Multiclass Classification using Least Squares Support Vector Machine

Nurkamila Jafar
Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University
Makassar, Indonesia
e-mail: nurkamila.jafar@gmail.com

Sri Astuti Thamrin
Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University
Makassar, Indonesia
e-mail: tuti@unhas.ac.id

Armin Lawi
Department of Computer Science, Faculty of Mathematics and Natural Sciences, Hasanuddin University
Makassar, Indonesia
e-mail: armin@unhas.ac.id

***Abstract*—In this paper, multiclass classification problem; One Against All and One Against One, with Least Squares Support Vector Machine (LS-SVM) will be used. There are three type of kernels were used in this paper; Radial Basis Function (RBF), polynomial and linear. One Against All method and One AgainstOne method will be compared to see the accuracy of each kernel, and the amount of misclassification using the confusion matrix. This is illustratedby using iris plant species dataset and the preferred method of contraception dataset. The results showed that the method of One AgainstOne is better than the One Against All based on the accuracy for kernelsRBF, polynomial, and linear.**

Keywords—Accuracy rate; Radial Basis Function;Linear; Multiclass; One Against All; One Against One; Polynomial

I. INTRODUCTION

Support vector machine (SVM) is one of algorithms that is very popular in the classification techniques used in data mining [13]. The basic idea of SVM is to make a hyperplane to separate the two class linearly [1, 2, 6, 7, 12, 13]. SVM is used to resolve two class classification, while in the real-world problems often requires classification of more than two class or called multiclass. To resolve the problem multiclass, some methods such as One Against All and One Against Onewas developed [7, 8]. One Against All method and One Against One method are based on the combining of some binary classification. Let the data consists of N class, with One Against All method, the training data will be as much as N times, while in the One Against One method, the training is conducted as a combination of two of N [8].

One-Against-All method with One-Against-One method using SVM was compared by Rahayu et al. [7]. They presented that One-Against-One method has a higher accuracy rate than One Against All method. Suykens and Vandawelle [9] introduced Least Squares SVM (LS-SVM). LS-SVM has better generalization capability and faster computing time than SVM [9].

The aim of thispaperis to compare the level of accuracy and the amount of misclassification of One Against All and One Against One methods using LS-SVM technique. The developed method will be applied in two datasets; iris plant species dataset and the preferred method of contraception.

The paper is organised as follows. In Section 2, material and methods related to LS-SVM will be explained. Results will be presented in Section 3. Section 4 will be explained the conclusion.

II. MATERIAL AND METHODS

A. Data Source

There are two datasets were used in this paper namely iris plant species and the preferred method of contraception. These data can be downloaded on the http://www.idvbook.com/wp-content/uploads/2010/02/iris_data_set.zip for iris plant species dataset and http://sci2s.ugr.es/keel/dataset/data/ classification/ contraceptive.zip for the preferred method of contraception dataset.

The iris plant species is called as dataset 1 and the preferred method of contraception is called as dataset 2. There is 150 items in dataset and it is divided into three classes; setosa, versicolor, and virginica. Then, the attributes for this dataset comprises two pieces; the length and the width of blade and the length and the width of calyx. They are measured in centimeters.

The preferred method of contraception is called as dataset 2 and it is consists of 1473 items. In this dataset, there are three classes of methods used namely non method, a long-term, and a short-term. The attributes for dataset 2 comprises three pieces; the amount of children was born, wife's employment status (0=yes, 1=no) and standard of life index (1=low, and 2,3,4=high).

