International Journal of Engineering & Technology, 8 (1.1) (2019) 173-176

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Classification Of Butterfly Species Based On Venasi Using Support Vector Machine

Asslia Johar Latipah[1]*, Gunawan Ariyanto[2], Rofilde Hasudungan[3], Nani Nurul Fatihah[4]

[1]*Universitas Muhammadiyah Kalimantan Timur*
[2]*Universitas Muhammadiyah Surakarta*
**Corresponding author E-mail: asslia@umkt.ac.id*

Abstract

Butterfly is one of the most frequent object in lab activities of taxonomic courses with venation as a key feature of classification. This research is conducted to see whether this key feature of insect classification can be utilized to classify the type of butterfly with image of venation computationally. Classification process begins with the preprocessing and features extraction, then proceed with data sharing as much as K. Finally the training and testing are conducted using Linear and Non Linear SVM models. Features utilized on this research is a vector with the standard deviation as element of vector as much as quadsplit cutting. 120 schemes were tested for each value of K where K = 2,3 and 5. The highest accuracy attained where K = 2 is 97.05% for Linear SVM and 94.41% for Non-Linear SVM, where K = 5 is 97.64% for Linear SVM and 96.76% for Non-Linear SVM. Lastly, when K = 10 is 97.94% for Linear SVM and 97.94% for Non-Linear SVM. We found that Linear SVM accuracy value remained stable at 1024 cutting image, and accuracy value decreases on Non Linear SVM. Also, The high value of the dimensions of the features can eliminate the non linear nature when is mapped to the kernel.

Keywords: *Linear SVM; Non Linear SVM; Classification; Butterfly; Venation.*

1. Introduction

Classification can be said as a way to classify an object into a group that has been defined by the similarity of the characteristics of the object to be grouped with the characteristics of objects in a group[1]. In the activity of classification of species of insects in the field of Biology, venation becomes the key classification that can be used to distinguish the type of insect[2].

Butterflies are often becomes the object of classification since that the butterfly is an insect which is large enough and easily found in the environment. One of the most commonly used species of butterflies is the butterfly of the Papilionidae family. The selection of butterflies coming from the Papilionidae family is mainly because that the butterfly of this family has wide enough wingspan with minimum level of brittleness. This eases the respective laboratory student to remove the outer layer of the wings to make transparent so that the venation becomes visible.

With the fact that the venation is a regarded as a key features in the butterfly species classification, especially in the field of biology. The author inspired to utilized this venation as a main visual features to perform classification computationally. With this, it can be tested whether venation of the butterfly, can be used as the main visual features on the classification of butterflies. By utilizing the machine learning of subfield of computer science. The example of butterfly wing venation image is shown in Figure. 1.

Fig. 1: Venation of butterfly wings

According Syafruddin[3] and Manimekalai[4], SVM classification algorithm is quite often used in classification of biological data, this is because SVM has the ability to generalize and avoid the occurrence of curse of dimensionality caused by biological data is usually very limited. This is evidenced in a study conducted by Vapnik[5] where the level of generalization obtained by SVM is not influenced by the dimensions of the input vector. This is the reason why SVM is one of the best methods to be used in solving high-dimensional problems with limited sample data.
Linear SVM and Non Linear SVM use the RBF kernel selected as a classification method. Nonlinear SVM uses the selected RBF kernel because according to ketut[6], the RBF kernel is very appropriate to use if the data mapping is not known. Meanwhile, according to Hsu[7], Linear SVM accuracy level can be comparable or better with Non Linear SVM using RFB kernel if cross validation test is elaborated.

2. Research Methods

The classification process begins with the data collection, with the feature extraction up to classification using the SVM method process will be conducted afterwards.
The data in this study is the image of the butterfly wings obtained in Bantimurung National Park, Makassar. The data specifications used are as follows:

1. Image format are .jpg .bmg .png
2. The position of the butterfly image is erect. Not tilted or flat.
3. Image retrieval can be done on the back or at the bottom of the wing.
4. ROI determination is done manually. Image ROI is the right wing.
5. In determining ROI, the empty space around the wings are minimized as small as possible.
6. Image data entered into the system already in the form of the right wing of the butterfly.
7. The butterfly wings must be intact, or in other words no missing pieces of venation.

The name of the butterfly species examined in this his study are shown in Table I below.

