ISBN : 978-602-99806-3-9

Classification of broiler chicken eggs using support vector machine (svm) and feature selection algorithm

Intan Y. Purbasari, Fetty T. Anggraeny and Nikolaus R. Harianto
Department of Informatics, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia 60294
E-mail: intanyuniar.if@upnjatim.ac.id

ABSTRACT: According to the National Standardization Agency of Indonesia, consumed chicken eggs are classified based on their eggshell color and weights. This research aimed to incorporate computer vision and machine learning technology to eggs' categorization process as an alternative to the standard and manual method. We used Hue Saturation Value (HSV) to store the eggs' color space and Support Vector Machine (SVM) as the classification algorithm because of its robustness in learning data. A feature selection algorithm, Wrapper, was also applied to increase classification accuracy. The dataset used consists of 60 egg data with eight noted attributes (four of numeric type and four of nominal type with the last attribute as the class): H-value, S-value, V-value, weight, color, density, area, and weight class. The feature selection algorithm evaluated a total number of 29 subsets and found one subset as the candidate, consisting of only one attribute: Area. There were six support vectors found, and the coefficients of the vectors were: 1, 0.668, 0.334, 0.1289, 0.0684, and 0.4688. The classification results with three experiment scenarios have accuracy values of 100%, which was an improvement of the result of the previous work by the authors. This shows that SVM is a good and robust algorithm for classification.

Keywords: Egg Classification, Support Vector Machine (SVM), Wrapper Feature Selection

1. INTRODUCTION

Chicken eggs are one of the many foods consumed by families in Indonesia. According to Statistics Indonesia (Badan Pusat Statistik), the average weekly consumption of chicken egg per capita is 0.194 kg (approximately two eggs per week per capita) in 2015 [1].

The National Standardization Agency of Indonesia classified consumed chicken eggs based on the eggshell's color and weights according to SNI 3926:2008 [2], by US Eggs Grading Manual. Two factors which define an egg's quality: exterior and interior. Exterior assessment includes size, shape, and eggshell's cleanness, while the interior assessment includes air pocket, albumen, and egg yolk's conditions. According to its shell's color, eggs can have white, light brown, or brown color. Egg weight is categorized into three classifications: small (<50g), medium (50-60g), and large (>60g).

Eggs' classification is usually an easy task for human experts, but it is subjective to each expert, and it might take quite some time to sort manually. An egg-grading machine was already invented to

ISBN : 978-602-99806-3-9

help this sorting process done automatically. It consists of several parts: roller, feed conveyor, steering conveyor, sort conveyor, egg sorter, and exit conveyor [3], [7]. In line with the development of image processing technology, several types of research have been done to include image processing and artificial intelligence in the classification process of eggs' sorting.

Research using K-Nearest Neighbor has already been performed to classify egg's quality based on its shell's cleanness with the highest accuracy of 88.89% [4]. Another experiment used image segmentation and regression analysis to predict the weight of broiler chicken eggs [5]. The classification accuracy reached 100%. However, the weight prediction correctness was only 42%. Another research was conducted, which also focused on egg's mass [7]. There were three classifications, and egg's mass was predicted using a function of some variables, and it recommended

some values of length and diameter of eggs to be considered in predicting egg's weight based on outer dimensions.

This research focused on the use of image processing tasks and a learning algorithm, Support Vector Machine (SVM) [8] to classify broiler chicken eggs based on the shell's color and the area of the images. SVM has been recognized as a high-performance algorithm for classification and thus is expected to give accuracy rate higher than 85% on the dataset used.

2. METHODOLOGY

Dataset was obtained by taking images of eggs using a digital camera Canon EOS 500D from 30 cm vertical distance. Egg's position was horizontal and on a black background. Twenty images were taken with a high resolution (3456 x 2304), and they were duplicated twice and had a resolution reduction (1800 x 1200 and 320 x 213), so a total of 60 (sixty) images were collected. A sample of the image is given in Figure 1.

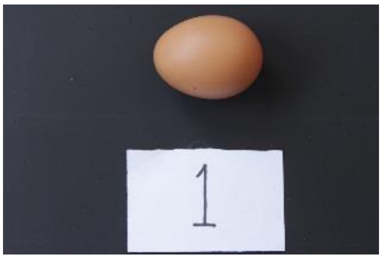

Fig.1 Sample of an egg's image

The images underwent several preprocessing steps:

1. Representing each egg's image with HSV color model
2. Grayscale conversion
3. Applying Sobel filter, dilation, and filling holes

We also took note of each egg's image resolution, color, area, and weight. Table 1 lists twenty four values (out of all sixty data) for all eight attributes where the last attribute is the class attribute. Resolution feature is categorized into three categories: Large (3456 x 2304), Medium (1800 x 1200), and Small (320 x 213). Area attribute is counted as the number of pixels in each image and grouped into three classes: Large (>41000), Medium (35000 - 41000), and Small (<35000). The Class attribute is made into two categories: True and False. Entries in True class are eggs which weigh less than 60g, while those in False class weigh more than 60g.

