Cocoa Beans Digital Image Classification Based On Color Features using Multiclass Ensemble Least-Squares Support Vector Machine

Yudhi Adhitya[1], Armin Lawi[2], Hartono[3] and Mursalin[4]
{armin@unhas.ac.id }

[1]Department of Informatics, Al-Asy'ariah Mandar University, Polewali Mandar, Indonesia
[2]Department of Computer Science, Hasanuddin University, Makassar, Indonesia
[3]Department of Computer Sciences, STMIK IBBI, Medan, Indonesia
[4]Department of Mathematics Education, Universitas Malikussaleh, Aceh Utara, Indonesia

Abstract. This research aims to determine the quality of cocoa beans through their fermentation status of their digital images. Samples of cocoa beans were scattered on a bright white paper under a controlled lighting condition. A compact digital camera was used to capture the images. The images were then processed to extract their color parameters. Classification process begins with an analysis of cocoa beans image based on color feature extraction. Parameters of visual classification of cocoa beans were obtained by extraction of color feature parameters, i.e.: Red (R), Green (G), Blue (B), Hue (H), Saturation (S) and Value (V). Then the beans are classified into 3 classes, i.e., Fermented Beans, Un-Fermented Beans and Moldy Beans. The classification process using the Multiclass Ensemble Least-Squares Support Vector Machine (MELS-SVM) method starts with the training process to get the model, and then the model is used in the testing process to get accuracy. The classification model of input parameters from our 1,604 cocoa beans images based on the color features obtained accuracy of 99.281%.

Keywords: The Cocoa Beans Digital Image, Multiclass Ensemble Least-Squares, Vector Machine

1 Introduction

Indonesia is one of the producers of cocoa beans [1], most of cocoa production exported to America, Singapore, Malaysia, Brazil, and China. Indonesia's cocoa beans production significantly increases, but the quality produced is very low and various of them are less fermented, not dry enough, beans size is evently high skin content, high acidity, taste is very diverse and inconsistent. This is reflected in the relatively low price of Indonesian cocoa beans and discounted prices compared to similar products from other producer countries [2]. Cocoa plant (Theobroma cacao L.) is one of the important plantation commodities as a source of industrial raw materials trade commodities that can increase the country's foreign exchange and income of cocoa farmers.

Indonesian cocoa farmers generally apply a variety of fermentation ways in terms of beans quantity, fermentation means and time. Fermentation is done in baskets, simple wooden crates

WMA-1 2018, January 20-21, Indonesia

DOI 10.4108/eai.20-1-2018.2281929

or plastic bags. What farmers do is not real fermentation because most farmers keep the harvested beans in plastic bags for 1-2 days then dried by drying in direct sunlight on cement floors, mats or woven bamboo. The requirements or conditions used to determine the quality of cocoa beans in Indonesia are contained in the Indonesian national standard of cocoa beans SNI 2323-2008. The Indonesian national standard regulates the classification of the quality of dry cocoa beans as well as general requirements and in particular to maintain the consistency of the quality of the cocoa beans produced. At the exporter level, the separation is carried out by using machinery primarily for the classification of cocoa beans. The results of this cocoa bean sorting will be determined by taking samples of cocoa beans to be analyzed in the laboratory in accordance with the standards of cocoa beans quality classification.[3]

Quality examination of cocoa beans is done using traditional and manual procedures, namely by using the visual method on cocoa beans by choosing one by one cocoa beans. The human vision must accurately see the object on the surface of the cocoa beans. In plain view, a human without special knowledge can differentiate and classify cocoa beans. Usually, they only armed with experience and knowledge gained ealier. However, manual checks have limitations such as tired eyes and different analytical results of each examiner.

This classification system has two important aspects, image analysis and pattern recognition. Image analysis has standard techniques for identifying, measuring, and acquiring large quantities of quantitative data. Image processing techniques include image capture, pre-processing, interpretation, quantization and image classification. But unfortunately, the resulting image is still not in accordance with the results expected by the user. Therefore, the existence of a process that can process an image is needed by the user. The discipline that gave birth to the techniques to process the image is called Digital Image Processing (Digital Image Processing). [4].

