Classification based on K-Nearest Neighbor and Logistic Regression method of coffee using Electronic Nose

by Dedy Prehanto

Submission date: 05-Oct-2021 08:15PM (UTC+0700)
Submission ID: 1665899167
File name: Prehanto_2021_IOP_Conf._Ser.__Mater._Sci._Eng._1098_032080.pdf (757.92K)
Word count: 3845
Character count: 19311

IOP Conference Series: Materials Science and Engineering

Classification based on K-Nearest Neighbor and Logistic Regression method of coffee using Electronic Nose

Classification based on K-Nearest Neighbor and Logistic Regression method of coffee using Electronic Nose

D R Prehanto[1,*], A D Indriyanti[1], I K D Nuryana[1] and G S Permadi[2]

[1] Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia
[2] Faculty of Information Technology, Universitas Hasyim Asy'ari, Indonesia

*dedyrahman@unesa.ac.id

Abstract. Coffee has its own scent of identity which can be felt directly with the ability of the human sense of smell. With a specific coffee aroma that can be used to identify the type of coffee. In this study we propose that E-Nose (Electronic Nose) can be used to identify coffee based on the aroma of coffee converted into value data used for the classification process. The initial step is the data validation process using the calculation of the average value, standard deviation, Minmax. After conducting the dataset validation process, the next step is to implement the Logistic Regression (LR) and K-Nearest Neighbor (KNN) classification methods. The accuracy value is derived from the Confusion Matrix evaluation method, TP, TN, FP and FN values. This study focuses on finding the best classification accuracy value with the criteria having the highest accuracy value. This system can be used to classify types of coffee with a mixture of coffee and milk. This study will compare the results of classification using the two classification methods. Based on the results of the accuracy of the two methods presented the best results using the KNN method with a statistical calculation is 97.7%.

1. Introduction

Coffee is the popular consumed beverage product, this is because coffee is an important commodity of international trade based on trade volume [1]. With so many types of coffee that exist in the world, there are several ways to determine the type of coffee that aims to distinguish the authenticity of the coffee. The aroma of coffee is the most commonly used method to determine the type of coffee, because with a high roasting process the coffee produces a stronger aroma [2].

E-Nose is a tool created to recognize odors by capturing gas with sensor aids, by using E-Nose which is applied in testing the aroma of coffee, it can produce a pattern that can be classified in identifying types of coffee [3]. By using K-Nears, the results of extraction of test data from E-Nose can be processed to produce a grouping based on the classification score achieved. This has been tested where K-Nears can classify types of coffee beans based on Image Processing with a classification score reaching 96.66% [3].

Another method for classification uses radial basis functions. This research succeeded in classifying coffee with an accuracy rate of 90.8% for integral feature extraction, 90% for maximum feature extraction and 94.1% for different feature extraction. In 2016 there was a study classifying coffee using backpropagation neural networks. The results showed that backpropagation neural networks were able to determine the difference between arabica and robusta with a 40% success rate. Other methods for

classifying using E-Nose are Support Vector Machine (SVM) and Perceptron method. SVM accuracy value is 71% and Perceptron 57% [4].

From the description of this research, it is expected to develop a classification of coffee using aroma, where it is more often used among traditional people. This study has three statistical calculations, among others, with the average value and standard deviation, the minimum and maximum values and compare between the standard deviation values and the maximum minimum value [5].

2. Methods

2.1. E-Nose processing

Fig. 1 is the first stage of E-Nose for processing aroma of the coffee. The aroma of coffee can be detected by MQ2, MQ3, MQ4, MQ7 and MQ135. The output of those sensors that in the form of an electrical signal is connected to Arduino as a microcontroller which functions as a sensor reader to make it easier.

Figure 1. Process of gas detection.

There are sensing element, sensor base and sensor cap that build the sensor which is shows in Fig. 3. The detector of elements divided into two common part, they are sensing material and heater that function is to heat up sensing element (e.g. 400°C). Depend on the target of gas itself, the detector will reprocess different gases such as Alcohol, NH4, CO2.

2.2. Logistic regression

Logistic Regression aims to test whether the probability of the occurrence of the dependent variable can be predicted with the independent variable. To assess the accuracy of the use of the logistic regression model, namely to find out whether the logistic regression model is in accordance with the data obtained, the model fit test is performed [6].

Regression analysis is one analysis that aims to determine the effect of a variable on other variables. The simplest regression model is a simple linear regression model with the form of equation (1):

$$Y = \beta 0 + \beta 1 X + \varepsilon \quad (1)$$

where :
Y = dependent variable (predicted value)
X = free variable
$\beta 0$ = constant
$\beta 1$ = regression coefficient (increase or decrease value)
ε = random error.