ISBN : 978-602-99806-3-9

Table 1 Eggs Dataset

H	S	V	Color	Resolution	Area	Weight	Class
0.4879	0.2162	0.3128	Light Brown	Small	Medium	70	F
0.513	0.1662	0.3286	White	Small	Small	60	T
0.5165	0.2266	0.3143	White	Small	Small	60	T
0.4728	0.1946	0.3421	White	Small	Large	80	F
0.4999	0.2324	0.3126	Light Brown	Small	Medium	70	F
0.4986	0.2418	0.3095	Light Brown	Small	Medium	70	F
0.4733	0.2004	0.3398	White	Small	Large	70	F
0.4978	0.2346	0.2914	Light Brown	Small	Medium	70	F
0.4785	0.2238	0.3268	Brown	Medium	Medium	70	F
0.5003	0.1706	0.3441	White	Medium	Small	60	T
0.5038	0.2359	0.3297	Brown	Medium	Small	60	T
0.4569	0.2014	0.363	White	Medium	Large	80	F
0.4878	0.244	0.329	Brown	Medium	Large	70	F
0.4823	0.2503	0.326	Brown	Medium	Medium	70	F
0.4567	0.207	0.3614	Brown	Medium	Large	70	F
0.4854	0.243	0.3059	Brown	Medium	Medium	70	F
0.4504	0.2346	0.3474	Brown	Large	Medium	70	F
0.4758	0.1691	0.3646	White	Large	Small	60	T
0.4789	0.2452	0.3472	Brown	Large	Small	60	T
0.4245	0.2062	0.391	Brown	Large	Large	80	F
0.4605	0.2544	0.3496	Brown	Large	Medium	70	F
0.4541	0.2567	0.3482	Brown	Large	Medium	70	F
0.4237	0.2069	0.391	Brown	Large	Large	70	F
0.459	0.2493	0.3257	Brown	Large	Medium	70	F

Before the classification algorithm was applied to the dataset, we performed a feature selection algorithm. The chosen algorithm was Wrapper since our previous review on various attribute selection algorithms in data mining classification [9] highlighted that Wrapper was a prominent method compared to five other methods investigated despite its time-consuming drawback. In this research, the Wrapper also used Support Vector Machine algorithm as its evaluator (Logistic function as calibration method, using Polynomial Kernel with exponent value of 2, and Best First as the search algorithm). Equation (1) shows the SVM scoring function to compute a score for new input for every data point from i to m [12]:

$$\sum_{i=1}^{m} \alpha_i y^{(i)} K\left(x^{(i)}, x\right) + b \qquad (1)$$

where

- $x^{(i)}$ and $y^{(i)}$ represent the i-th training example (x is an input vector, and y is the class label)
- α_i is the coefficient related with the i-th training example
- x is the input factor to be classified
- K is the kernel function
- b is a scalar value

ISBN : 978-602-99806-3-9

From all sixty data, we experimented with three test options:

1) Use all 60 data as training set and also test set
2) Use cross-validation with ten folds
3) Use 66% data (about 40 data) as training set and 34% (about 20 data) as test set

3. RESULTS AND DISCUSSION

The whole feature selection and classification process were performed using WEKA (Waikato Environment for Knowledge Analysis) version 3.8.1.

3.1. Attribute Selection Result

The Wrapper attribute selection algorithm used in pre-processing was first run on the dataset. At this stage, the process started with no attributes (empty set) and gradually added the primary attribute into the set, also known as forward search direction. We also used SVM as the learning scheme for the Wrapper class with a Linear Kernel. The SVM in WEKA is an implementation of the Sequential Minimal Optimization (SMO) algorithm for training a support vector classifier [10]. All other parameters were set to default values.

The algorithm evaluated a total number of 29 subsets and found one subset as the candidate, consisting of only one attribute: Area. Thus, the number of non-class attributes had been reduced from six to one attribute. Using only this attribute, we then proceeded to the classification process. Table 2 lists the selected attribute:

Table 2 Selected Attribute *Area*

	Label	Count	Weight
1	Medium	32	32.0
2	Small	18	18.0
3	Large	10	10.0

Table 2 showed that 32 data have Medium area, 18 data have Small area, and 10 data have Large area, with each data accounts for one to the total weight.