2 Literature Review

Image processing techniques have been widely used in the field of agriculture such as determining the type of defects of coffee beans, edamame quality determination, quality inspection of RSS rubber, mango quality determination, identification of maturity level of lemon and mangosteen, identification of defective cocoa beans, quality determination of cocoa beans.

I Wayan Astika, Mohamad Solahudin, Andri Kurniawan, Yunindri Wulandari (2010), use ANN structures to develop the relationship between input parameters, quality components of cocoa beans and outputs. ANN classified cocoa beans into 3 parameters namely: Fermented Beans, Un-Fermented Beans and Moldy Beans.The classification of fermentation status has an accuracy of 88.54%, which consists of95.62% for fermented beans,81.72% for moldy beans, and71.43% for non fermented beans.[5]

S. Nurmuslimah (2016), created a software system that starts with taking pictures of files to display on the system interface, image is processed using edge detection sobel to get numeric data value. Furthermore, these data are used as data input training Neural Network Backpropagation. After training data is obtained, then the data is used for the testing process. From the testing process, the output of the created system is able to provide information about the quality of cocoa beans. Using Backpropagation method with alpha = 0.6, hidden layer = 3, fault tolerance = 0.0001, target = 0.9, resulting in a system that has a level of accuracy of (76%) has an error rate (24%) in determining the quality of cocoa beans.[6]

Data mining is a process that uses statistical techniques, calculations, artificial intelligence and machine learning to extract and identify useful information and related knowledge from large databases [7]. SVM initially can only classify data in two classes. However, further research SVM was developed so that it can classify data over two classes (multiclass) [8]. Classifying M-classes means predicting the class label $C_m, m = 1, \dots, M$ one way to solve the M-class problem by formulating it into binary L classification problems [9].

SVM concept is simply described as trying to find the best hyperplane that serves as a separator of two classes in the input space. Pattern which is a member of two classes: +1 and -1 and share alternate field separators. The best dividing fields can not only separate the data but also have the largest margins. Margin is the distance between the field of separator (hyperplane) with the closest pattern of each class.

Let $\{x_1, \dots, x_n\}$ be the dataset and $y_i \in \{+1, -1\}$ is the class label of the x_i data. The two classes are separated by a pair of parallel bounding plane. The first delimiter field limits the first class while the second delimiter field limits the second class, so it is obtained [9]:

$x_i.w + b \geq +1$ for $y_i = +1$

$x_i.w + b \leq -1$ for $y_i = -1$ (1)

The best dividing fields with the largest margin values can be formulated into quadratic programming problems:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i \qquad (2)$$

with constraints $y_i(x_i.w + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where $l = 1, .., n$ is a slack slack variable that determines the level of misclassification of the data samples. While $C > 0$ is a parameter.

Method for classifying data that can not be separated linearly is kernel method. The kernel method transforms the data into the feature space dimension so that it can be linearly separated on the feature space. The kernel method can be formulated:

$$K(\boldsymbol{x_i}.\boldsymbol{x_j}) = \varphi(\boldsymbol{x_i}).\varphi(\boldsymbol{x_j}) \qquad (3)$$

Commonly used kernel functions are as follows:

a. The linear kernel: $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{x_i^T x_j}$

b. Kernel polynomial:

$$K(x_i, x_j) = (\gamma.x_i^T x_j + r)^p, \gamma \geq 2$$

c. RBF Kernel (Radial Basis Function): $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp\left(-\gamma \|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2\right), \gamma > 0$

LS-SVM was first introduced by Suykens and Vandewalle in 1999. LS-SVM is one of the SVM modifications that solves linear equations. If the SVM separator field is given as in (3), then for LS-SVM is given as follows:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + \frac{C}{2} \xi^T \xi \qquad (4)$$

with constraint $y_i(x_i.w + b) \geq 1 - \xi_i$

The above equation can be solved after forming Lagrangian:

$$L = \frac{1}{2} \|w\|^2 + \frac{C}{2} \xi^T \xi - \sum_{i=1}^{l} \alpha_i \left(y_i(\varphi(x_i).w + b) - 1 + \xi_i\right)$$

(5)

where α_i is a Lagrangian multiplier whose value can be either positive or negative.