B. SVM Classifier

The concept of SVM is to find the best hyperplane to separate two classes in input space. The pattern form of the element from two classes; +1 and -1. The best hyperplane is not only separate data but also has the maximum margin. Margin is the distance between the hyperplaneand nearby pattern of each classes. For example, $\{x_1,..,x_n\}$is the input

978-1-5090-5548-7/16/$31.00 ©2016 IEEE

of pattern and $y_i \in \{+1,-1\}$ is the output of pattern of x_i. The classifier is constructed as follow [8]:

$$x_i.\mathbf{w} + b \geq +1 \text{ for } y_i = +1,$$
$$x_i.\mathbf{w} + b \leq -1 \text{ for } y_i = -1. \quad (1)$$

The best hyperplane with the largest margin value can be formulated into a quadratic programming problem, that is:

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

with the constraint $y_i(\mathbf{x}_i.\mathbf{w} + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where $i = 1,...,n$, and ξ_i is the slack variables that determinate the level of classification error from sample data. While $C > 0$ is a parameter [14].

For classifying data that cannot be separated linearly, we can use thekernel method. Kernel method can transform the data into dimensional feature space,so that it can be separated linearly on the feature space. Kernel method can be formulated:

$$K(\mathbf{x}_i.\mathbf{x}_j) = \varphi(\mathbf{x}_i).\varphi(\mathbf{x}_j). \quad (3)$$

The kernel functions are commonly used are as below:

- Linear kernel: $K(\mathbf{x}_i.\mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$
- Polynomial kernel: $K(\mathbf{x}_i.\mathbf{x}_j) = (\gamma.\mathbf{x}_i^T\mathbf{x}_j + \gamma)^p, \gamma \geq 2$
- RBF kernel: $K(\mathbf{x}_i.\mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \gamma \geq 0$

C. LS-SVM Classifier

LS-SVM was first introduced by Suykens and Vandewalle in 1999[3, 4, 5, 9, 11]. LS-SVM is one of modification of SVM that can solve the linear equations. If the SVM's hyperplane in the equation (3), then for LS-SVM is given as follows:

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

with the constraint $y_i(\mathbf{x}_i.\mathbf{w} + b) \geq 1 - \xi_i$.

The (4) can be solved after forming the Lagrange[14]*:*

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\xi^T\xi - \sum_{i=1}^{l}\alpha_i(y_i(\varphi(\mathbf{x}_i).\mathbf{w} + b) - 1 + \xi_i), \quad (5)$$

where α_i is Lagrange multiplier which value may be positive or negative.

To optimize condition of the equation (5), do differential on ***w****, b,* ξ*,* and α with equal to zero. The results of the process are as below:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{l}\alpha_i y_i \varphi(\mathbf{x}_i), \quad (6)$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow \sum_{i=1}^{l}\alpha_i y_i = 0, \quad (7)$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow \alpha = \gamma\xi_i, i = 1,...,N, \quad (8)$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow (y_i(\varphi(\mathbf{x}_i).\mathbf{w} + b) - 1 + \xi_i) = 0, i = 1,...,N. \quad (9)$$

D. Multiclass

One Against All method is used by constructing the k binary classification model, which k is the sum of classes [7, 8]. Each classification model *i*-th is trained to use all the data. For example, there are classification problem with three classes, for training, we used three binary classification. This can be seen in Table 1. The objective function [10]:

$$\min_{\mathbf{w}_i,b_i,\xi_{t,i}} \frac{1}{2}\sum_{i=1}^{m}(\mathbf{w}_i)^T\mathbf{w}_i + \frac{C}{2}\sum_{i=1}^{m}\xi_{t,1}^2, \quad (10)$$

with constraint $y_{t,i}(\varphi_i(\mathbf{x}_t).(\mathbf{w}_i)^T + b_i) \geq 1 - \xi_{t,i}$.

TABLE I. THREE BINARY CLASSIFICATION MODELS OF ONE AGAINST ALL METHOD

$\mathbf{y}_i = 1$	$\mathbf{y}_i = -1$	**Hyperplane Function**
Class 1	Not Class 1	$f^1(\mathbf{x}_t) = \mathbf{x}_t(\mathbf{w}^1) + b^1$
Class 2	Not Class 2	$f^2(\mathbf{x}_t) = \mathbf{x}_t(\mathbf{w}^2) + b^2$
Class 3	Not Class 3	$f^3(\mathbf{x}_t) = \mathbf{x}_t(\mathbf{w}^3) + b^3$

By using the One Against One method, we construct $\frac{k(k-1)}{2}$ binary classification model. Each classification model is trained from the *i*-th class and *j*-th class. The classification models for One Against One method can be seen in Table 2. The objective function is given in equation (10).