2.3. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a advanced development innovation from the Nearest Neighbor classification technique. Based on the idea each new instance can be classified by the most votes from neighbor k, where k is a positive integer, and usually by a small number [7]. The K-NN classification algorithm predicts the category of the test sample according to the training sample k which is the closest neighbor to the test sample, and puts it in the category that has the largest probability category [8]. KNN is one of the types of insance-based learning, or lazy learning where this function is only approached locally and all calculations are postponed until classification [9].

The K-NN classification method has several stages, as follows [10]:

- The value of k is the determination of a new query that belongs to a class based on the number of closest neighbors.
- Calculates the distance of the query point from the training point to determine the value of the nearest neighbor.
- See the smallest value from the distance of each training point to the query point.
- Take the smallest k value and then look at the class of the new request.

Euclidean distance is used to calculate distance near or far points with neighbors represented as follows [10]:

$$J(a,b) = \sqrt{\sum_{k=1}^{kn} (a_k - b_k)^2} \quad (2)$$

J (a, b) can be interpreted as the distance between point a whose class of points is known and b new points. The distance between the new point and the training point is calculated and taken to the nearest k point. New points are predicted to enter the class with the most classifications of these points [10].

Discrete multiclass classification must assign each observation to one of the standard classes C1, ..., Ck. Can do probabilistic classification where each observation distributes probabilities that describe the probabilities of one of the classes produced. Before a new classification can be used, its accuracy must be evaluated experimentally, by comparing implied classifications (measured by new devices or methods) against real classifications or gold standard classifications (measured by benchmark methods) with a sizeable test data set representative of future data that is large enough expected [11].

Table. 1 presents the basic form of sion matrix configuration for multi-class classification tasks. In the confusion matrix, Nij represents the number of samples actually belonging to the Ai class but is classified as classAj [12]

Table 1. Confusion matrix.

		Predicted				
		A_1	⋮	A_j	⋮	A_n
Actual	C_1	N_{11}		N_{1j}		N_{1n}
	⋮			⋮		
	C_k	N_{i1}	⋮	N_{ij}	⋮	N_{in}
	⋮			⋮		
	C_n	N_{n1}		N_{nj}		N_{nn}

Confusion matrix is a method for compute accuracy of data mining. The formula below is to do calculations with 4 outputs, namely: recall, precision, accuracy and error rate [13]. Recall is the ratio of positive cases that are correctly identified [14]. The result of Precision for the ratio of cases is true positive. The comparison of accurately in identified cases and a total of cases is the meaning of Accuracy. Error Rate is a case that is identified incorrectly with a number of cases [15].

2.4. Data collection

Collection of data used in the form of coffee flavor aromas. Comparison of various coffee scents and coffee with milk mixture as the object of research. This study uses 9 types of coffee and coffee and milk

mixture as a class, which is the data in the study. Comparison of coffee with milk done 50 times for nine classes is presented in Table 2.

Table 2. Proportion coffee with milk.

Proportion Coffee with milk	Classes	Proportion Coffee with milk	Classes
Coffee 0% with Milk 100%	L0-NL100	Coffee 75% with Milk 25%	L75-NL25
Coffee 10% with Milk 90%	L10-NL90	Coffee 80% with Milk 20%	L80-NL20
Coffee 20% with Milk 80%	L20-NL80	Coffee 90% with Milk 10%	L90-NL10
Coffee 25% with Milk 75%	L25-NL75	Coffee 100% with Milk 0%	L100-NL0
Coffee 50% with Milk 50%	L50-NL50		

Data collection was carried out for 15 minutes at room temperature for each experiment. During these 15 minutes, 300 data were obtained. Each class has conducted 50 tests. The coffee in this experiment used processed ground coffee with an ideal grinder level, from coarse to medium size, with the weight of coffee for each data collection of 15 grams. The sensor will detect based on the aroma of coffee which produces digital value.

2.5. Processing data

The next step is to process the data detected by E-Nose on the coffee it detects. Using machine learning to estimate the accuracy of the best model in the invisible data by evaluating the actual data that is not visible. So, the accuracy will estimated using statistical methods that purpose for validating data [14]. The first process carried out in this study is to calculate the Avg and SD value. Then the second is to calculate the Minmax value. The last is to calculate using the Avg, SD and Minmax statistical methods.

The first calculation starts by collecting data on Avg values and SD values of nine classes of the mixture Coffee with Milk into tables and stored that contains the recapitulation results of the nine classes. The amount of data collected from the recapitulation is 450 data with details of 50 data from each class that show in Table 3.