To optimize the conditions in (5), a decrease of w, b, ξ, and α is equal to zero. The results of the process are as follows:

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{l} \alpha_i y_i \varphi(x_i) \qquad (6)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (7)$$

$$\frac{\partial L}{\partial \xi} = 0 \rightarrow \alpha = \gamma \xi_i, \; i = 1, \dots, N \quad (8)$$

$$\frac{\partial L}{\partial \alpha} = 0 \rightarrow y_i(\varphi(x_i).w + b) - 1 + \xi_i = 0, i = 1, \dots, N$$

Using the One Against All (OAA) method, a binary binary model is constructed k (k is the number of classes). Each i-class model is trained by using the entire data. For example, there is a classification problem with 3 classes. For training use 3 pieces of binary classification. Its objective function:

$$\min_{w_i, b_i, \xi_{t,i}} \frac{1}{2} \sum_i^m (w_i)^T w_i + \frac{C}{2} \sum_i^m \xi_{t,i}^2 \quad (10)$$

with constraint $y_{t,i}(\varphi_i(x_t)(w_i)^T + b_i) \geq 1 - \xi_{t,i}$.

Confusion matrix is a table that states the amount of test data that is correctly classified and the amount of test data being misclassified

True Posstive (TP), ie the number of documents from class 1 is correctly classified as class 1.

True Negative (TN), ie the number of documents from class 0 is are correctly classified as class 0.

False Positive (FP), ie the number of documents from class 0 incorrectly classified as class 1.

False Negative (FN) is the number of documents from class 1 that are misclassified as class 0.

The calculation of accuracy is expressed by the equation:

$$\text{Akurasi} = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%$$

The calculation of sensitivity is expressed by the equation:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\%$$

The calculation of false discovery rate is expressed by the equation:

$$\text{False Discovery Rate} = \frac{FP}{FP + TP} \times 100\%$$

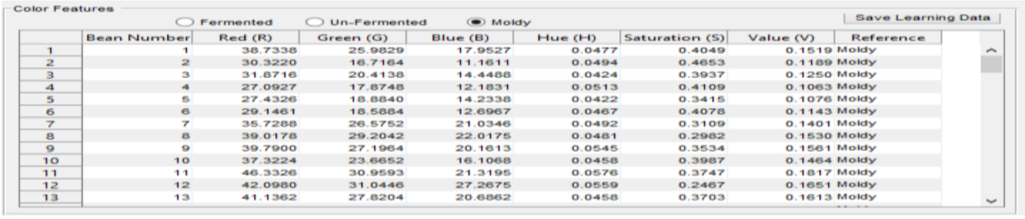
3 Methods

The process of classifying cocoa beans is explained by the following activities:

1. Cocoa bean image capture
2. Extraction of morphological features using image processing.
3. Conducting the training process to obtain the classification model
4. Conducting the classification process using test data
5. From the classification process in obtaining the results of the classification of cocoa beans based on morphological features.

3.1 Cocoa Beans Color Features

The most commonly used color feature models are RGB and HSV components. Digital image of cocoa beans that have been taken, processed using grayscale and thresholding and then done edge detection by using kernel sobel to detect the edge of the image. After that the process of segmentation by masking and separating the image of cocoa beans per unit of seed. The separated cocoa bean image was then extracted by color extracting feature by extracting RGB mean data and HSV mean data



	Bean Number	Red (R)	Green (G)	Blue (B)	Hue (H)	Saturation (S)	Value (V)	Reference
1	1	38.7338	25.9829	17.9527	0.0477	0.4049	0.1519	Moldy
2	2	30.3220	16.7164	11.1611	0.0494	0.4653	0.1189	Moldy
3	3	31.8716	20.4138	14.4488	0.0424	0.3937	0.1250	Moldy
4	4	27.0927	17.8748	12.1831	0.0513	0.4109	0.1063	Moldy
5	5	27.4326	18.8840	14.2338	0.0422	0.3415	0.1076	Moldy
6	6	29.1461	18.5884	12.6967	0.0467	0.4078	0.1143	Moldy
7	7	35.7288	26.5752	21.0346	0.0492	0.3109	0.1401	Moldy
8	8	39.0178	29.2042	22.0175	0.0481	0.2982	0.1530	Moldy
9	9	39.7900	27.1964	20.1613	0.0545	0.3534	0.1561	Moldy
10	10	37.3224	23.6652	16.1068	0.0458	0.3987	0.1464	Moldy
11	11	46.3326	30.9593	21.3195	0.0576	0.3747	0.1817	Moldy
12	12	42.0980	31.0446	27.2675	0.0559	0.2467	0.1651	Moldy
13	13	41.1362	27.8204	20.6862	0.0458	0.3703	0.1613	Moldy

