Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

The visualization and classification method of support vector machine in lymphoma cancer

To cite this article: B C Kristina *et al* 2020 *J. Phys.: Conf. Ser.* **1613** 012065

The visualization and classification method of support vector machine in lymphoma cancer

B C Kristina[1], A F Hadi[1], A Riski[1], A Kamsyakawuni[1], D Anggraeni[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Science, University of Jember, Jember, Indonesia

E-mail: dian_a.fmipa@unej.ac.id

Abstract. In the classical-classification multivariate process, it becomes an interesting topic to be discussed in the research area because of the larger variables with smaller observations. For this we need a method that can handle this problem. One answer is to use machine learning. SVM is a classification method in machine learning that is able to classify these data types. In addition, SVM can also model and classify relationships between variables efficiently and easy interpretation. This paper aims to create a visualization of SVM classifiers, then obtain an accuracy value to have an optimal classification with a misclassification of small numbers. This study aims to find good SVM input parameters by assessing the importance of variables using visual methods. This visualization will distinguish groups of people who contract diffuse lymphoma cancer and follicular lymphoma cancer with data on the genetic expression of lymphoma cancer. The classification using kernel Linear, Gaussian RBF, Polynomial and Sigmoid. The best classification accuracy using linear kernel functions with training data has a classification accuracy of 100% and testing data has a classification accuracy of 94, 73%.

1. Introduction

Every year there are estimated 12 million people worldwide sufferings from cancer and 76 million of them die. In Indonesia, cancer is one of the causes of death and is a very big disease problem because until now there is no known definite cause of cancer. The types of cancer are breast cancer, neck cancer, lung cancer, liver cancer, ovarian cancer and lymph node cancer (lymphoma cancer). This study raised the case of lymphoma cancer because lymphoma cancer ranked 10th most cancer in Indonesia in 2010 and 2011 although the number of lymphoma cases is actually still relatively low but the number of lymphoma cases continues to increase rapidly each year [1]. When compared with normal people, the number of patients with cancer is very small, while the gene variables obtained in patients who have cancer are more numerous. In this case, the observation variable is bigger than the observation data so that it takes machine learning method. Therefore, the need for related analysis of contracting lymphoma cancer risks classification by mutation of genetic expression in lymphoma cells to be measured from the Affymetrix Gene Chip software. The form of genetic mutation data is microarray datasets, where the number of observed variables much more than individuals who were observed [2]. Support Vector Machine is a classification method in machine learning that is able to classify data types. SVM can classify relationships between variables without the need for strict assumptions, efficient and easy interpretation [3].

Based on previous research, SVM often classifies cancer data and gets high accuracy values. A high level of accuracy is believed to support cancer diagnosis. Research by Novianti [4] is a diagnosis of breast cancer patients using logistic regression and SVM based on the results of mammography has a classification accuracy of 94.34% while logistic regression of 84.90%. Research by Shofi [5] is the result of breast cancer classification showing that the use of Smooth Support Vector Machine (SSVM) has a value with an accuracy of 99.63% better than the Multivariate Adaptive Regression Splines (MARS) model of around 95.88%. Research from Puspitasari [6] is a method of SVM in classifying dental and oral diseases with an accuracy of 94.44%. Based on the explanation above, many studies have discussed the classification with the SVM method, in this study will use visualization of the results of the classification lymphoma cancer gene expression with the SVM method.

These experiments primarily consist of either monitoring each gene multiple times under many conditions [7], or alternately evaluating each gene in a single environment but in different types of tissues, especially cancerous tissues [8]. Of the thousands of variables will be visualized into 2 classes including diffuse large B-cell lymphoma cancer and follicular lymphoma so that the best hyperplane will be obtained from lymphoma cancer data [9]. Visualization of SVM output can help to understand the results of the algorithm SVM [10].

