Optimization Features Using GA-SVM Approach

Andy[1], Michael Fernando[2], Kristanto Halim[3], Gradiyanto Sanjaya[4]

[1, 2, 3, 4]Bina Nusantara University, Master of Information Technology Student, Kebon Jeruk Raya No. 27, Jakarta Barat, Indonesia

Abstract: ***Feature selection often used to choose the feature that maximizes the prediction of classification accuracy. Feature selection is one of the most important factor that influence classification accuracy rate. In this paper we proposed the combination of Genetic Algorithm (GA) and Support Vector Machine for feature optimization. In this research we compare the result with K Nearest Neighbor, Decision Tree, and Linear Discriminant Analysis. For better comparison, the experiment was conducted using 6 different dataset. The result shows that GA-SVM gives better accuracy than using all features or other method on 3 of 6 dataset.***

Keywords*:* Feature Optimization, Genetic Algorithms (GAs), Support Vector Machine (SVM);

1. Introduction

In machine learning, there are cases where many features used are irrelevant or redundant. This unnecessary feature not only increase the search space size but make generalization more difficult and the process to capture rules that specify the classification more difficult [1].

Feature selection often used to choose the feature that maximizes the prediction of classification accuracy [1]. Features selection is one of the most important factor that influence classification accuracy rate. Feature selection can eliminate noisy, irrelevant, and redundant data in the dataset used for classification [2]. Feature selection approach prefers a model with the smallest possible number of parameter/features that adequately represents the data. Because feature selection need an exhaustive search of all possible subset of features to guarantee the best subset of features can be found, the computational cost to produce the best subset is usually large. The trade-off of features selection is the computational time in exchange of the best subset features that can be generated from the selection [1].

Feature selection has 2 broad categories that have been proposed: filter and wrapper. Filter categories is based on statistical criteria. Some of the methods used in filter categories are t-test, chi-square test, or principal component analysis. In wrapper, feature selection is generated from a learning algorithm that search for an optimal subset of features. The forefront of research in wrapper features selection nowadays is in stochastic algorithms such as ant colony optimization (ACO), genetic algorithm (GA), particle swarm optimization (PSO), and simulated annealing (SA) [1].

Recent researches have proposed a hybrid approach that combines filter and wrapper method. Some of hybrid techniques include t-statistics and a GA, a correlation-based feature selection algorithm and a genetic algorithm, and mutual information and a GA [1].

GA is a stochastic algorithm with global search heuristic that mimics natural evolution. GA able to quickly scan a vast population to find the best possible solution available and often work well with highly constrained problem [3]. Besides that, GA was shown to be very efficient for optimum solution search in a great variety of problems and can avoid problem in traditional optimization algorithm such as returning the local minimum [4].

As mention in above, wrapper uses a learning algorithm to assess the accuracy of all the potential subset available. Currently, the most popular learning algorithm used in wrapper is support vector machine (SVM) [1]. SVM is a learning method based on statistical learning theory. SVM can solve over fitting problem, local optimal solution and low-convergence rate that exist in ANN and the difference in risk minimization leads to better generalization performance for SVM than ANN [5].

In this research, we proposed a hybrid method of GA and SVM. We believe GA as one the wrapper method that were shown to very efficient combined with SVM advantages can produce a feature selection techniques that can generated a optimized subset
of features.

2. Materials and Methods

A. Dataset

In this research more than one dataset is used to prove that our method can be used for general usage. This research uses Parkinson Dataset [6], Breast Cancer Wisconsin (Diagnostic) Dataset [7], Ionosphere Dataset [8], Climate Model Simulation Crashes Dataset [9], SPECTF Heart Dataset [10], and Cylinder Bands Dataset [11] from UC Irvine Machine Learning Repository. Parkinson dataset, breast cancer dataset, ionosphere dataset, climate model simulation crashes dataset, SPECTF heart dataset, and cylinder bands dataset respectively contain 195 data with 22 features, 569 data with 30 features, 351 data with 34 features, 540 data with 18 features, and 512 data with 38 features. For cylinder bands dataset there is 512 data with 234 missing value, after missing value is removed the remaining 278 data is used. For every dataset 50% from total data was taken for training and the rest 50% was taken for testing purpose. The detailed composition for each dataset is shown in Table I.

