CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 7(1) (2021), Pages 1-12
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Genetic Algorithm for Variable Selection and Parameter Optimization in SVM and Fuzzy SVM for Colon Cancer Microarray Classification

Irhamah[1], Elok Faiqah[2], Heri Kuswanto[3], NLP Satyaning Pradnya Paramita[4]

[1,3]Department of Statistics,
Faculty of Mathematics, Computing, and Data Sciences
Institut Teknologi Sepuluh Nopember, 60111 Surabaya, Indonesia

Email: irhamah@statistika.its.ac.id, elokfaiqoh212@gmail.com,
heri_k@statistika.its.ac.id, sat.pradnyaparamita@gmail.com

ABSTRACT

Colon cancer is the second leading cause of cancer-related deaths globally; hence, research on that topic, such as disease diagnosis, needs to be undertaken with improvement. Microarray has an essential role in biomedical research as a tool for identifying and classifying diseases, especially cancer. This study aims to develop a classification model using Fuzzy Support Vector Machines (FSVM) hybridized with Genetic Algorithm (GA) for classifying individuals based on gene expression into two classes, i.e., normal and cancer, and compare classification performance based on the use of selection method. Fuzzy memberships were used in SVM to deal with the case of imbalanced microarray data. Meanwhile, the role of the genetic algorithm is, firstly, to select the relevant genes as the features and, secondly, to optimize the parameter of FSVM as GA can handle the problem of nonlinear optimization that has a high dimension, adaptable, and easily combined with other methods. The results show that classification using FCBF selection has a higher accuracy value than those without the selection. FSVM optimized using GA has the highest accuracy value compared to other classification methods used in this study.

Keywords: Feature Selection; Fuzzy SVM; Genetic Algorithm; Parameter Optimization; SVM

INTRODUCTION

Cancer is a leading cause of death globally: an estimated 7.6 million people died of cancer in 2005, and 84 million people will die in the next ten years if action is not taken. More than 70% of all cancer deaths occur in low- and middle-income countries, where cancer prevention, diagnosis, and treatment resources are limited or non-existent [1]. The cancer diagnosis can be done based on its morphological structure but has difficulties due to very thin differences in morphological structures between different types of cancer [2]. In 1999, Alon researched gene expression related to colon cancer which contained data microarray data. Microarray data is a technology in Molecular and Medical Biology that can see differences in gene expression. Microarray Data is a type of data with very high dimensions used in bioinformatics. The characteristics of data microarray are the small amount of data and a large number of features. Microarray data consists of thousands of spots (attributes), and from each spot consists of millions of

Submitted: September 18, 2020 Reviewed: May 21, 2021 Accepted: November 02, 2021
DOI: http://dx.doi.org/10.18860/ca.v7i1.10337

copies of DNA molecules that respond to a gene. The Collection of genes will be used to classify a class of diseases [3]. Therefore, the study was conducted using microarray data from colon cancer, which will be carried out by typing genes, features, or variables using the Support Vector Machine (SVM). The SVM is one of the classification methods which gives the best results with accuracy values reaching 80% to 90% and can be used to deal with high dimensional data. [4] implemented SVM on the microarray data and showed promising results.

SVM has a better performance in classification than other methods [5], but the effect of imbalanced data on SVM will be a drawback in the paradigm of maximizing margins. Some researchers propose FSVM, which applies fuzzy membership to each sample and formulates SVM so that different sample inputs have other contributions and can handle imbalanced data. In [6], SVM and NB are used to classify microarray data where the methodologies involve dimension reduction of microarray data using ICA, followed by the feature selection using FBFE. Still, there was no specific method used to deal with imbalanced data. Therefore, this study compares the classification performance of FSVM and SVM analysis on colon cancer microarray imbalanced data.

The biggest problem in setting the SVM model is in determining the hyper-parameter values of the SVM [7]. Setting the parameter values will increase the classification accuracy of the SVM model [8]. Thus, in this study, GA is used to optimize the value of parameters on the SVM model to increase the classification performance. GA can handle high-dimensional nonlinear optimization problems [9]. This study also compares the effect of Fast Correlation Based Filter (FCBF) in classification performance with variable selection. The FCBF algorithm is based on the idea that good features are a feature that is relevant to the class but not redundant to pertinent other features. The methods implemented in this research are SVM without variable selection, SVM with FCBF variable selection, SVM GA without variable selection, SVM GA with FCBF variable selection, FSVM without variable selection, FSVM with FCBF variable selection, FSVM GA without variable selection, and FSVM GA with FCBF variable selection. The best method is obtained from the highest classification performance value.