Table 3. Avg dataset and SD of each class.

The test	Avg MQ2	Avg MQ3	Avg MQ4	Avg MQ7	Avg MQ135	Class
1-50	31.85	55.99	49.50	98.25	11.76	L0-NL100
51-100	54.50	76.08	104.48	188.26	9.77	L10-NL90
101-150	54.52	94.50	97.67	193.61	10.51	L20-NL80
151-200	59.31	82.87	121.51	210.53	10.96	L25-NL75
201-250	50.78	88.93	96.86	215.30	11.47	L50-NL50
251-300	62.15	90.84	118.19	201.01	10.36	L75-NL25
301-350	60.04	79.63	110.75	182.16	9.38	L80-NL20
351-400	87.92	116.46	93.76	189.45	9.90	L90-NL10
400-450	48.80	106.35	87.60	175.05	14.61	L100-NL0

The second data processing is by calculating the Minmax values of nine classes of the mixture Coffee with Milk [2]. The data is the recapitulation of 450 data with 50 data details from each class.

Calculations carried out at the data processing stage will be classified using the Logistic Regression (LR), and K-Nearest Neighbor (KNN) methods. The aim is to estimate the accuracy of the statistical calculation when processing data [17].

At the classification stage, the dataset that has been stored will be divided into two types of data, including training data and testing data. The data stored is data nine classes of the mixture Coffee with

Milk. The data, amounting to 50 data from each class, is divided into 80% for training data and 20% for testing data from the total input data [18].

3. Results and discussion

This research was conducted with the aim of utilizing E-Nose in carrying out the detection of the aroma of coffee and continued with the calculation of statistics included in the initial process. The end result is expected that E-Nose can classify the mixture Coffee with Milk by applying a machine learning algorithm. The Classes divided by nine classification based on the value of percentages of coffee mixture that show in Table 4.

Table 4. Classification of the mixture coffee with milk.

No.	Coffee	Milk	Class
1.	0%	100%	L0-NL100
2.	10%	90%	L10-NL90
3.	20%	80%	L20-NL80
4.	25%	75%	L25-NL75
5.	50%	50%	L50-NL50
6.	75%	25%	L75-NL25
7.	80%	20%	L80-NL20
8.	90%	10%	L90-NL10
9.	100%	0%	L100-NL0

After getting the accuracy from average and standard deviation, the next step is trying to find the other accuracy value from another statistical calculation [19]. We will see the result from the calculation of the min value and the max value. From here we get the highest accuracy value, it came from KNN method with 95.27%.

From those data we will get the value of accuracy. We use the formulation in python. First, we will try to get the accuracy from average calculation and standard deviation. We can see the result, the accuracy from LR method of average and standard deviation statistical calculation is 91.38%, then for KNN methods which result is 96.11%.

The results of the classification accuracy using confusion matrix of the mixture coffee with Milk obtained an accuracy of 97.77% from the KNN algorithm and Avg-SD statistical calculation. The coffee mixture data of the L0-NL100 prediction class is classified into 10 data in the target class. The L10NL90 prediction class classifies 9 data in its target class. The L20NL80 prediction class classified 8 data in its target class. The L25NL75 prediction class classified 15 data in its target class. The L50NL50 prediction class classified 12 data in its target class. The L75NL25 prediction class classified 5 data in its target class. The L80NL20 prediction class classified 14 data in its target class. The L90NL100 prediction class classified 7 data in its target class. The L100NL0 prediction class classified 8 data in its target class.

The result of accuracy using confusion matrix that L100NL0 class classification there are 2 data included in the L90NL10 class target, meaning that the detection data between the aroma of coffee from the 2 types of mixture there are similarities in some data when detecting the aroma of coffee. The similarity in the aroma detection data will bring the predicted data to classes that are not in harmony with the target class. So, the results of the classification of the mixture Coffee with Milk obtained an accuracy of 97.77% from the KNN algorithm and Minmax statistical calculation.

Similarly, with Table 5, the results of the classification accuracy using confusion matrix in Table 5 is coffee mixture between Coffee with Milk from the KNN algorithm and statistical calculation of Avg-SD-Minmax value, obtained an accuracy percentage of 97.77%.

Table 5. Confusion matrix of KNN for Avg-SD-Minmax.