Table 1: Dataset Detail

Dataset Details				
Dataset Name	Total Data	Missing Value	Used Data	Features
Parkinson	195	-	195	22
Breast Cancer	569	-	569	30
Ionosphere	351	-	351	34
Climate Model	540	-	540	18
SPECTF Heart	267	-	267	44
Cylinder Bands	512	234	278	38

B. Methods

In this research the methods consists of a feature optimization using GA with Support Vector Machine (SVM) to determine the fitness function, and classification step.

1) Feature Optimization

Feature optimization is the primary topic in this research. In this research, Genetic algorithm (GA) is used to optimize the features from dataset. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), or Linear Discriminant Analysis (LDA) were used to find fitness function for GA. After applying GA selected feature should be the best feature that could produce the best accuracy.

a) Genetic Algorithm

According to Pengfei Guo, Xuezhi Wang, and Yingshi Han [12] genetic algorithm (GA) is a powerful stochastic algorithm, the initial idea of GA is the application of natural selection and natural genetics in machine learning and optimization problems. To solve a problem GA maintains a population called strings or chromosomes and GA modifies the population by some genetic operators to seek a near optimal solution to the problem. An example of genetic operators is selection, crossover, and mutation [3].

- Selection: According to previous research [13] a selection scheme is applied for every individual to determine how individual are chosen for mating based on their fitness value. Fitness itself can be defined as capability of an individual to survive and reproduce in environment. After selection is applied, it generates a new population from the old one thus starting a new generation. Each chromosome is evaluated in this generation to determine its fitness value. This fitness value is used to determine which chromosomes to be used from the population for the next generation.
- Crossover: After selection, crossover operation is applied to the selected chromosomes. Crossover involves in swapping of genes or sequences between two individuals. Crossover operation is repeated with different parent individuals until next generation has enough individuals [13].
- Mutation: After crossover, mutation operator is applied to some randomly selected subset of the population. Mutation alters chromosomes and introduces new traits. Mutation is applied to bring diversity in the population [13].

b) Fitness Function

According to Engelbrect [14] fitness function is the ability of an individual of an evolutionary algorithm to survive. To calculate fitness function. The formula to calculate fitness function can be seen below.

$$w = error(M)$$

$$FV = w/(N - n1)$$

w is the classification error , M is a classification model produced by machine learning classification algorithm, FV is the Fitness Value, N is the number features in the data, and n1 is the number of features selected that used in the classification.

The process step to find the fitness value in general can be described below:

1. A population consist of one row and all features with 0 and 1 value is passed to the fitness function
2. Find (FI) which is the features has value 1 in the population
3. Get X1 which is all row of the data used but only with the features in F1
4. Calculate n1 which is the number of features selected in F1
5. Calculate M by classify X1 using selected classification algorithm
6. Calculate the classification error(w)
7. Calculate the fitness value (FV)