2. Method

2.1. Lymphoma Cancer Gene Expression

Frozen diagnostic nodal tumor specimens from 58 DLBCL (Diffuse Large B-Cell Lymphoma) patients and 19 FL (Lymphoma Follicle) patients were selected for this study. Scans were carried out on the Affymetrix scanner and the expression values for each gene were calculated using Affymetrix Gene Chip software [11]. DNA microarray experiments generating thousands of gene expression measurements are being used to gather information from tissue and cell samples regarding gene expression differences that will be useful in diagnosing disease [12].

2.2. Classification

Classification is a method of grouping data that will learn training data using a classification algorithm. As for several classification algorithms, including Bayesian Classification, K-Nearest Neighbor, Decision Tree Induction, Case-Based Reasoning, Genetic Algorithms, Discriminant Analysis, and Support Vector Machines [13]. Experimental and evaluation shows that SVM, KNN and NB are traditional classification texts. Experiments and evaluations show valid clarification texts [14].

Measurement of classification performance can describe how well the classifier is in classifying data. Confusion matrix is one method that can analyze how well the classifier recognizes tuples from each class classification [15]. Confusion matrix is a table recording the results of classification work, can be seen in Table 1.

Table 1. Confusion Matrix

Original class	Prediction class	
	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

Based on the table, the classification performance of accuracy and precision can be calculated. The accuracy value is the value of how big the accuracy of the data classification results, while precision is the ratio between the true positive class and all positive result classes. Comparing precision and accuracy parameters can be used as a reference in determining the best classification method [16].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} * 100\% \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} * 100\% \quad (2)$$

2.3. Machine Learning

Machine learning is a method that allows machines to gain knowledge for problem solving by showing old cases accordingly. Machine learning considers the use of artificial intelligence (AI) methods although some of these technologies do not show intelligence directly but they are very useful for the design of Intelligent Decision Support Systems [17]. The method in machine learning consists of two approaches, namely supervised learning and unsupervised learning. Supervised learning is the process of inducing knowledge from a series of observations where the expected results are known beforehand. Usually supervised learning uses existing data [18].

2.4. Support Vector Machine (SVM)

Support Vector Machines (SVM), a supervised machine learning technique, have been shown to perform well in multiple areas of biological analysis including evaluating microarray expression data [19], detecting remote protein homologies [20], and recognizing translation initiation sites [21]. We have also recently become aware of another effort that uses SVM in analyzing expression data [22]. Support Vector Machine (SVM) is a learning system that uses a hypothetical space in the form of linear functions in a high-dimensional feature and is trained by using learning algorithms that are based on optimization theory [23]. The level of accuracy in the model that will be generated by the transition process with SVM is very dependent on the kernel function and parameters used [24]. Based on its characteristics, the SVM method is divided into two, including Linear SVM and Non-Linear SVM. Linear SVM is linearly separated data, which separates the two classes on the hyperplane with soft margins. While Non-Linear SVM is implementing the function of kernel tricks on high-dimensional space [25].

The basic concept of SVM is to find the best hyperplane separating two classes. As Figure 1 below shows the separation of 2 classes, namely class -1 and class +1 with the SVM method. Data is denoted as $x_i \in \mathbb{R}^D$ While each label is denoted $y_i \in \{-1, +1\}$ for i = 1,2, ..., n, and i is the amount of data.

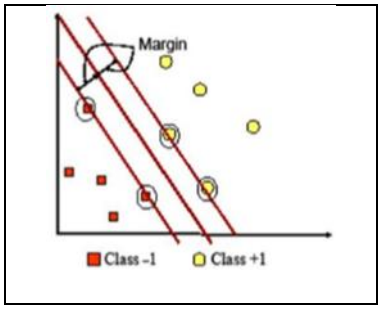

Figure 1. Classification of two classes with the SVM method

Both classes are assumed -1 and +1 can be separated completely by hyperplane dimension d, which is defined

$$\vec{w}.\vec{x} + b = 0 \quad (3)$$

Information:

w: weighting vector
x: training data
b: bias

Pattern $\vec{x}_1$ which belongs to the class -1 (negative samples) can be formulated as a pattern that satisfies inequality

$$\vec{w}.\vec{x}_1 + b \leq -1 \quad (4)$$

while the pattern that belongs to the class +1 (positive samples)

$$\vec{w}.\vec{x}_1 + b \geq -1 \tag{5}$$

The margin can be found by maximizing the value of the distance between the hyperplane and the closest point, that $\frac{1}{||\vec{w}||}$ ($\|w\|$ is the norm of the vector w) [26].