Table 1: The list of butterfly classes

Class	Name Species
0	Papilio Polytes
1	Troides Halipron
2	Graphium Androcles
3	Papilio Bluemei
4	Papilio Perantus
5	Graphium Agomaninon
6	Lamptropus Meges
7	Troides Hypolitus
8	Pachilopta Polyphontes
9	Graphium Milon
10	Graphium Meyeri
11	Graphium Deucdlion
12	Papilio Gigon
13	Troides Helena
14	Papilio Sataspes
15	Papilio Ascalapus
16	Chilasa Veiovis

3. Data Preprocessing

Image preprocessing is conducted to produce a good image in the process of feature contraction. In this stage, RGB image conversion is conducted to reduce the image channel to one (grayscale), next the normalization of light intensity with pixel operation followed with the noise removal using median filter. Data preprocessing stage is as follows:

1. Image channel conversion to simplify RGB image attributes from 3 channel or dimension (Red, Green, Blue) into one (Grayscale). This to make the process feature contractions easier.
2. The image is resized to have the same dimensions to speed up the computing process. This process is done by calculating the entire pixel value of the input data data and retrieving the average value of its dimensions as a model. The image of this model will be the reference in the process of resizing on all input images.
3. Normalization of light intensity to get the image with the ideal light intensity.
4. Noise reduction by using median filter. This method is used to remove the salt-papper type noise, ie, disturbances in the image of small dots[8]. The window size is modified in such a way that the number of pixels in the window is odd. When the number of pixels is even, the median value should be taken an average of two pixels in the middle.

4. Feature Estraction

In this process, the image will be pixeled into the vector as a feature with the following stages:

1. Line detection by the Canny method. The image are filtered with a Gausian filter. Which uses a simple matrix smaller than the image size. Then the line edge are detected by the determination of the direction of the line. The image then goes into the hysteresis process to get the binary image using the otsu method. The result of this process is a binary image with a black background and white venation line[8].
2. Cutting the image into blocks by using quadsplit cutting model. The concept of this quadsplit is to cut the image into several blocks

according to the cut level which is a multiple of the power of 4. If the cutting level is one, then the image will be cut into 4 parts whereas if the cut level two then the image will be cut as much as 16 or 42 and so on. The cutting process of the image will be governed by the level of cutting to see the effect of cutting the image with the level of accuracy.

3. Calculating the standard deviation value (σ) of each cutting block. This standard deviation value is the feature in which a vector (V) feature is formed using equation (1).

$$V = (\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n) \quad (1)$$

With σ_n denoting the default value of the n_{th} deviation, where n is the number of cutting blocks in the previous stage, whereas M x n denotes the number of pixels of an image. Thus, a vector will have N elements as much as the number of deductions as the x value, which pairs up with the class label as the y value. The vector formed can be written in equation (2):

$$x_m = (x_1, x_2, \ldots, x_n; y_i) \quad (2)$$

where,
xi: Training data
n: many cutting blocks
yi: Class label (i: -1, + 1)
m: 1,2,3,.... amount of training data
This resulted to more cuts equated to the higher the data dimension of the training. The example of the vector with the number of cuts four shown in equation (3) below:

$$x_m = (x_1, x_2, x_3, x_4; y_i) \quad (3)$$

5. Classification with SVM

Classification was performed using two SVM algorithms : Linear SVM and Non Linear SVM. In this stage, the data is divided into two, namely the training data and test data. The first step is to do the training, the training is done to form the hyperplane model and get the required parameter value[9] which is then followed by the test using the test data. On this stage, we used available LibSVM Library.

6. Result And Discussion

The test scheme is a scenario in testing to see which process plays a major role on the accuracy level.
The testing process is done by the cross validation scheme with 120 test schemes in each value k. The k values used are k = 2, k = 5 and k = 10. Later from each scheme, the accuracy results will be analyzed. The scheme is divided into four sections. Each section has its own process sequence. Table 2 shows the scheme used.

Table 2: Scheme of the classification process of butterfly classification

No Schemes	Process Sequences	Processes in trial	Many experi-ments
1	○ Image Conversion ○ Resize ○ Canny edge detection ○ Quadsplit Cutting ○ Classification with SVM	• Cutting level 1 - 5 • Linear and Non Linear SVM	10 experi-ments
2	○ Image Conversion ○ Resize ○ Normalization ○ Canny edge detection ○ Quad Split Cutting ○ Classification with SVM	• Normalization • Cutting level 1 - 5 • Linear and Non Linear SVM	10 experi-ments
3	○ Image Conversion ○ Resize ○ Median Filter ○ Quadsplit Cutting ○ Classification with SVM	• Median filter with window 1 - 5 • Cutting level 1-5 • Linear and Non Linear SVM	50 experi-ments
4	○ Image Conversion ○ Resize ○ Normalization ○ Median Filter ○ Quadsplit Cutting ○ Classification with SVM	• Normalization • Median filter with window 1 - 5 • Cutting level 1-5 • Linear and Non Linear SVM	50 experi-ments
Total scheme			120 experi-ments

From the entire testing processes, we obtained the results with various accuracy values. The highest accuracy values of each scheme are shown in Table 3 and table 4.