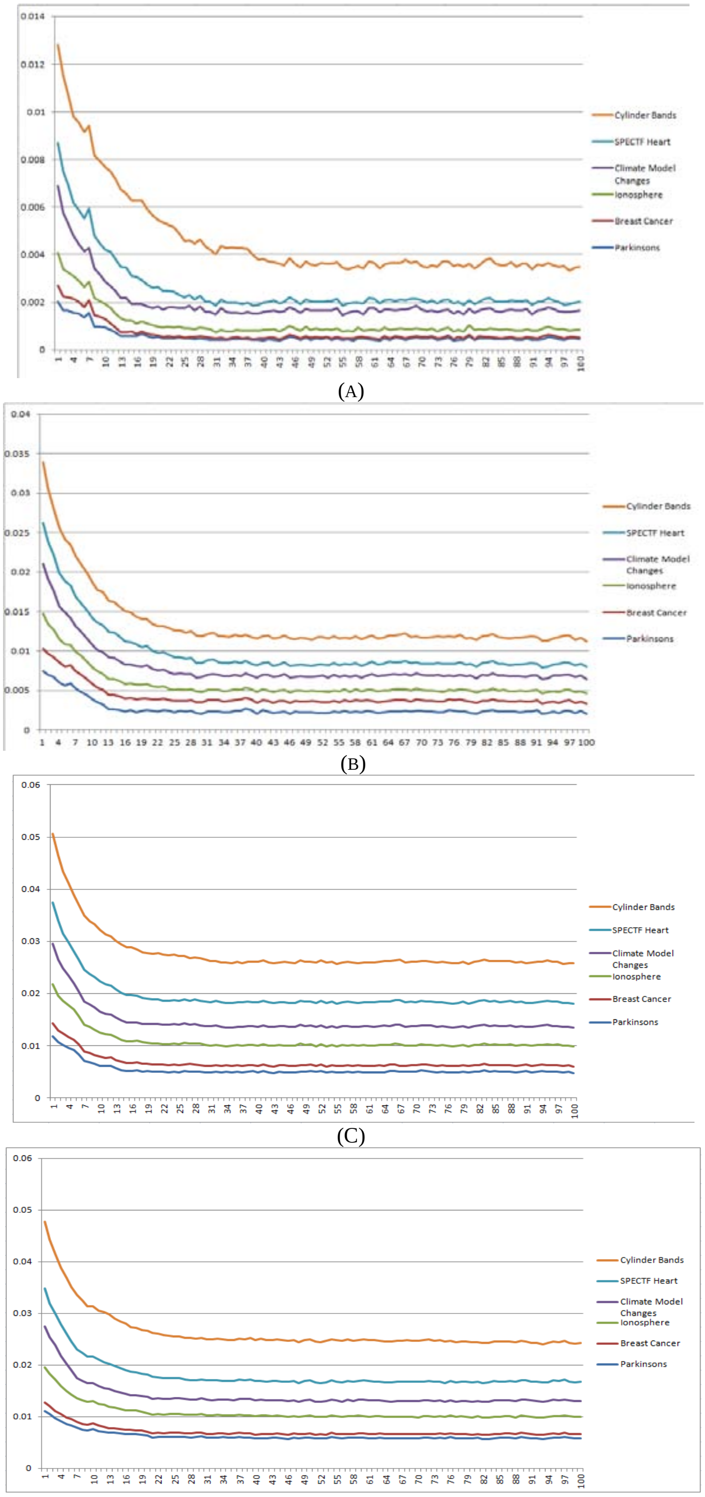
2) Classification

In this research Adaptive Neuron- Fuzzy Inference System (ANFIS) is used as classifier. ANFIS combine properties of fuzzy logic (membership function and fuzzy rules) and neural network to produce a system that are expected to be able to interpreting the relationship between extracted features [12]. ANFIS model maps input through input Membership Functions (MF) and then maps MF outputs to outputs. The membership function and fuzzy rules in fuzzy inferences system can be set by human that have expertise about the targeted system model. The fuzzy rules and membership function will be used by ANFIS to describe input-output behavior of the system [13].

The classification result from ANFIS is used to determine whether the features selected in features selection optimize or not. The features selected from features selection can be said to have optimized if the features selected are smaller than initial number of features and the accuracy is higher.

3. Results & Discussion

In this research we observe the usage of GA-SVM as feature optimization with ANFIS as classifier. Each of the dataset obtained from UC Irvine Machine Learning Repository is divided into two parts. Each part contains 50% of the dataset, the first part is training data and the second part is testing data. This test was conduct on ANFIS classifier with 100 epochs and the best accuracy was achieved after 500 loops. For comparison, we use GA combined with K-Nearest Neighbor (KNN), Decision Tree (DT), or Linear Discriminant Analysis (LDA). The detailed of the results is show in Table II until Table VII, while the fitness function for each dataset can be seen in Fig 1.