Figure 1. Results of the image analysis process for color features using matlab software

3.2 Dataset

Based on the color features are divided into three classes namely, first class for fermentedbeans, second grade for unfermented beans, and third grade for moldy beans. Attributes for dataset 1 consist of six pieces: Red (R), Green (G), Blue (B), Hue (H), Saturation (S) and Value (V). The data sample for dataset 1 is 1,604 items.

Variables used in this research are:

yi=class of dataset that is class 1, class 2, class 3 and class 4.

xi =data processed by LSSVM technique, xi consists of Red (R), Green (G), Blue (B), Hue (H), Saturation (S) dan Value (V).

α=Lagrange multiplier.

w =normal field to support vector.

b =distance of the bounding plane to the center point.

Values of *α* and *b* will be obtained after the training is completed. Value of α is then used to find the value of *w*.

3.3 Classification Process

The classification process is divided into two, training and testing process. The data used is the data of color feature extraction and morphological features as much as 100% data so it is expected to get the accuracy of multiclass classification on LS-SVM. Training process aims to build the MELS-SVM model by finding the parameters, ie values α, w, and b. After forming the MELS-SVM model, proceed with the testing process on the data to see the accuracy of the LS-SVM technique using the One Against All method.

Based on the workflow, the classification process begins with the training process on the train data to obtain α and b values using the Matlab r2016a software and the additional toolbox LSSVMlabv which can be downloaded on the site http://www.esat.kuleuven.be/sista/lssvmlab/toolbox.html. In the training process that generates α and b values, the RBF kernel is used. Furthermore, after the obtained values of α and b, the next step is to find the value of w then the value of w and b are used to arrange the separator function.

After getting the separation function model from the training post, proceed to the testing process. The testing process begins by inserting the item values (xi) of the test data, then obtaining the prediction class. From this prediction class will be calculated the accuracy level of the method by finding the total class that is correctly predicted by the class of the test data.

3.4 Multiclass Algorithm Methods

The analysis used in this multiclass classification is the One Against All classification method. The multilingas method algorithm as follows:

1. Dataset input.
2. Identify the input dataset
   a. The values of the training data feature (x_i)
   b. Class of training data (*yi*)
   c. Values feature test data (xt_i)
   d. Class of test data (yt_i)
3. Initiate objects on LS-SVM before performing the training process with the initlssvm function
   a. Specify data of training data feature (x_i)
   b. Specifies the training data class (y_i),
   c. Choose a classifier to classify data
   d. Selects the *kernel* and its parameters to use
4. Selecting the multilingual method code used (*code_OneVsAll* for *One Against All*)
5. Conduct training process with trainlssvm function
6. Calculating values *w*
7. Make predictions based on the model obtained and determine data feature test data (xt_i) with simlssvm function
8. Create a *confusion matrix*
9. Calculate the level of accuracy with the formula:

$$\lambda = \frac{C}{N} \times 100\%$$

where *C* is the correct total of predictions and *N* is the total of all data tested.

3.5 Implementation Results on Dataset

The separator function for the one against all method with the RBF (Radial Basis Function) kernel using the parameter $\sigma = 0.5$ for dataset is as follows:

$$f_1(\boldsymbol{x}) = \boldsymbol{x}_i.\boldsymbol{w}_1 + b = \boldsymbol{x}_i.\begin{pmatrix} 1{,}125.403 \\ 721.711 \\ 435.615 \\ -1{,}554.422 \\ 1{,}125.771 \\ 1{,}095.667 \end{pmatrix} + (-0{,}3984) \qquad (1.1)$$