Table 3: Highest accuracy results for Linear SVM in all four test schemes

K-Fold	Scheme 1	Scheme 2	Scheme 3	Scheme 4
K=2	95,88 %	97,05 %	87,35 %	88,52 %
K=5	96,76 %	97,64 %	97,05 %	97,53 %
K=10	97,94 %	97,94 %	97,64 %	97,94 %

Table 4: Highest accuracy results for Non Linear SVM in all four test schemes

K-Fold	Scheme 1	Scheme 2	Scheme 3	Scheme 4
K=2	93,52%	94,41%	85,88%	88,52%
K=5	95,00%	95,29%	96,76%	96,76%
K=10	95,00%	94,41%	95,88%	97,64%

From the experiments conducted, we found conclusive some results:

a. The smaller the Gamma value, the higher the level of accuracy gained when classifying the Non Linear SVM using the RBF kernel. However it should still be adjusted to the value of parameter C.
b. For Linear SVM, the combination of the best parameter values is C = 0.1 and the parameter value Gamma = 1. With the highest accuracy value of 97.94%.
c. For Non Linear SVM, the combination of the best parameter values is C = 62.5 x 106 and the parameter value Gamma = 1 x 10-5. With the highest accuracy value 97.94%.

When the image region cutting are proceed as much as 1024 sub-images, the performance of Non-linear SVM decreases. This impairment of accuracy occurs in all test schemes and for all K values. This proves that no other process affects impairment accuracy other than the cutting process, with the image entered in accordance with the provisions.

From the result, we observed that this cutting process is importnat since that the standard deviation value formed from the cutting result has a value that is not unique for each image data. The standard deviation value produces many support vectors of equal value. This results in the number of slack variables that can decrease the accuracy value because SVM Non Linear with RBF kernel must combine parameter value C with Gamma parameter. The greater the parameter value of C, the number of data that gets pinalty will be greater as well.

Also considering the effect of σ on the parameter of Gamma which must be adjusted carefullly. That if it is too large, especially on the process of mapping the data into the higher dimension, will resulted to tendenciton away from its non-linear form. In other hand, if it is too small, the function will be irregular. This is in accordance with the one mentioned by Hsu [7] where the ability of the RBF kernel will decrease if the size of the data is much smaller (less) than the large feature.

From the whole experiment, the process of determining ROI, resizing process, and cutting process is a very influential process on the level of accuracy. Errors in classification are largely influenced by the three processes. Here are some factors that can make the accuracy value decrease based on our finding on the test results:

1. Determination of ROI is diverse.
2. Enter data has an invisible noise
3. Improper image selection of the normalization model.
4. The input image has a dimension too high with a small amount of data.

7. Conclusion

From the research that has been done, we obtained some conclusions, among others, as follows:

1. The highest level of accuracy that can be achieved by Linear SVM and Non Linear SVM with the RBF kernel is the same ie 97.94% with different schemes.
2. At the cut of 1024, the Linear SVM performance remains stable while the Non Linear SVM decreases. This is because Linear SVM only needs to set the parameter value of C, whereas in Non Linear SVM, must combine the parameter value of Gamma and parameter C. Where in mapping is too high, Non Linear data will be lost if setting σ value on Gamma parameter is too high but the separator function will be unstable if it is too small.
3. The process of determining ROI, Resizing and cutting process, are three processes that contribute the greatest influence to the value of accuracy in the classification.

References

[1] Prasetyo, E., 2012, *Data mining. Konsep dan Aplikasi menggunakan MATLAB,* ANDI, Yogyakarta.
[2] Borror, D. J., Triplehorn, C. A., and Jhonson, N. F., 1996, *Pengenalan Pelajaran Serangga,* Ed.6, diterjemahkan oleh Partosoedjono, Gadjah Mada University Press, Yogyakarta.
[3] Syafruddin, S,. 2013. *Model sistem cerdas untuk deteksi awal penebangan liar kawasan hutan pada daerah aliran sungai.* Universitas Hasanuddin, Makassar.
[4] Manimekalai, K dan Vijaya, MS., 2014, Support Vector Machine Based Tool For Plant Species Taxonomic Classification, *Journal of Asian Scientific Research*, Vol 4, hal 159-173.
[5] Vapnik V.N. 2014. *"The Nature of Statistical Learning Theory"*, Springer-Verlag, New York Berlin Heidelberg.
[6] Ketut, I, P., 2015, *Support Vector Machine Pada Information Retrieval*. Universitas Pendidikan Ganesha. Bali.
[7] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin.,2016, *A Practical Guide to Support Vector Classification*. Department of Computer Science and Information Engineering, National Taiwan University, Taipe, Taiwan.
[8] Kadir, A., and Susanto, A., 2013, *Teori dan Aplikasi Pengolahan citra*, ANDI, Yogyakarta.
[9] Fan, R.-E., Chen, P.-H., Lin, C.-J., 2005. Working set selection using second order information for training support vector machines. J. Mach. Learn. Res. 6, 1889–1918.