It can be formulated as a Quadratic Programming (QP) problem, that is finding the minimum point equation 4, with these limits

$$\text{By minimizing:} \|w\|^2 = w^T w \tag{6}$$

$$\text{With the provision of : } y_i(w.x_i + b) - 1 \geq 0, \forall_i \tag{7}$$

This problem can be solved by a variety of computational techniques, including Lagrange Multiplier.

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i (w.x_i + b - 1) \tag{8}$$

α_i are Lagrange multipliers, which is zero or positive ($\alpha_i \geq 0$). The optimal value of the equation 5 can be calculated by minimizing L to and b, and maximize L against α_i. With regard to the nature that at the optimal point gradient L = 0, equation 5 can be modified as the maximization problem only contain α_i. Just as the equation 2.6 below.

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j x_i x_j \tag{9}$$

$$\alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{10}$$

From the result of this calculation is obtained α_i the most positive value. Data were correlated with positive α_i called a support vector [27].

The explanation is the assumption that both classes can be separated perfectly by the hyperplane. However, generally two classes in the input space cannot be separated completely. SVM technique reformulated by introducing soft margins to solve this problem. In soft margin, equation 3 is modified by slack variable ξ_i ($\xi_i \geq 0$) as,

$$y_i(w.x_i + b) \geq 1 - \xi_i, \forall_i \tag{11}$$

Thus the equation becomes:

$$\text{minimize: } \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \, 1 \tag{12}$$

C selected parameters to control the trade-off between the margins and the error classification. Great value C means it will provide a greater penalty against the classification error.

To solve non-linear problems, SVM is modified by including kernel functions. In the non-linear SVM, the data $\vec{x}$ mapped by the function $\phi(\vec{x})$ vector space into a higher dimension. In this new vector space, a hyperplane that separates these two classes can be constructed. Illustration of this concept can be seen in Figure 2.

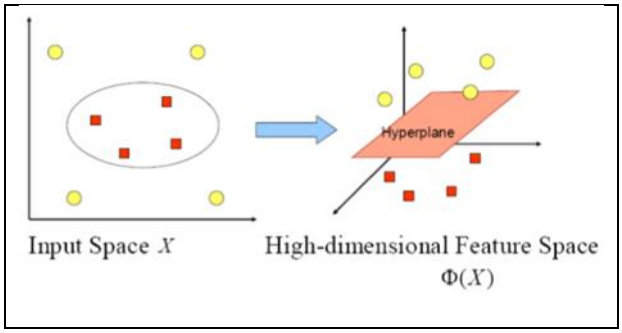

Figure 2. The function maps ϕ data to vector spaces of higher dimensions

Some kernel functions that are commonly used for the function of a Linear kernel, Polynomial, Gaussian and Sigmoid [26]. Summary of some kernel functions are included in Table 1.

Table 2. Commonly Used Kernels in SVM

Kernel Function	Definition
Linear	$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i . \vec{x}_j)$
Polynomial	$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i . \vec{x}_j + 1)^D$
Gaussian	$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\left\|\vec{x}_i - \vec{x}_j\right\|}{2\sigma^2}\right)$
Sigmoid	$K(\vec{x}_i, \vec{x}_j) = tanh(\alpha\vec{x}_i . \vec{x}_j + \beta)$

2.5. Principal Component Analysis

Principal Component Analysis (PCA) is a method that used to reduce a number of data dimensions to smaller dimensions. The purpose of PCA is to explain part of the variation in the set of variables observed on the basis of several dimensions. From a variable that is changed to many variables [28]. Principal component is a linear combination of variables that are observed, information contained in the principal component is a combination of all variables with a certain weight. The selected linear combination is a linear combination with the largest variety that contains the most information [29]. The package used to be able to run PCA on the R program is the psych package.