3.2. Classification Result

After we applied the attribute selection algorithm, the classification process continued by using Support Vector Machine (SVM) algorithm. Again, the implementation used is the SMO algorithm.

There were six support vectors found, and the coefficients of the vectors were: 1, 0.668, 0.334, 0.1289, 0.0684, and 0.4688. Of those six support vectors, two of them belong to the positive class (support vectors with coefficient 1 and 0.334), and the rest belong to the negative class. The value of is 0.333. Figure 2 showed the output of support vector values:

```
     1      * <0 1 0 > * X]
-    0.668  * <1 0 0 > * X]
+    0.334  * <0 1 0 > * X]
-    0.1289 * <0 0 1 > * X]
-    0.0684 * <0 0 1 > * X]
-    0.4688 * <0 0 1 > * X]
-    0.333
```

Fig.2 Output of Support Vector Values

ISBN : 978-602-99806-3-9

Plugging in the values back to Eq. (1) results in Eq. (2):

$$\begin{aligned}
&\boldsymbol{Score\ Function} \\
&= (\mathbf{1})(\mathbf{1})(\boldsymbol{x_2})^{\mathbf{2}} + (-\mathbf{1})(\mathbf{0.668})(\boldsymbol{x_1})^{\mathbf{2}} \\
&+ (\mathbf{1})(\mathbf{0.334})(\boldsymbol{x_2})^{\mathbf{2}} \\
&+ (-\mathbf{1})(\mathbf{0.1289})(\boldsymbol{x_3})^{\mathbf{2}} \\
&+ (-\mathbf{1})(\mathbf{0.0684})(\boldsymbol{x_3})^{\mathbf{2}} \\
&+ (-\mathbf{1})(\mathbf{0.4688})(\boldsymbol{x_3})^{\mathbf{2}} + \mathbf{0.333} \qquad (2)
\end{aligned}$$

If we look at the data, entries in the Positive class (True value) are those having Area=Small, while entries in the Negative class (False value) have Area=Medium and Area=Large. Therefore, we can assume that Area=Medium and Area=Large fall into the same attribute value. Let us suppose that Area=Small equals $\boldsymbol{x}$ and (Area=Medium and Area=Large) equals $\boldsymbol{y}$**.** Thus, Eq. (2) can be rewritten in Eq. (3):

$$\begin{aligned}
&\mathbf{Score\ Function} \\
&\qquad = \mathbf{1.334(x)^2} \\
&\qquad - \mathbf{1.3341(y)^2} \\
&\qquad + \mathbf{0.333} \qquad (3)
\end{aligned}$$

Figure 3 showed the plotting of Score Function which revealed that the decision boundary is a linear plane:

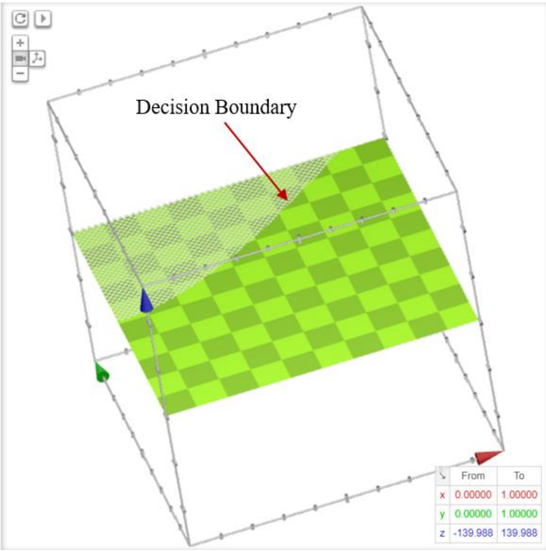

Fig.3 Plotting of Scoring Function

Data above the decision boundary (in the positive y direction) falls into Positive Class (Area=Small), while those under the decision boundary falls into Negative Class (Area=Medium and Area=Large).

(1) The result of using all 60 data as training set and also test set

Table 3 listed the confusion matrix of experiment scenario 1:

ISBN : 978-602-99806-3-9

Table 3 Confusion Matrix of Experiment Scenario 1

	PREDICTED False	PREDICTED True
ACTUAL False	42	0
ACTUAL True	0	18

The accuracy value of scenario 1 is 100%, precision is 100%, and recall is 100%.

(2) The result of using cross-validation with ten folds

Table 4 listed the confusion matrix of experiment scenario 2:

Table 4 Confusion Matrix of Experiment Scenario 2

	PREDICTED False	PREDICTED True
ACTUAL False	42	0
ACTUAL True	0	18

The accuracy value of scenario 2 is 100%, precision is 100%, and recall is 100%.

