. S. Machado, and R. M. Moraes (2015). Assessment of Fuzzy Gaussian Naive Bayes for Classification Tasks. *PATTERNS 2015*; 73

Foody, G. M. (2008). Harshness in image classification accuracy assessment. *International Journal of Remote Sensing*, **29**(11); 3137–3158

Hsu, S.-C., I.-c. Chen, and C.-L. Huang (2017). Image Classification Using Naive Bayes Classifier With Pairwise Local Observations. *Journal of Information Science & Engineering*, **33**(5)

Kavila, S. D., , and R. Y (2016). Research Domain Selection using Naive Bayes Classification. *International Journal of Mathematical Sciences and Computing*, **2**(2); 14–23

Liu, J.-e. and F.-P. An (2020). Image Classification Algorithm Based on Deep Learning-Kernel Function. *Scientific Programming*, **2020**

Mansour, A. M. (2018). Texture classification using Naïve Bayes classifier. *IJCSNS Int. J. Comput. Sci. Netw. Secur*, **18**(1); 112–121

Moraes, R. M. (2015). A new generalization for naive bayes style fuzzy probabilistic classifier

Park, D.-C. (2016). Image classification using Naïve Bayes classifier. *International Journal of Computer Science and Electronics Engineering (IJCSEE)*, **4**(3); 135–139

Rastogi, N., S. Rastogi, and M. Darbari (2019). A Novel Software Reliability Prediction Algorithm Using Fuzzy Attribute Clustering and Nave Bayesian Classification. *International Journal of Computer Sciences and Engineering*, **7**(2); 73–82

Resti, Y., F. Burlian, I. Yani, and D. Rosiliani (2020). Analysis of a cans waste classification system based on the CMYK color model using different metric distances on the k-means method. *Journal of Physics: Conference Series*, **1500**; 012010

Resti, Y., A. S. Mohruni, F. Burlian, I. Yani, and A. Amran (2017). A probability approach in cans identification. *MATEC Web of Conferences*, **101**; 03012

Resti, Y., A. S. T. Mohruni, T. Rodiana, and D. Zayanti (2019a). Study in Development of Cans Waste Classification System Based on Statistical Approaches. *Journal of Physics: Conference Series*, **1198**(9); 092004

Resti, Y., F. Nasution, A. S. Mahruni, F. A. Almahdinil, and D. A. Zayanti (2019b). A cans waste classification system based on RGB images using different distances of k-means clustering. *The 5th International Conference on Science, Technology and Interdisciplinary Research, September 23-25 September.*

Sequera, M. S., S. A. Guirnaldo, and P. J. R (2017). Naïve Bayes Classifie and Fuzzy Logic System for Computer-Aided Detection and Classification of Mammamographic Abnormalities. *Journal of Theoretical & Applied Information Technology*, **95**(2)

Sharma, R. C., K. Hara, and H. Hirayama (2017). A machine learning and cross-validation approach for the discrimination of vegetation physiognomic types using satellite based multispectral and multitemporal data. *Scientifica*, **2017**

Soares, E. A. d. M. G. and R. M. Moraes (2018). Fusion of Online Assessment Methods for Gynecological Examination Training: a Feasibility Study. *TEMA (São Carlos)*, **19**(3); 423–436