METHODS

Fast Correlation Based Filter (FCBF)

Fast Correlation Based Filter or FCBF is a variable selection method developed by [10]. In general, a feature is good if it is relevant to the class concept and not redundant to any of the other relevant characteristics. Suppose we adopt the correlation between two variables as a goodness of fit measure. In that case, the above definition means that a feature is good if it is highly correlated to the class but not correlated to any other components. In other words, if the correlation between an element and the class is high enough to make it relevant to (or predictive of) the class and the correlation between it and any other suitable features does not reach a level, then it can be predicted by any of the other relevant features, it will be regarded as a good feature for the classification task. In this sense, the problem of feature selection boils down to find a suitable measure of correlations between components and a sound procedure to select features based on this measure.

There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation, and the other is based on information theory. This study used a second approach: choosing a correlation degree based on the entropy theory information [10]. Decresing of entropy in X reflects

additional information of X by Y called Information Gain [11]. To measure information gain, symmetrical uncertainty is used. Information retrieval is symmetrical from two random variables X and Y. Symmetrical uncertainty values are carried out in the range 0 to 1 with a value of 1 which indicates another value or X and Y dependent and 0, which shows that X and Y are independent.

Support Vector Machines (SVM)

SVM was first introduced by Vapnik in 1992. SVM is a technique for finding a separation function in classifications that can separate two or more different data groups. SVM can find the best hyperplane function between unlimited functions to separate objects. The best hyperplane is located right in the middle between two object sets of two classes. Finding the best hyperplane is equivalent to maximizing the margin or distance between two sets of objects from two different classes. SVM works to find a separation function with maximum margins [12]. SVM is a classification technique with a training process (supervised learning) to find the legitimate line of the best hyperplane with *f(x),*

$$f(\mathbf{x}) = sign(g(\mathbf{x})) \tag{1}$$

where $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \mathbf{b}$;$\boldsymbol{x}, w \in R^n$ and $b \in R$

If the value of g (x) is negative then the observation goes into the negative class, so if g (x) is positive then obstruction will enter the positive class. The concept of hyperplane on SVM is as follows,

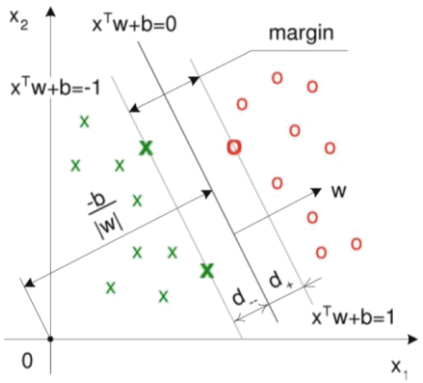

Figure 1. The concept of hyperplane in SVM

In Figure 1. the margin is equal to d- + d+. The classification function is a hyperplane plus a margin zone. This separates points from the two classes with the highest distance (margin) between the two classes. $\mathbf{x}_\mathbf{i}^T\mathbf{w} + b = 0$ is a separator hyperplane, d- + d+ will be the shortest distance on the closest object from class +1 (-1). Because separation can be solved without error, all observations i = 1, 2, ..., *n* is measured as,

$$\begin{aligned} \mathbf{x}_\mathbf{i}^T\mathbf{w} + b \geq +1 \text{ for } y_i = +1 \\ \mathbf{x}_\mathbf{i}^T\mathbf{w} + b \geq -1 \text{ for } y_i = -1 \end{aligned} \tag{2}$$

where $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0, i = 1, 2, .., n$

In general, in the real-world domain (real world problem) rarely are linearly separable, mostly nonlinear. The method for classifying data that cannot be separated from linear functions is by transforming data into feature space dimensions so that they

can be separated linearly in feature space. Input space with 2 dimensions cannot separate data into two classeslinearly. Therefore, it is necessary to map the input vector by function $\varphi(x)$ into a new vector space with a higher dimension (3 dimensions). The way to classify data is to use the transformation function $x_i \rightarrow \phi(x_i)$ into the feature space so that there is a separate field that can separate data according to its category. By using the $x_i \rightarrow \phi(x_i)$ transformation function, the value is generated,