2.6. Support Vector Machine in R Program

R is a language computer and an interactive programming environment for data analysis and graphs. The main objective of environmental R is for enable and encourages the creation of good data analysis. In the R program, there are several packages that can be used to facilitate the processing of data as desired. One package in R used in this research that the package Support Vector Machine (SVM).There are four related packages existing SVM in R program is a package e1071, kernlab, Klar, and svmpath. Each package SVM has a different function. E1071 package works for completion method parameters and visualization. Kernlab package serves to optimize the kernel base so as to provide the result of SVM implementation is flexible and expandable. Klar package serves to implement the analysis SVM classification as a regular separator. Svmpath package serves to provide the appropriate algorithm with the results of the SVM solution [30].

2.7. Dataset

This study will use microarray gene expression data from Shipp [11] produced at the Whitehead Institute. The data used by the authors include secondary data obtained from the website

[https://ico2s.org/datasets/microarray.html]. The Data obtained has a sample of 77 objects for 2647 lymphoma cancer genes. The variables in this study are the independent variable (x) and the dependent variable (y). The independent variable (x) in this study is in the form of gene expression data of 2647 variables with a total observation (n) of 77 objects. The dependent variable (y) is a variable that contains classes consisting of two categories, namely the Diffuse (D) category and the Follicle (F) category.

2.8. Application

The research steps in this study are attached as follows:

1. Take observational data on lymphoma cancer gene expression.
2. Transforming gene expression data using Principal Component Analysis.
3. Install the SVM package and a number of packages needed in the R program.
4. Import the data that has been transformed into the R program.
5. Data is divided into two stages, namely the training and testing phase by dividing the training data by 75% and testing data by 25% randomly.
6. Determine the kernel functions to be used in SVM, namely the *Linear, Polynomial, Gaussian, and Sigmoid* kernels.
7. Obtained the results of SVM classification predictions with the best kernel.
8. Visualization with the SVM classification plot results with the kernel used

3. Result and Discussion

The SVM results in linear kernels obtained training predictions of 100% and testing of 94,73%. In the RBF kernel the training prediction is 100% and the testing test is 84,21%. In Sigmoid kernel, training prediction is obtained 96,55% and testing is 89,47%. In the Polynomial kernel the training prediction is 86,20% and the testing test is 68,42%. Thus, the best classification accuracy uses the linear kernel function are included in Table 2.

Table 3. Classification Using SVM in Linear Kernel

Parameter		
SVM – Type	:	C – Classification
SVM – Kernel	:	Linear
Cost	:	1
Number of *Support vector*	:	31

Based on the Table 2 the classification results obtained using a linear kernel and the value of cost parameter 1 with a support vector of 31. From the SVM training process we get the confusion matrix in Table 3.

Table 4. Confusion Matrix Training Process with Linear Kernel

Confusion matrix		Predictive Value	
		Diffuse	*Follicle*
True Value	*Diffuse*	42	0
	Follicle	0	16
Accuracy		1	

Table 3 explains the results of the classification process there are 16 training data classified into 42 diffuse data class and 16 follicle data class. The level of prediction accuracy in training data is 100%. From the SVM testing process we get the confusion matrix in Table 4.

Table 5. Confusion Matrix Testing Process with Linear Kernel

Confusion matrix		Predictive Value	
		Diffuse	*Follicle*
True Value	*Diffuse*	15	1
	Follicle	0	3
Accuracy		0,9473684	

Table 4 explains the results of the classification process there are 58 training data classified into 15 diffuse data class and 3 follicle data class with 1 data misclassification. The level of prediction accuracy in testing data is 94,73%. For visualization, dimension reduction must be done using PCA method. After dimension reduction, created SVM process and plotting data is generated in Figure 3.