(A)

(B)

(C)

(D)

Figure 1: Fitness function for each dataset (a)Decision Tree (b) K Nearest Neighbor (c) Linear Discriminant Analysis (d) Support Vector Machin

Table 1: Performance Evaluation Parkinson Dataset

Performance Evaluation from Parkinson Dataset				
Method	*Training Accuracy*	*Testing Accuracy*	*MSE*	*Features*
GA-KNN	92.86%	84.54%	9.58E-02	3
GA-SVM	94.90%	94.85%	9.88E-02	6
GA-DT	97.96%	89.69%	4.70E-02	5
GA-LDA	96.94%	93.81%	7.33E-02	6
Non-GA	91.84%	89.69%	1.77E-01	22

Table 2: Performance Evaluation Breast Cancer Dataset

Performance Evaluation from Breast Cancer Dataset				
Method	*Training Accuracy*	*Testing Accuracy*	*MSE*	*Features*
GA-KNN	98.60%	93.66%	3.34E-02	4
GA-SVM	99.30%	95.07%	1.45E-02	9
GA-DT	98.25%	94.37%	3.83E-02	10
GA-LDA	96.84%	94.01%	3.08E-02	11
Non-GA	98.60%	95.42%	NaN	30

Table 3: Performance Evaluation Ionosphere Dataset

Performance Evaluation from Ionosphere Dataset				
Method	*Training Accuracy*	*Testing Accuracy*	*MSE*	*Features*
GA-KNN	96.02%	94.29%	6.88E-02	5
GA-SVM	97.73%	94.86%	4.80E-02	8
GA-DT	96.59%	94.29%	6.89E-02	8
GA-LDA	94.89%	92.57%	1.03E-01	4
Non-GA	44.32%	27.43%	NaN	34

Table 4: Performance Evaluation SPECTF Heart Dataset

Performance Evaluation from SPECTF Heart Dataset				
Method	*Training Accuracy*	*Testing Accuracy*	*MSE*	*Features*
GA-KNN	100.00%	85.71%	3.56E-03	14
GA-SVM	97.76%	86.47%	5.11E-02	10
GA-DT	98.51%	84.21%	3.52E-02	11
GA-LDA	97.76%	84.96%	5.12E-02	7
Non-GA	98.51%	82.71%	3.12E-02	44

Table 5: Performance Evaluation Climate Model Simulation

Performance Evaluation from Climate Model Simulation				
Method	*Training Accuracy*	*Testing Accuracy*	*MSE*	*Features*
GA-KNN	99.63%	94.81%	9.25E-03	5
GA-SVM	100.00%	94.07%	8.24E-04	5
GA-DT	100.00%	94.81%	4.49E-03	9
GA-LDA	100.00%	94.07%	8.24E-04	5
Non-GA	100.00%	96.30%	1.44E-09	18

Table 6: Performance Evaluation Cylinder Bands Dataset

Performance Evaluation from Cylinder Bands Dataset				
Method	*Training Accuracy*	*Testing Accuracy*	*MSE*	*Features*
GA-KNN	75.54%	68.35%	3.68E-01	5
GA-SVM	77.70%	71.22%	3.32E-01	3
GA-DT	84.89%	70.50%	2.56E-01	9
GA-LDA	82.73%	73.38%	2.50E-01	7
Non-GA	64.03%	64.03%	NaN	38

From the accuracy shown on table II to table VII can be seen that SVM has the best accuracy on 3 dataset which is Parkinson dataset, ionosphere dataset, and SPECTF heart dataset. LDA has best accuracy on 1 dataset which is Cylinder bands dataset. Even though GA-SVM produce highest accuracy in the 3 of 6 dataset used, other method produce higher accuracy than initial accuracy in 4 of 6 dataset.

The other 2 dataset have highest accuracy when using all features available. The accuracy when using all features on the 2 dataset exceeds 95%. From this result, we can conclude that applying feature selection when the accuracy of using all Features already past 95% is not really necessary. This because we can conclude that all features that produce such result were already optimized.