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \phi(x_j) + b \tag{3}$$

Feature space in practice usually has a higher dimension than input vectors. This causes the computation of feature space to be very large, because there is a possibility that feature space can have an unlimited number of features. In addition, it is difficult to know the right transformation function. The transformation function in SVM is to use kernel tricks [13]. The kernel trick is to calculate the scalar product in the form of a kernel function.

$$K(\mathbf{x_i}, \mathbf{x_j}) = \phi\left(x_i^t\right)\phi\left(x_j\right) \tag{4}$$

Then the transformation function in the above equation can be used without the need to know the transformation function explicitly. Thus the resulting function is,

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K\left(\mathbf{x_i}, \mathbf{x_j}\right) + b \tag{5}$$

where $0 \le \alpha_i \le C \; ; i = 1, 2, ..., n$

The requirement for a function to function as a kernel is to fulfill the Mercer theorem, which is a necessary and sufficient condition for symmetric functions $K(\mathbf{x_i}, \mathbf{x_j})$. The most commonly used kernel functions are as follows,

a. *Kernel Gaussian (RBF)* : $K\left(\mathbf{x_i}, \mathbf{x_j}\right) = \exp\left(-\gamma\left(\left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2\right)\right)$

b. *Kernel Linear* : $K\left(\mathbf{x_i}, \mathbf{x}_j\right) = \mathbf{x_i}^T \mathbf{x_j}$

c. *Kernel Polynomial* : $K\left(\mathbf{x_i}, \mathbf{x_j}\right) = \left(\delta \mathbf{x_i^t x_j} + r\right)^p$

Fuzzy Support Vector Machines

Fuzzy Support Vector Machine (FSVM) is a development of Support Vector Machine for multiclass problems. By using the decision function obtained from SVM for a class pair, for each class a polyhedral pyramidal membership function is defined. FSVM uses membership functions to classify variable that cannot be classified by decision functions [14]. There are several applications that only want to focus on accuracy for class classification. For this purpose fuzzy membership can be determined as a function of each class. Suppose given a series of training [15]:

$$(y_1, x_1, s_1), ..., (y_n, x_n, s_n)$$

Fuzzy membership becomes a function in the class $s_i = 1$ if $y_i = 1$ and $s_i = 0.1$ if $y_i = -1$ by using lagrangian, the decision function for FSVM is stated as follows

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \mathbf{K(x_i, x_j)} + b \tag{6}$$

when $0 \le \alpha_i \le s_i C \; ; i = 1, 2, ..., n$

Pre-Processing Data

Pre-processing datais a process to improve the quality of raw data, so that it can improve accuracy and efficiency for the next data mining process. In this study, pre-processing of data was carried out using transformations. One method of transformation is scaling where one of the advantages of scaling is that it can avoid features with a greater range of values dominating features with a smaller range of values. Each data is linearly transformed into a range [0, 1] using the following equation,

$$v' = \frac{v - \min_a}{\max_a - \min_a} \tag{7}$$

where$v'$is the transformed value, v is the initial value, max_a is maximum value on variable, and min_a is minimum value on the variable.

Genetic Algorithm

Genetic algorithm (GA) was first discovered by John Holand in 1975. The GA concept is based on the theory of evolution with the principle of natural selection developed by Darwin. GA is a technique for identifying solutions to optimization problems. The steps taken in the GA method are as follows:

Step 1 : Define, which defines the operator in GA that corresponds to the problem. At this research, the selection and optimization variables were carried out using genetic algorithm

Step 2 : Initialize, which is to form an initial population consisting of N chromosomes. Set N = 100

Step 3 : Fitness, which evaluates the fitness of each chromosome in the population

Step 4 : Selection, which applies the roulette wheel selection method which gives a set of M mating populations with size N

Step 5 : Crossover, which is the recombine process. This process randomly pairs all M chromosomes to form N / 2 pairs. If a random number [0.1] is less than Pc, then crossovers occur

Step 6 : Mutation, which is using mutation probability (Pm) to carry out the process of inheritance mutation

Step 7 : Replace, which is replacing the old population with the new population. The new population is obtained by selecting the best N chromosomes obtained by evaluating the fitness value of parents and new breeds

Step 8: Test, that is, if the criteria has been met, then the process is stopped and return to the best solution of the current population. If the criteria have not been met, then go back to step 2. Furthermore, elitism is one of the techniques that is done to maintain the best individual who has the highest fitness value to survive for the next generation

Microarray Data

Microarray data is one of the technologies used to measure the level of expression of thousands of genes simultaneously in one observation and appears as a device of the microarray which is usually summarized in the list of genes and expressed in two conditions or classifications based on the phenotype. Microarray data is a type of high dimensional data because it has a number of genes (variables) hundreds or even thousands, while the number of observations that usually do not reach 100 or far smaller than the number of variables. Two common methods for analyzing data microarray are clustering and classification [16]. Based on the information they possess,

microarray has an important role in biomedical research as a tool for identifying and classifying diseases, especially cancer.