Figure 3. Plotting Data Visualization

From Figure 3 it is known that the red pattern is the classification of diffuse class data, while the blue pattern is the classification of follicle class data. Purple pattern is a support vector. Then get a hyperplane by maximizing the margin of the support vector that supports in Figure 4.

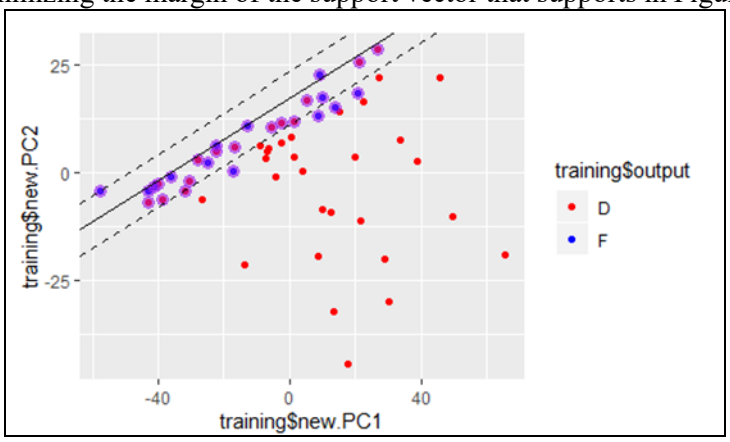

Figure 4. Process Training Data Visualization

In Figure 4 that a straight line is a hyperplane that separates the diffuse class and the follicle class So that the best hyperplane is obtained to fix 2 classes.

4. Conclusion

The best classification results for lymphoma cancer gene expression are using a 100% linear kernel in training data and 94.73% in testing data. To visualize the results of the classification, it is necessary to do a PCA which aims to reduce the data dimensions. Furthermore, using the SVM method, it is

obtained the plot of Diffuse class and the Follicle class, so that the best hyperplane is obtained that separates the two classes.

Acknowledgement
For the preparation of this paper, we thank to all member Data Science Research Group, and all member of the Statistical Laboratory, Department of Mathematics, University of Jember, Indonesia.

5. References

[1] Departemen Kesehatan 2015 *Menkes canangkan komitmen penanggulangan kanker di Indonesia*. Diakses pada tanggal 21 Mei 2019 dari [http://www.depkes.go.id/].
[2] Singh D, Febbo P G, Ross K, Jackson D G, Manola J, Ladd C, Tamayo P, Renshaw A A, D'Amico A V, Richie J P, Lander E S, Loda M, Kantoff P W, Golub T R, and Sellers W R 2002 Gene Expression Correlates of Clinical Prostate Cancer Behavior Cancer Cell 1 pp 203-09.
[3] Darsyah M Y 2013 Menakar Tingkat Akurasi Support Vector Machine Study Kasus Kanker Payudara Statistika Jurnal Statistika 1 pp 15-20.
[4] Novianti F A and Purnami S W 2012 Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik Dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi Jurnal Sains dan Seni ITS 1 pp D147- 51.
[5] Andari S, Purnami S W and Otok B W 2013 Smooth Support Vector Machine Dan Multivariate Adaptive Regression Splines Untuk Mendiagnosis Kanker Payudara Jurnal Statistika 1 pp 37-47
[6] Puspitasari A M, Ratnawati D A and Widodo A W 2018 Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 2 pp 802-10.
[7] Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D and Futcher B 1998 Comprehensive identification of cell cycle-regulated genes of the yeast Saccha- romyces cerevisiae by microarray hybridization Mol. Biol. Cell 9 pp 3273–97.
[8] DeRisi J, Penland L, Brown P, Bittner M, Meltzer P, Ray M, Chen Y, Su Y and Trent J 1996 Use of a cdna microarray to analyse gene expression patterns in human cancer Nat. Genet 4 pp 457–460.
[9] Glaab E, Bacardit J, Garibaldi J M, and Krasnogor N 2012 Using Rule-Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data *PLoS ONE* **7** pp 1-18.
[10] Caragea D, Cook D and Honavar V G 2005 Visual Methods for Examining Support Vector Machine Results, with Applications to Gene Expression Data Analysis *Computer Science Technical Reports* pp 1-25.
[11] Shipp M A, Ross K N, Pablo T, Weng A P, Kutok J L, Aguiar R C T, Gaasenbeek M, Angelo M, Reich M, Pinkus G S, Ray T S, Koval M A, Last K W, Norton A, Lister T A, Mesirov J, Neuberg D S, Lander E S, Aster J C, and Golub T R 2002 Diffuse large Bcell lymphoma outcome prediction by gene-expression profiling and supervised machine learning *Nature medicine* **8** pp 68–74.
[12] Terrence S, Furey N, Cristianini N, Duffy D, Bednarski W, Schummer M, and Haussler D 2000 Support vector machine classification and validation of cancer tissue samples using microarray expression data *Bioinformatics* **16** pp 906–914.
[13] Khan, Aurangzeb B, Baharudin L, Lee H, and Khan K 2010 A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology* **1** pp 4-20.
[14] Yao and Zhi-Min 2012 An Optimized NBC Approach in Text Classification. *Physics Procedia* **24** pp 1910-14.