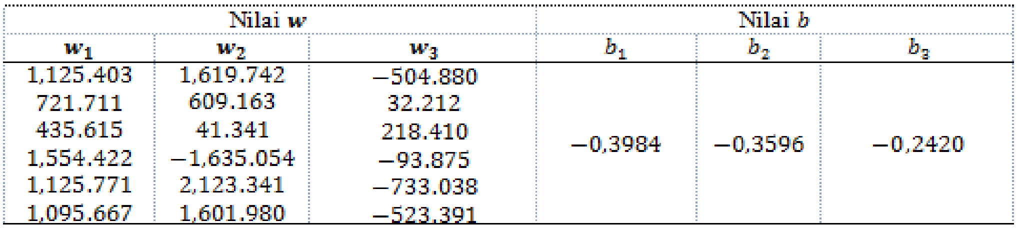
$$f_2(\boldsymbol{x}) = \boldsymbol{x}_i.\boldsymbol{w}_2 + b = \boldsymbol{x}_i.\begin{pmatrix} 1{,}619.742 \\ 609.163 \\ 41.341 \\ -1{,}635.054 \\ 2{,}123.341 \\ 1{,}601.980 \end{pmatrix} + (-0{,}3596) \qquad (1.2)$$

$$f_3(\boldsymbol{x}) = \boldsymbol{x}_i.\boldsymbol{w}_3 + b = \boldsymbol{x}_i.\begin{pmatrix} -504.880 \\ 32.212 \\ 218.410 \\ -93.875 \\ -733.038 \\ -523.391 \end{pmatrix} + (-0{,}0372) \qquad (1.3)$$

4 Results And Discussion

Values of *w* and *b* are based on RBF kernel and its parameters for the use of the One Against All method on dataset can be seen on table 1.

Table 1: *w* and *b* values



Nilai w			Nilai b		
w_1	w_2	w_3	b_1	b_2	b_3
1,125.403	1,619.742	−504.880	−0,3984	−0,3596	−0,2420
721.711	609.163	32.212			
435.615	41.341	218.410			
1,554.422	−1,635.054	−93.875			
1,125.771	2,123.341	−733.038			
1,095.667	1,601.980	−523.391			

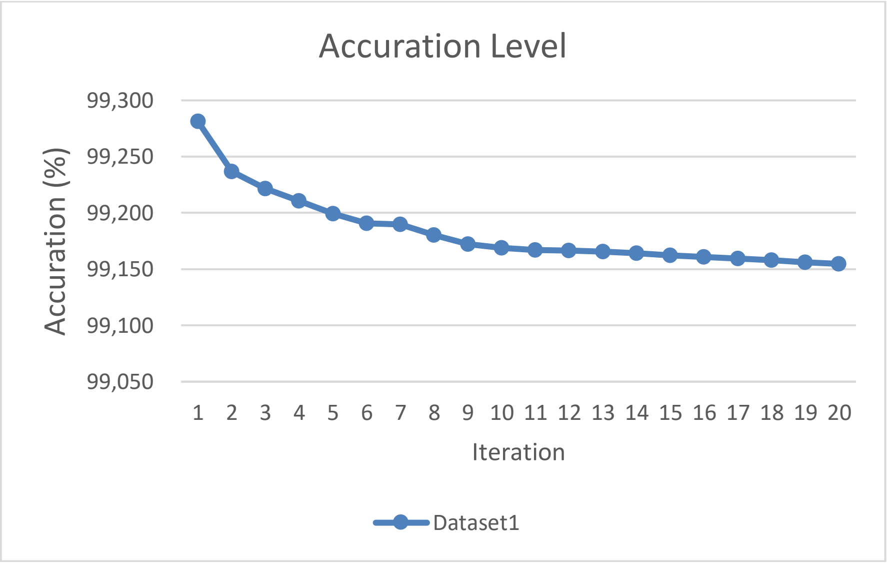


Figure 2. Accuracy level graph of One Against All method for each kernel in dataset

Based on Table 1 and Figure 2, the use of RBF kernel types and using the parameter σ = 0.5 has the highest accuracy on dataset has the highest accuracy (99.281). Accuracy for classification with the number of classes : 3 classes, i.e.: Fermented Beans, Un-Fermented Beans and Moldy Beans.

.

5 Conclusions

This research, the cocoa beans image is processed using grayscale and thresholding and then edge detection is done by using a sobel kernel to detect the edges of the image. After segmentation process which can extract feature that produces data based on the extraction of color feature, which resulting color feature parameters : Red (R), Green (G), Blue (B), Hue (H), Saturation (S) and Value (V).