TABLE II. THREE BINARY CLASSIFICATION MODELS OF ONE AGAINST ONE METHOD

$\mathbf{y}_i = 1$	$\mathbf{y}_i = -1$	**Hyperplane Function**
Class 1	Class 2	$f^{12}(\mathbf{x}_t) = \mathbf{x}_t(\mathbf{w}^{12}) + b^{12}$
Class 1	Class 3	$f^{13}(\mathbf{x}_t) = \mathbf{x}_t(\mathbf{w}^{13}) + b^{13}$
Class 2	Class 3	$f^{23}(\mathbf{x}_t) = \mathbf{x}_t(\mathbf{w}^{23}) + b^{23}$

E. Modified Confusion Matrix

Confusion matrix is a table that states the amount of test data that is correctly classified and the amount of test data incorrectly classified. For multiclass problem, modified confusion matrix is given in Table III.

From Table III, K_{11} is the amount of items from class 1 which was correctly classified as class1. K_{12} is the amount of items from class 1 which was incorrectly classified as class 2. Moreover, K_{13} is the amount of items from class 1 which was incorrectly classified as class 3 and so on.

TABLE III. CONFUSION MATRIX FOR CLASSIFICATION THREE CLASS

		Prediction Class ($\hat{y}_i$)		
		Class 1	**Class 2**	**Class 3**
Actually Class (y_i)	**Class 1**	K_{11}	K_{12}	K_{13}
	Class 2	K_{21}	K_{22}	K_{23}
	Class 3	K_{31}	K_{32}	K_{33}

The accuracy can be calculated by using the equation (11) as follow:

$$Accuracy = \frac{K_{11} + K_{22} + K_{33}}{K_{12} + K_{13} + K_{21} + K_{23} + K_{31} + K_{32}} \times 100\% \quad (11)$$

III. RESULTS

A. Misclassification

Total misclassification using One Against All method and One Against One method in both datasets with three types of kernel can be seen in Table IV and Table V.

TABLE IV. TOTAL MISCLASSIFICATION USING ONE AGAINST ALL METHOD FOR EACH TYPE KERNELS AT DATASET 1 AND DATASET 2

Kernel Type and Parameter		**Total Misclassification**	
		***Dataset* 1**	***Dataset* 2**
RBF	$\alpha = 0.5$	0	324
Polynomial	Degree 3	0	350
Linear		2	374

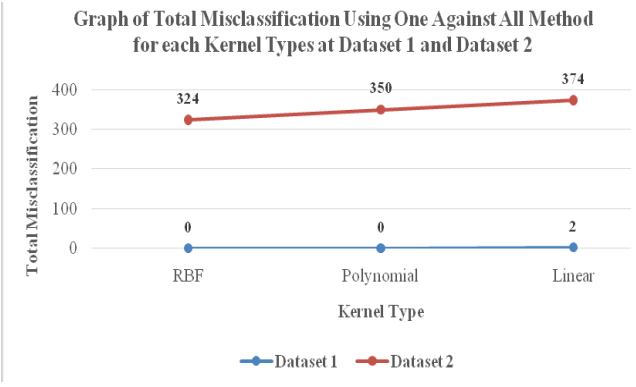

Fig. 1. Graph of total misclassification using One Against All method for each kernel type at dataset 1 and dataset 2

Table IV and Fig. 1 explain the sum of the smallest misclassification using One Against All at dataset 1 with kernel RBF (0) and polynomial (0). At dataset 2 RBF kernel gives the smallest sum of misclassification (324). Linear kernel gives the biggest sum of misclassification at dataset 1 (2) and dataset 2 (374).