	TARGET									
P R		L0-NL 100	L10-NL 90	L20-NL 80	L25-NL 75	L50-NL 50	L75-NL 25	L80-NL 20	L90-NL 10	L100-NL0
E	L0-NL100	**10**	0	0	0	0	0	0	0	0
D	L10-NL90	0	**9**	0	0	0	0	0	0	0
I	L20-NL80	0	0	**8**	0	0	0	0	0	0
C	L25-NL75	0	0	0	**15**	0	0	0	0	0
T	L50-NL50	0	0	0	0	**12**	0	0	0	0
I	L75-NL25	0	0	0	0	0	**5**	0	0	0
O	L80-NL20	0	0	0	0	0	0	**14**	0	0
N	L90-NL10	0	0	0	0	0	0	0	**7**	0
	L100-NL0	0	0	0	0	0	0	0	2	**8**

If seen from the average of all accuracy produced, the classification of Coffee with Coffee can be done using data from the aroma detection results conducted by E-Nose. In doing the classification of the data that is used greatly affects the value of accuracy produced, the more attributes used in the classification process, it will show a higher accuracy value.

Table 6. Algorithm performance of confusion matrix.

	Precision	Recall	F1-Score
Mix L0-NL100	1.00	1.00	1.00
Mix L10-NL90	1.00	1.00	1.00
Mix L20-NL80	1.00	1.00	1.00
Mix L25-NL75	1.00	1.00	1.00
Mix L50-NL50	1.00	1.00	1.00
Mix L75-NL25	1.00	1.00	1.00
Mix L80-NL20	0.78	1.00	0.88
Mix L90-NL10	1.00	1.00	1.00
Mix L100-NL0	1.00	1.00	1.00
Accuracy		**0.98**	

Table 6 shows that the accuracy results are very good if the dataset classified has the amount of False Negative data and the value of False Positive produces very close or symmetric values.

4. Conclusion

This research utilizes the Electronic Nose (E-Nose) system can detect the aroma of coffee mixture between Coffee with Milk. The aroma detection results are shown by the sensor signals which are displayed in digital data. The value that come from each sensor are calculated for validating using average (Avg), standard deviation (SD), minimum and maximum value (Minmax). After that, the next step is to classify the result from validation dataset value for classification using Logistic Regression (LR) and K-Nearest Neighbor (KNN) algorithm. The result from classification will be evaluating using confusion matrix. Accuracy is the main purpose in this research using confusion matrix that generate the best method using KNN classification method. So, the highest classification in this research is KNN method which value is 97.77%. From confusion matrix the result also shown that the best method of validating data is use all of statistical calculation.

References

[1] Kurniawan F, Budiastra I W and Widyotomo S 2019 Classification of Arabica Java Coffee Beans Based on Their Origin using NIR Spectroscopy *IOP Conference Series: Earth and Environmental Science* **309**(1) 012006

[2] Arboleda E R 2018 Discrimination of civet coffee using near infrared spectroscopy and artificial

neural network *International Journal of Advanced Computer Research* **8**(39) 324-334
[3] Arboleda E R, Fajardo A C and Medina R P 2018 Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors *2018 IEEE International Conference on Innovative Research and Development (ICIRD)* pp 1-5
[4] Rahmandani M, Nugroho H A and Setiawan N A 2018 Cardiac Sound Classification Using Mel-Frequency Cepstral Coefficients (MFCC) and Artificial Neural Network (ANN) *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)* pp 22-26
[5] Badriyah T, Tahrir M and Syarif I 2018 Predicting The Risk of Preeclampsia with History of Hypertension Using Logistic Regression and Naive Bayes *2018 International Conference on Applied Science and Technology (iCAST)* pp 399-403
[6] Ohyver M, Moniaga J V, Yunidwi K R and Setiawan M I 2017 Logistic Regression and Growth Charts to Determine Children Nutritional and Stunting Status: A Review *Procedia computer science* **116** 232-241
[7] Khamis H S, Cheruiyot K W and Kimani S 2014 Application of k-nearest neighbour classification in medical data mining *International Journal of Information and Communication Technology Research* **4**(4)
[8] Suguna N and Thanushkodi K 2010 An improved k-nearest neighbor classification using genetic algorithm *International Journal of Computer Science Issues* **7**(2) 18-21
[9] Imandoust S B and Bolandraftar M 2013 Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background *International Journal of Engineering Research and Applications* **3**(5) 605-610
[10] Pramesti R P A 2013 *Identifikasi karakter plat nomor kendaraan menggunakan ekstraksi fitur ICZ dan ZCZ dengan metode klasifikasi KNN* (Scientific Repository of Bogor Agricultural University)
[11] Ruuska S, Hämäläinen W, Kajava S, Mughal M, Matilainen P and Mononen J 2018 Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle *Behavioural processes* **148** 56-62
[12] Deng X, Liu Q, Deng Y and Mahadevan S 2016 An improved method to construct basic probability assignment based on the confusion matrix for classification problem *Information Sciences* **340** 250-261
[13] Permadi G S, Vitadiar T Z, Kistofer T and Mujianto A H 2019 The Decision Making Trial and Evaluation Laboratory (Dematel) and Analytic Network Process (ANP) for Learning Material Evaluation System *E3S Web of Conferences* **125** 23011
[14] Rodríguez J, Durán C and Reyes A 2010 Electronic nose for quality control of Colombian coffee through the detection of defects in "Cup Tests" *Sensors* **10**(1) 36-46
[15] Xu M, Wang J and Zhu L 2019 The qualitative and quantitative assessment of tea quality based on E-nose, E-tongue and E-eye combined with chemometrics *Food chemistry* **289** 482-489
[16] Permadi G S, Adi K and Gernowo R 2018 Application Mail Tracking Using RSA Algorithm As Security Data and HOT-Fit a Model for Evaluation System *E3S Web of Conferences* **31** 11007
[17] Prehanto D R, Indriyanti A D, Nuryana K D, Soeryanto S and Mubarok A S 2019 Use of Naïve Bayes classifier algorithm to detect customers' interests in buying internet token *Journal of Physics: Conference Series* **1402**(6) 066069
[18] Seka D, Bonny B S, Yoboué A N, Sié S R and Adopo-Gourène B A 2019 Identification of maize (Zea mays L.) progeny genotypes based on two probabilistic approaches: Logistic regression and naïve Bayes *Artificial Intelligence in Agriculture* **1** 9-13
[19] Lawi A and Adhitya Y 2018 Classifying physical morphology of cocoa beans digital images using multiclass ensemble least-squares support vector machine *Journal of Physics: Conference Series* **979**(1) 012029