The colon cancer data stores information in the form of gene expressions obtained from patient of tumor colon tissues and normal colon tissues. Colon data consists of 62 patients, 40 of which are patients of the tumor class and the other 22 are normal classpatients. The number of variables in colon cancer data is 2000 variables. The steps of microarray experiment to get a colon cancer microarray data are as follows [17].

1. Obtaining mRNA from the observed cell (for example in the case of a tumor, the sample observed is a cell that has a tumor.
2. mRNA is converted to cDNA using reverse tranciptase enzyme.
3. Marking cDMA from tumor cells in red and cDNA from normal cells in green.
4. The sample is hybridized, iecDNA binds to DNA.
5. The sample is scanned to measure the expression of each gene through fluorescence contained (fluorescence is related to the amount of cDNA in the sample for the gene).
6. A bright red glowing point is a gene that is highly expressed in tumor cells, while a bright green glow is a gene that is highly expressed in normal cells. If the gene is expressed in both samples (tumors and normal), the color produced is bright yellow.

From the process, the final data is obtained which consists of thousands of points that have different colors and need to be interpreted. Color dots must be changed to a certain value to be analyzed later. Here is an illustration of the image to get microarray data,

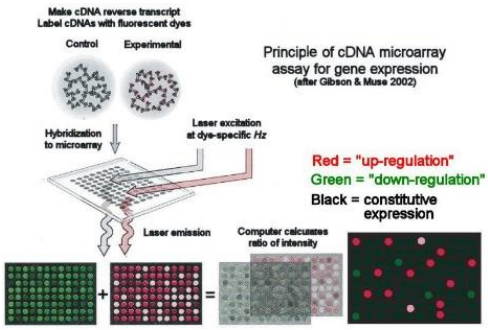

Figure 2. Microarray data process (source: Gibson & Muse, 2002)

RESULTS AND DISCUSSION

Description of Data

The data used in this study is the microarray type colon cancer data by U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine in 1999 [18]. Data were taken from human intestinal tissue. The expression of these genes is stored in 2000 variables. This data consists of 62 patients, where 40 patients are having tumor colon tissue (tumor/tissue) while 22 are normal patients (normal colon tissue). The research variable used are shown in Table 1,

Table 1. Research Variables

Variable	Information	Measurement Scale
Dependent (Y)	*Colon cancer class:* 0 = normal 1 = tumor	Nominal
Independent (X)	Gene expression	Ratio

Characteristics of Colon Cancer Data

The data used are observations made on 62 samples taken from human intestinal tissue. The data is divided into two classes, namely the tumor class and normal class. The tumor class is described as negative by coding 1 and the normal class is described as positive by coding 0. The description of the two colon cancer classes is shown in Figure 3.

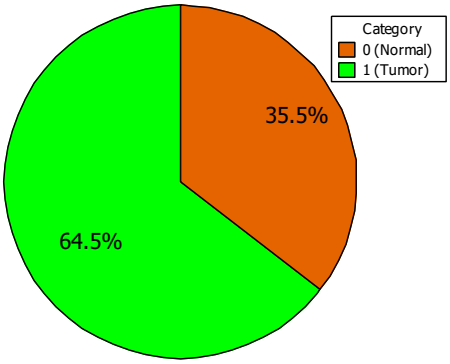

Figure 3. Percentage of Normal Class and Tumor in Colon Cancer Data

In this observation, a positive sample is stated as a normal sample with code 0 and a negative class is stated as a tumor sample with code 1. Figure 3. shows that of the 62 patients taken there were 64.5% of the total samples stated as tumors and the remaining 35.5% were declared normal . Based on the proportion of samples of normal and tumor class classes, it is known that colon cancer data is imbalanced data so that the appropriate method to classify colon cancer class is the Fuzzy Support Vector Machine (FSVM) method. Colon cancer data has a number of genes or variables as many as 2000 variables so that if it will complicate the classification because the data distribution pattern will become very complex. Following is the pattern of data distribution shown in Figure 4.