The accuracy level is derived from the number of data items categorized into the correct class by the LSSVM model. The level of accuracy using one against all method on datasets with RBF type (Radial Basis Function) using parameter σ = 0.5 resulting 99,281%. With three class classification, ie. Fermented Beans, Un-Fermented Beans and Moldy Beans.

Acknowledgements

The authors would like to express heartfelt thanks to Department of Computer Science, Hasanuddin University and Department of Informatics, Al-Asy'ariah Mandar University, Polewali Mandar, Department of Computer Science, Malikussaleh University, for the research support and assistances.

References

[1] ICCO Annual Report 2012/2013, International Cocoa Organization, (2013).

[2] M. Haryadi, dan Supriyanto, Pengolahan Kakao Menjadi Bahan Pangan, Pusat Antar Universitas Pangan dan Gizi. Universitas Gajah Mada, Yogyakarta, (2001)

[3] Biji kakao [SNI] Standard Nasional Indonesia, SNI; (SNI 2323:2008, ICS 67.140.30 Badan Standardisasi Nasional), (2008)

[4] T. Sutoyo, Edy Mulyanto, Vincent Suhartono, Dwi Nurhayati, Wijanarto Oky, Teori Pengolahan Citra Digital, Yogyakarta: ANDI, (2009).

[5] I.W. Astika, M. Solahudin, A. Kurniawan, Y. Wulandari, Determination of Cocoa Bean Quality with Image Processing and Artificial Neural Network, Department of Agricultural Engineering Bogor Agricultural, University Bogor, Indonesia, (2010)

[6] S. Nurmuslimah, Implementasi Metode Backpropagation Untuk Mengidentifikasi Jenis Biji Kakao Yang Cacat Berdasarkan Bentuk Biji, Jurusan Sistem Komputer, Fakultas Teknologi Informasi, Institut Teknologi Adhi Tama Surabaya (ITATS), (2016).

[7] Turban, E., dkk. 2005. Decision Support Systems and Intelligent Systems. Yogyakarta: Andi Offset.

[8] Sembiring, K. 2007. Tutorial SVM Bahasa Indonesia. Bandung: Institut Teknologi Bandung.

[9] Gestel, T., Suykens, J., Lanckriet, G., Lambrechts, A., Moor, B., & Vandewalle, J. 2002. Multiclass LS-SVMs: Moderated Outputs and Coding-Decoding Schemes. Neural Processing Letters, 45-58.

[10] Jafar, N., Thamrin, S.A., Lawi, A., 2016, Multiclass Classification using Least Squares Support Vector Machine, Proceeding of International Conference on Computational Intelligence and Cybernetics.

[11] Kurniawan, D dan Supriyanto, C. 2013. Optimasi Algoritma Support Vector Machine (SVM) Menggunakan AdaBoost untuk Penilaian Risiko Kredit. Jurnal Teknologi Informasi.

[12] Marliani R, R., Lawi, A., & Thamrin, S. A. 2016. Algoritma Adaboost untuk Optimasi Ensemble Least Squares Support Vector Machine. Jurusan Matematika Fakultas

Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

[13] Nurkamila, J. Lawi, A., & Thamrin, S. A. 2016. Least Squares Support Vector Machine Untuk Klasifikasi Data Multikelas. Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

[14] Lawi, A., Adhitya, Y. 2017. Classifying Physical Morphology of Cocoa Beans Digital Images using Multiclass Ensemble Least-Squares Support Vector Machine. Proceeding of International Conference on Science. (to appear)

[15] Vapnik, V., & Cortes, C. 1995. Support-Vector Networks. Machine Learning, 273-297.

[16] Pratt, William K. 2007, Digital image processing, PIKS Scientific inside, John Wiley, 4th Edition.

[17] Xu, Y., Lv, X., Wang, Z., & Wang, L. 2014. A Weighted Least Squares Twin Support Vector Machine. Journal of Information Science and Engineering, 1773-1787.

[18] Zhou, L., Lai, K. K., dan Yu, L. 2010. Least Squares Support Vector Machines Ensemble Models for Credit Scoring. Expert Systems with Applications, 127-133.