TABLE V. TOTAL MISCLASSIFICATION USING ONE AGAINST ONE METHOD FOR EACH TYPE KERNELS AT DATASET 1 AND DATASET 2

Kernel Type and Parameter		**Total Misclassification**	
		***Dataset* 1**	***Dataset* 2**
RBF	$\alpha = 0.5$	0	233
Polynomial	Degree 3	0	217
Linear		0	246

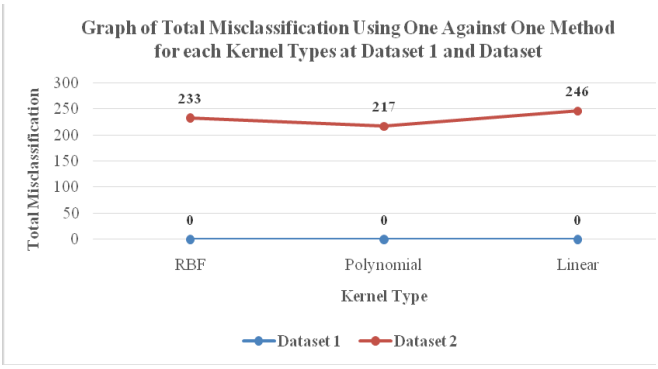

Fig. 2. Graph of total misclassification using One Against One method for each kernel type at dataset 1 and dataset 2

Table V and Fig. 2 show that there is no misclassification at dataset 1 (0) by using One Against One method with RBF kernel, polynomial, and linear, respectively. While in the dataset 2, the smallest sum of misclassification is polynomial kernel (217) and the largest sum of misclassification is the linear kernel (246).

B. Accuracy Rate

Accuracy rate is obtained from sum of data items that are categorized into the right classes by LS-SVM method. Accuracy rate using one against all method and one against one method at both datasets with three types of kernel can be seen in Tables VI and VII.

TABLE VI. ACCURACY RATE KERNEL TYPES AT DATASET 1 AND DATASET 2 WITH USING ONE AGAINST ALL METHOD

Kernel Type and Parameter		**Accuracy Rate (%)**	
		***Dataset* 1**	***Dataset* 2**
RBF	$\alpha = 0.5$	100	26.531
Polynomial	Degree 3	100	20.635
Linear		95.556	15.193

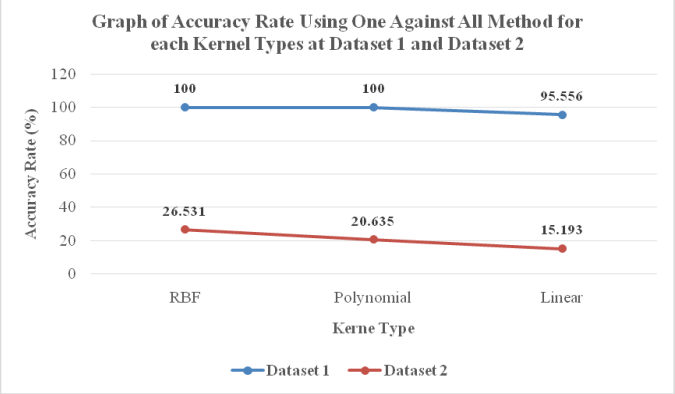

Fig. 3. Graph of accuracy rate of One Against All method for each kernel type at dataset 1 and dataset 2

As can be seen in Table VI and Fig. 3, based on the accuracy rate, the RBF kernel and polynomial kernel have the highest accuracy rate at dataset 1 (100).However, in the dataset 2, the RBF kernel has the highest accuracy rate (26.531). Meanwhile, the linear kernel has the lowest accuracy rate on dataset 1 (95.556) and dataset 2 (15.193).