(3) The result of using 40 data (66%) as training set and 20 (34%) as test set

Table 5 listed the confusion matrix of experiment scenario 3:

Table 5 Confusion Matrix of Experiment Scenario 3

	PREDICTED False	PREDICTED True
ACTUAL False	15	0
ACTUAL True	0	5

The accuracy value of scenario 3 is 100%, precision is 100%, and recall is 100%.

In all three experiment scenarios, the classifier gave a perfect accuracy value. There is no difference in results for these scenarios. The Area attribute selected by the feature selection algorithm can predict the class for each instance accurately. It is expectable because Area is regarded to define the size of each egg. Thus, the less area an egg has, the lighter the egg's weight.

The size of the small dataset used also contributed to the high accuracy value, despite the fact that Support Vector Machine is already known for its robust performance. This strong result was also an improvement of the previous work by the authors which used the same dataset but with ID3 classification algorithm [11]. In the research report, accuracy value achieved was 80%, precision 100%, and recall 75%.

The selection of which kernel functions used apparently also determines the classification results. Polynomial Kernel and Pearson VII function-based Universal Kernel gave 100% accuracy, while Gaussian Kernel only gave 70% accuracy. Considering Gaussian Kernel uses a normal distribution at each data point, the low accuracy percentage might happen because the data distribution is not normal in the dataset used. Therefore, it is necessary to perform a statistical summary on the dataset before applying any feature selection or classification algorithms.

ISBN : 978-602-99806-3-9

4. CONCLUSION

We have experimented on predicting broiler chicken eggs using feature selection algorithm Wrapper and also Support Vector Machine as the classification algorithm. The achieved accuracy value is 100%, which was an improvement of the result of the previous work by the authors.
The Wrapper feature selection algorithm used has successfully selected a single feature most important to determine the class, which was Area. There was no known run time problem because the dataset's size was small.

5. REFERENCES

[1] Statistics Indonesia (Badan Pusat Statistik), Per Capita Weekly Average Consumption of Several Food Items, 2007-2015, 2017, https://www.bps.go.id/linkTabelStatis/view/id/950
[2] National Standardization Agency of Indonesia, Consumed Fresh Eggs, 2008 http://sisni.bsn.go.id/index.php?/sni_main/sni/detail_sni/4363
[3] Sidiq S.A. and Irmawati D., Image processing for egg's identification based on size, J. Electr., Inf., and Voc. Edu. *(ELINVO)* Vol. 1, 2016, pp. 151-156
[4] Trisnaningtyas P.R. and Maimunah, Egg's classification based on shell's cleanness using K-Nearest neighbor, Proc. of Informatics National Conf. Oct 22, 2015 (Bandung: ITB) pp. 241-245
[5] Wijaya T.A. and Prayudi Y., Computer vision implementation and image segmentation to classify broiler chicken eggs, Proc. of National Seminar of Appl. Inf. Tech. (SNATI), June 19, 2010 (Yogyakarta: UII) pp. G1-G5
[6] Rashidi M. and Gholami M., Modeling of egg mass-based on some geometrical attributes American-Eurasian J. Agric. & Environment Sci., 2011, Vol. 10, Issue 1, pp. 9-15
[7] Nopriandi F., Design and Performance of Egg Grading Machine, Postgraduate Thesis, 2015 (Bogor: Postgraduate School-Bogor Agricultural Institute)
[8] Cortes C. and Vapnik V., Support-Vector Networks V. Mach. Learn, 1995, 20:273. https://doi.org/10.1007/BF00994018
[9] Purbasari I.Y. and Nugroho B., Attribute Selection Algorithms Benchmarking in Data Mining Classification, Proc. of National Seminar on Inf. Tech. and Multimedia (SNASTIA), Sept 21, 2013 (Surabaya-University of Surabaya) pp. C47-C54
[10] Platt J.C., Fast training of Support Vector Machines using Sequential Minimal Optimization, Adv. in Kernel Methods – Support Vector Learning, 1998, (Cambridge, Massachusetts: MIT Press) https://www.microsoft.com/en-us/research/publication/fast-training-of-support-vector-machines-using-sequential-minimal-optimization/
[11] Harianto N.R. and Purbasari I.Y., Classification of Broiler Chicken Eggs using Iterative Dichotomizer 3 (ID3) Algorithm, Undergraduate Thesis, 2015 (Surabaya: UPN "Veteran" Jawa Timur)
[12] McCormick C., SVM Tutorial – Part 1, Web article retrieved from http://mccormickml.com/2013/04/16/trivial-svm-example/ April 16, 2013 Accessed: August 29, 2017