Besides the accuracy, from the result can be seen that smaller number of features selected from a method is not always guarantee the result is the best among the other method used.

4. Conclusion

From the research, we can see that feature selection using GA-SVM approach produce best result on most dataset among other GA combination approach. This means features used in classification can be optimized using GA-SVM approach. There was some exception where the accuracy from the feature selection result produces lower accuracy than using all features available. But there was some similarity in the 2 dataset where the accuracy of all features are above 95%. This is because the features in dataset that have accuracy above 95% are already optimized and feature selection is not necessary. In this research, all dataset have 2 classes which is binary classification, for future research the feature selection can be applied in multi-class classification to determine whether features optimizing can produce better result in multi-class classification.

5. Acknowledgment

We would like to thank Dr. Ir. Diaz D. Santika, M.Sc from Bina Nusantara University for his guidance and UC Irvine Machine Learning Repository for providing the dataset.

References

[1] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," Pattern Recognition, 2009.

[2] Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu and S. Wang, "An Improved Particle Swarm Optimization for Feature Selection," Journal of Bionic Engineering, vol. 8, no. 2, 2011.

[3] J. Guo, J. White, G. Wang, J. Li and Y. Wang, "A Genetic Algorithm for Optimized Feature Selection with Resource Constraints in Software Product Lines," *Journal of Systems and Software*, 2011.

[4] A. L. Oliveira, P. L. Braga, R. M. Lima and M. L. Cornélio, "GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation," *Information and Software Technology*, vol. 52, pp. 1155-1166, 2010.

[5] H. Li and Y. Xin Zhang, "An Algorithm of Soft Fault Diagnosis for Analog Circuit Based on The Optimized SVM by GA," *IEEE*, pp. 1023-1027, 2009.

[6] M. A. Little, P. E. McSharry, E. J. Hunter and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE*, 2008.

[7] W. H. Wolberg, W. N. Street and O. L. Mangasarian, "UCI Machine Learning Repository," [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

[8] V. Sigillito and S. P. Group, "UCI Machine Learning Repository," Space Physics Group , [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Ionosphere.

[9] D. Lucas, R. Klein, J. Tannahill, D. Ivanova, S. Brandon, D. Domyancic and Y. Zhang, "UCI Machine Learning Repository," Lawrence Livermore National Laboratory, [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Climate+Model+Simulation+Crashes.

[10] K. J. Cios, L. A. Kurgan and L. S. Goodenday, "UCI Machine Learning Repository," [Online]. Available: http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart.

[11] B. Evans, "UCI Machine Learning Repository," [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Cylinder+Bands.

[12] P. Guo, X. Wang and Y. Han, "The Enhanced Genetic Algorithms for the Optimization Design," *IEEE,* pp. 2990-2994, 2010.

[13] C. Sharma, S. Sabharwal and R. Sibal, "A Survey on Software Testing Techniques using Genetic Algorithm," *IJCSI,* vol. 10, no. 1, pp. 381-393, 2013.

[14] A. P. Engelbrect, Computational Intelligence An Introduction, USA: John Wiley & Sons, 2007.

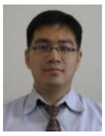
[15] Z. Chunhong and J. Licheng, "Automatic parameters Selection for SVM based on GA," *IEEE,* pp. 1869-1872, 2004.

Author Profile

Andy was born in Medan, 1991. His major is Computer Science and interest in game application and Artificial Intelligent major. Currently study in Bina Nusantara University for master degree.

Michael Fernando was born in Jakarta, 1992. He studied his bachelor degree in Bina Nusantara University with computer science major and currently study for master degree in Bina Nusantara University.

Kristanto Halim was born in Jakarta, 1992. He studied his bachelor degree in Bina Nusantara University with Computer Science Major and currently study for his master degree in Information Technology in Bina Nusantara University.



Gradiyanto Sanjaya was born in Jakarta, 1992. He received bachelor degree in Computer Science from Bina Nusantara University in 2014 and currently study at same university for his master degree in Computer Science. His interest are game applications and Artificial Intelligent.