TABLE VII. ACCURACY RATE KERNEL TYPES AT DATASET 1 AND DATASET 2 WITH USING ONE AGAINST ONE METHOD

Kernel Type and Parameter		Accuracy Rate (%)	
		Dataset 1	*Dataset* 2
RBF	$\alpha = 0.5$	100	47.166
Polynomial	Degree 3	100	50.794
Linear		100	44.218

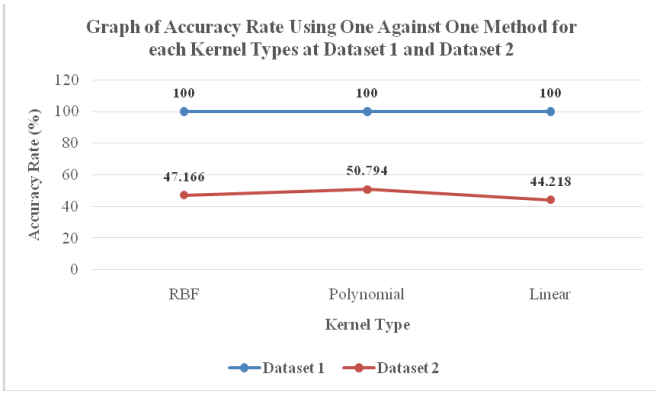

Fig. 4. Graph of accuracy rate of One Against One method for each kernel type at dataset 1 and dataset 2

From Table VII and Fig. 4, by using One Against One method with the RBF kernel, polynomial kernel and linear kernel, the accuracy rate of dataset 1 reaches 100% at datasets 1. For the dataset 2, the highest accuracy rate was resulting in the choice of a polynomial kernel (50.794) and the smallest in linear kernel (44.218).

IV. CONCLUSION

This paper compared the One Against All method and One Against One method using LS-SVM classification technique. This method was applied in two datasets; iris plant species and the preferred method of contraception, respectively.

Overall the results showed that in term of the accuracy, the method of One Against One is better than the One Against All for these two datasets. This result is associated with the work of [9] that is LS-SVM has better generalization capability and faster computing time than SVM. This one of the advantage for using LS-SVM in multiclass classification problem. For further research, it is recommended to use other classification techniques of variation SVM and other multiclass methods to see the capability of the classification method.

REFERENCES

[1] Anggraini S, R., Lawi, A., & Thamrin, S. A. 2015. *Optimasi Ensemble Support Vector Machine dengan Algoritma Adaboost*.Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

[2] Burges, C. J. 1998. *A Tutorial on Support Vector Machine forPattern Recognition.* Boston: Kluwer Academic Publisher.

[3] Gestel, T. V., & Suykens, J. A. 2004. Benchmarking Least Squares Support Vector Machine Classifiers. *Machine Learning*, 5-32.

[4] Gestel, T., Suykens, J., Lanckriet, G., Lambrechts, A., Moor, B. D., & Vandewalle, J. (2002). Bayesian Framework for Least-Squares Support Vector Machine Classifiers, Gaussian Processes, and Kernel Fisher Discriminant Analysis. *Neural Computation*, 1115-1147.

[5] 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.

[6] Gunn, S. 1998. *Support Vector Machines for Classification and Regression.* Southampton: University of Southampton.

[7] Rahayu, M. S., Lawi, A., & Thamrin, S. A. 2015. *Perbandingan Teknik Klasifikasi Multiclass Menggunakan Support Vector Machine*. Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

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

[9] Suykens, J., & Vandewalle, J. 1999. Least Squares Support Vector Machine Classifier. *Neural Processing Letters*, 293-300.

[10] Suykens, J., & Vandewalle, J. 1999. *Multiclass Least Squares Support Vector Machine*.

[11] Suykens, J., Gestel, T., Brabanter, J., Moor, B., & Vandewalle, J. 2002. *Least Squares Support Vector Machines.* Belgium: World Scientific Publishing Co. Pte. Ltd.

[12] Tomasouw, B. P., & Irawan, M. I. 2012. Multiclass Twin Bounded Support Vector Machine untuk pengenalan Ucapan. *Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan*.

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

[14] 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.