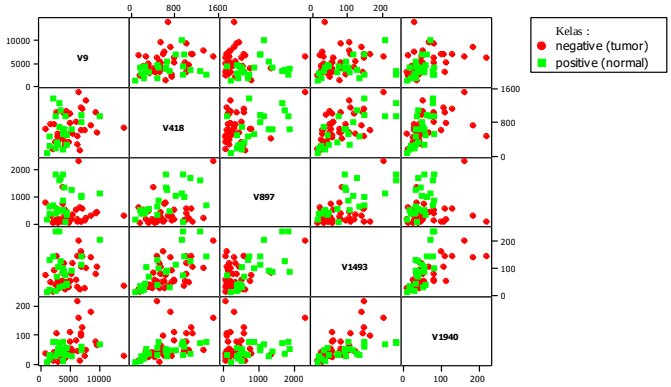

Figure 4. Distribution patterns of several variables in colon cancer data

The pattern of colon cancer data distribution shown in Figure 4. in red is the tumor class and the green color is normal class is spread evenly so that it is difficult in classification. The function separator or hyperplane is expected to be able to help overcome the classification problems in colon cancer data so that in this study an analysis will be conducted using fuzzy support vector machine (FSVM) classification method.

Pre-Processing Data

Data pre-processing is a process carried out to improve the quality of raw data by producing precise accuracy values and efficient classification results. In this study, pre-processing is done using variable transformation and selection. One method of transformation is scaling where one of the advantages of scaling is that it can avoid features with a greater range of values dominating features with a smaller range of values.

Table 2. Results of transformation of colon cancer data scaling

Variable	Variable Name	Before Transformation		After Transformation	
		Mean	Varians	Mean	Varians
X_1	H55933	7016	9566467	0.3936	0.0569
X_2	R39465	4967	4791241	0.4087	0.0623
X_3	R39465_1	4095	3305418	0.3851	0.0614
X_4	R85482	3988	4076712	0.2784	0.0403
X_5	U14973	2937	1841267	0.2556	0.0384
...	...	...	...	...	...
...	...	...	...	...	...
X_{1999}	R77780	53.25	1479.39	0.2477	0.0404
X_{2000}	T49647	42.97	806.28	0.3070	0.0551

Table 2 shows that after transformation using scaling, the average value of each variable becomes smaller and ranges between [0,1]. The variance value of the transformed variable is also small which indicates that the observation value with small data distribution or diversity between observations of one with other observations is quite small.In this study, the selection of variables used is FCBF (Fast Correlation Baser Filter) and the variables obtained after selecting variables with FCBF are 15 variables from 2000 variables. The 15 variables obtained were further analyzed using SVM and FSVM classifications.

Colon Cancer Classification Performance

The SVM classification consists of SVM classifications with or without FCBF variable selection. The results of classification methods performance are shown in Table 3. In SVM without FCBF selection, the optimum valueof cost and gamma parameter are 128 and 0.001953125. Based on the AUC, the SVM without selection was good because the AUC value was 86.42%. The accuracy value obtained is also quite high, that is 83.57%. Whereas for SVM classification with FCBF variable selection, the optimal value of cost and gamma parameter are 2048 and 0.03125. The AUC value obtained in the SVM classification with FCBF selection is very good because it is located in a range of 90-100%. It means that the SVM model with FCBF selection has been very good in classifying colon cancer data. the accuracy obtained is 91.90% which means that by using a cost value 2048and gamma 0.03125, SVM models can classify 91.90% of observations correctly. In addition, the model can also classify positive or normal classes correctly at 88.33% which can be seen from the sensitivity value and 95.50% of the model can classify negative classes or tumors with a specificity value.