[15] Han J, Kamber M, and Pei J 2012 *Data Mining: Concept and Techniques, Third Edition*. Whaltam: Morgan Kauffman Publisher.
[16] Ali J, Khan R, Ahmad N, and Maqsood I 2012 Random Forest and Decision Trees *International Journal of Computer Science* **9** pp 272-78.
[17] Turban E, Aronson J E, and Liang T P 2005 D*ecision Support System and Intelligent System Edisi 7 Jilid 1* Yogyakarta: Andi.
[18] Abbott D 2014 *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst* Indianapolis: Wiley.
[19] Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares M J, and Haussler D 2000 Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97** pp 262–67.
[20] Jaakkola T, Diekhans M and Haussler D 1999 Using the Fisher kernel method to detect remote protein homologies *In Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology* AAAI Press Menlo Park CA.
[21] Zien A, Ratsch G, Mika S, Scholkopf B, Lemmen C, Smola A, Lengauer T and Muller K 2000 Engineering support vector machine kernels that recognize translation initiation sites *Bioinformatics* **16** pp 799-807.
[22] Mukherjee S, Tamayo P, Mesirov J, Slonim D, Verri A and Poggio T 1999 Support vector machine classification of microarray data *Technical Report CBCL Paper* 182/AI Memo 1676 MIT.
[23] Susilowati E, Sabariah M K and Alfian A G 2015 Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter *e-Proceeding of Engineering*. **2** 1478.
[24] Siagian R Y 2011 Klasifikasi Parket Kayu Jati Menggunakan Metode Support Vector Machines (SVM) *Jurnal* Pengembangan *Teknologi Informasi dan Ilmu Komputer* **3** pp 7844-50.
[25] Rachman F and Purnami S W 2012 Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer dengan Menggunakan Regresi Logistik Ordinal dan Support Vector Machine *Jurnal Sains dan Seni ITS* **1** pp D130-35
[26] Nugroho A S, Witarto A B and Dwi H 2003 *Support Vector Machine, Teori dan Aplikasinya dalam Bioinformatika*. Diakses pada tanggal 21 Mei 2019 dari [http://www.ilmukomputer.com].
[27] Prasetyo E 2012 *Data Mining-Konsep dan Aplikasi Menggunakan MATLAB* Yogyakarta: Andi.
[28] Dunteman G H 1989 *Principal Components Analysis Quantitative Applications in the Social Sciences*. California: Sage Publications.
[29] Mattjik A A, Sumartajaya I M, Wibawa G N A and Hadi A F 2011 *Sidik Peubah Ganda Dengan Menggunakan SAS* Departemen Statistika Institut Pertanian Bogor: IPB Press.
[30] Karatzouglou A, Meyer D, and Hornik K 2006 Support Vector Machine in R *Journal of Statistic Software* **15** pp 1:28.