Classification based on K-Nearest Neighbor and Logistic Regression method of coffee using Electronic Nose

ORIGINALITY REPORT

19%	16%	17%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	e-journal.uajy.ac.id Internet Source	4%
2	eprints.umsida.ac.id Internet Source	4%
3	Rosa Andrie Asmara, Mustika Mentari, Nadia Salsabila Herawati Putri, Anik Nur Handayani. "Identification of Toga Plants Based on Leaf Image Using the Invariant Moment and Edge Detection Features", 2020 4th International Conference on Vocational Education and Training (ICOVET), 2020 Publication	2%
4	www.researchgate.net Internet Source	2%
5	iopscience.iop.org Internet Source	2%
6	erepo.uef.fi Internet Source	1%

7	A D Indriyanti, D R Prehanto, I G L P E Prismana, I K D Nuryana. "The web-based estimation of motorcycles sales using linear regression method", IOP Conference Series: Materials Science and Engineering, 2021 Publication	1 %
8	Xinyang Deng, Qi Liu, Yong Deng, Sankaran Mahadevan. "An improved method to construct basic probability assignment based on the confusion matrix for classification problem", Information Sciences, 2016 Publication	1 %
9	Muhammad Rivai, Eddy Lybrech Talakua. "The implementation of preconcentrator in electronic nose system to identify low concentration of vapors using neural network method", Proceedings of International Conference on Information, Communication Technology and System (ICTS) 2014, 2014 Publication	1 %
10	Edi Sutoyo, Ahmad Almaarif. "Twitter sentiment analysis of the relocation of Indonesia's capital city", Bulletin of Electrical Engineering and Informatics, 2020 Publication	1 %
11	repository.president.ac.id Internet Source	1 %

12	Kaveh Khalili-Damghani, Farshid Abdi, Shaghayegh Abolmakarem. "Solving customer insurance coverage recommendation problem using a two-stage clustering-classification model", International Journal of Management Science and Engineering Management, 2018 Publication	<1 %
13	"Trends in Data Engineering Methods for Intelligent Systems", Springer Science and Business Media LLC, 2021 Publication	<1 %
14	sic.ici.ro Internet Source	<1 %
15	"Innovative Data Communication Technologies and Application", Springer Science and Business Media LLC, 2021 Publication	<1 %
16	Armin Lawi, Yudhi Adhitya. "Classifying Physical Morphology of Cocoa Beans Digital Images using Multiclass Ensemble Least-Squares Support Vector Machine", Journal of Physics: Conference Series, 2018 Publication	<1 %
17	pertambangan.fst.uinjkt.ac.id Internet Source	<1 %

18	"Computational Intelligence in Pattern Recognition", Springer Science and Business Media LLC, 2020 Publication	<1 %
19	Sumita Wardani, Sawaluddin, Poltak Sihombing. "Hybrid of Support Vector Machine Algorithm and K-Nearest Neighbor Algorithm to Optimize the Diagnosis of Eye Disease", 2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT), 2020 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		