Table 3. Performance of Colon Cancer Classification

Method	Variable Selection	Performance Classification (%)					Optimal Parameters	
		Accuracy	Sensitivity	Specificity	G-Means	AUC	Cost	Gamma
SVM	without variable selection	83.57	83.33	89.5	85.31	86.42	128	0.001953125
	with FCBF variable selection	91.9	88.33	95.5	91.92	91.4	2048	0.03125
SVM GA	without variable selection	83.57	83.33	89.5	85.31	86.42	1272.99	0.0011941
	with FCBF variable selection	95	91.67	97.5	94.29	94.58	18652	0.0161011
FSVM	without variable selection	75.71	40	95	47.13	67.5	32768	0.0000610
	with FCBF variable selection	90.24	81.67	95	87.09	88.33	2048	0.125
FSVM GA	without variable selection	87.38	-	-	-	86.25	43011.2	0.000088722
	with FCBF variable selection	90.03	-	-	-	97.5	2968.8	0.15275

On SVM-GA without selection, the optimal cost parameter is 1272.99 and gamma optimal parameters is 0.0011941 with the highest fitness value of 94.64%.The model is also classified as good in classifying based on AUC value in the range of 80 to 90%. The cost and gamma optimal parameters from SVM-GA-FCBF is 18652.00 and 0.0161011. The model can classify the tumor class and normal correctly by 95% which can be known from the constellation value. All classification performance values which include sensitivity, specificity, g-means, and AUC values have values above 90%.

Before conducting FSVM classification, firstlywe have to determine the optimal value of cost and gamma parameter from training data as has been done in SVM analysis from the optimal parameter range. The determination of optimal parameters is based on AUC value and accuracy. This study uses 10 fold and the highest average for the AUC value and the accuracy value in the training data was obtained from the cost value 32768 and gammavalue 0.000061035 with the accuracy value obtained is 98.74% and the AUC value is 98.22%. These parameters will be usied in FSVM analysis using data testing. The following classification performance is obtained using the optimal parameter values in training data that is cost value 32768 and gamma value 0.000061035.According to Gorunescu 2011, the AUC value in the range 60-70 shows a poor classification. This is also supported by the Gmeans value of 47.13% which means that the stability between the performance of the classification of minority classes and the majority class is 47.13%. Meanwhile for FSVM with FCBF selection, the cost and gamma optimal values are 2048 and 0.125 with an average value of 98.39% accuracy and an average AUC value of 97.89%. Optimal parameters will be used to find performance testing data classification. Classification performance on testing data will be calculated using the optimal parameter value in training data: cost 2048 and gamma 0.125. The AUC value obtained is 88.33%, which means that the model is good in classifying. The model can also classify positive or normal classes and negative or tumor classes correctly by sensiticity value is 81.67% and specificity value is 95%.

In the classification using FSVM without selection and with GA optimization, optimization will be performed on each fold with a range of costs 2^{15}-2^{16} and the range of gamma is 2^{-3}-2^{-2}. The average fitness value value generated for the range cost 2^{15}-2^{16}

and the range gamma is 2^{-3} - 2^{-2} in the training data is 98.22%. Average Fitness Value is used to get the highest AUC value from the K-fold Cross Validation (KCV) method which is a reliable method for predicting errors in a classification. This method is usually used by researchers to reduce the bias that occurs because of sampling data to be used. The average fitness value value for testing data is 86.25%. There are 7 folds that have a fitness value of up to 100% then the cost and gamma optimal parameters of GA optimization results can be taken from one of the seven cost values and gamma values which has a fitness value of 100%.

The cost and gamma optimal parameters obtained from FSVM FCBF selection will be optimized using GA optimization. The range used in the FSVM GA optimization FCBF selection is 2^{11}-2^{12} for the cost parameter and 2^{-3}-2^{-2} for the gamma parameters. By using this range of parameters, the fitness value value generated in the training data reaches 98.35% so that the value range 2^{11}- 2^{12} for the cost parameter and 2^{-3}-2^{-2} for gamma parameters can be used to find the optimal para-meter using data testing. The average fitness value value generated using a range of cost 2^{11}-2^{12} and the range of gamma 2^{-3}-2^{-2} is 97.5%. The value of fitness-value obtained to fold 1 to fold 9 is 100% so that for cost and gamma optimal can be taken from one of the folds that have a fitness value of 100%.

The results of SVM grid search method and GA for SVM optimization without selection have the same accuracy and AUC values. This can occur because the variables have not been selected. When compared to SVM grid search and SVM-GA for optimization, it can be seen that with GA optimization can increase the accuracy and AUC value. GA for SVM optimization has an AUC value of 94.58% and an accuracy value of 95%. Based on AUC value, the SVM-GA optimization classification performance is very good.Classification method using FCBF selection has a value higher accuracy than without selection, both for for SVM method or FSVM according to a comparison table of classification in Table 3. In addition, it is known that the best method that can be used to classify the class of colon cancer is FSVM using FCBF with GA optimization, since it yields the highest AUC compared to other methods (that is 97.50%). AUC is considered first than Accuracy because this study deals with imbalance data.

CONCLUSIONS

Classification analysis has been carried out using SVM and FSVM on colon cancer data thatconsist of 35.5% normal class and 64.5% tumor class. After selecting variables using FCBF, from about 2000 variables are then reduced to 15 variables. The results of SVM classification with variable selection give higher accuracy values than the SVM without variable selection. In addition, SVM GA optimization with variable selection also produces better results than SVM without selection with GA optimization. FSVM classification method also produces the same results, where using variable selection produces higher accuracy values than without variable selection. Among the four methods used, GA-FSVM classification method produces the highest fitness value compared to other classification methods investigated by this study. The fitness value for FCBF selection is 97.50% and the one without selection is 86.25%.

ACKNOWLEDGMENTS

The authors thank to the Ministry of Research, Technology, and Higher Education, Republic of Indonesia and Institut Teknologi Sepuluh Nopember Surabaya Indonesia for the financial support under "Penelitian Dasar Unggulan Perguruan Tinggi"

REFERENCES

[1] WHO, World Cancer Day : Global Action to Avert 8 Million Cancer-Realted Deaths By 2015. Retrieved from https://www.who.int/mediacentre/news/releases/2006/pr06/en/, 2002

[2] T. R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, & J.P. Mesirov, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science,* 286, 531-537, 1999

[3] M. M. Babu, *Introduction To Micoarray Data Analysis*, U.K : Horizon Press, 2013

[4] T.S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, & D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data". *Bioinformatics*, Vol. 16, No. 6 , 906-914, 2000

[5] Y.-N. Chen, C.-A. Lu, and C.-Y. Huang, *Anti Spam Filter Based on Naive Bayes, SVM and KNN Model*. Sillicon Valley: Carnegie Mellon School, 2009.

[6] R. Aziz, C. K. Verma, & N. Srivastava, "A Fuzzy Based Feature Selection from Independent Component Subspace for Machine Learning Classification of Microarray Data", *Genomics Data*, Vol. 8, 4-15, 2016

[7] S. Yenaeng, S. Saelee, & W. Samai, "Automatic Medical Case Study Essay Scoring by Support Vector Machine and Genetic Algorithm", *International Journal of Information and Education Technology*, Vol. 4, No. 2, 132-137, 2014

[8] C. L. Huang & C. J. Wang, "A GA-based Feature Selection and Parameters Optimization for Support Vector Machines", *Expert Systems with Application*, Vol. 31 , 231- 240, 2006

[9] H. Roubos & M. Setnes, *Compact Fuzzy Models and Classifiers through Model Reduction and Evolutionary Optimization*. In L. Chambers, The Practical Handbook of Genetic Algorithms, 2001

[10] L. Yu, and H. Liu, Feature Selection for High Dimentional Data : A Fast Correlation-Based Filter Solution, *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*. Washington DC, 2003

[11] J. Quinlan, *C4.5: Programs for machine learning*. Morgan Kaufmann, 1993

[12] S. Abe and T. Inoue, *Fuzzy Support Vector Machines for Multiclass Problems,* Jepang : Kobe University, 2002

[13] V. N. Vapnik, *The Nature of Statictical Learning Theory 2nd Edition*, Springer-Verlag: New York Berlin Heidelberg, 1999

[14] B. Scholkopf and A. Smola, *Learning with Kernel: Support Vector Machines, Regulerization, Optimization, and Beyond*. Cambridge: MA: MIT Press, 2002

[15] C. Lin and S. Wang, "Fuzzy Support Vector Machines", *IEEE Trans. Neural Network* , 464-471, 2002

[16] S. Selvaraj and J. Natarajan, "Microarray Data Analysis and Mining Tools", *Bioinformation*, 6(3), 95-99, 2011

[17] A. P. Kusumaningrum, *Optimasi Parameter Supprort Vector Machine*

Menggunakan Genetic Algorithm Untuk Klasifikasi Microarray Data. Surabaya : DepartemenStatistika FMKSD ITS, 2018

[18] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad Patterns Of Gene Expression Revealed By Clustering Analysis Of Tumor And Normal Colon Tissues Probed By Oligonucleotide Arrays", *ProcNatlAadSci* U S A. Jun 8;96(12):6745-6750, 1999