한국정보전자통신기술학회논문지(jkiiect)'15-10, Vol.8 No.5
논문 15-08-05-384 http://dx.doi.org/10.17661/jkiiect.2015.8.5.384

A Design of an Optimized Classifier based on Feature Elimination for Gene Selection

Byung-Kwan Lee, Seok-Gyu Park*, Yusrina-Tifani

유전자 선택을 위해 속성 삭제에 기반을 둔 최적화된 분류기 설계

이병관, 박석규*, 유슬리나 티파니

Abstract This paper proposes an optimized classifier based on feature elimination (OCFE) for gene selection with combining two feature elimination methods, ReliefF and SVM-RFE. ReliefF algorithm is filter feature selection which rank the data by the importance of the data. SVM-RFE algorithm is a wrapper feature selection which wrapped the data and rank the data based on the weight of feature. With combining these two methods we get less error rate average, 0.3016138 for OCFE and 0.3096779 for SVM-RFE. The proposed method also get better accuracy with 70% for OCFE and 69% for SVM-RFE.

요 약 본 논문은 두 가지 속성 삭제 방법인 ReliefF와 SVM-REF를 조합하여 유전자 선택을 위한 속성 삭제에 기반을 둔 최적화된 분류법(OCFE)을 제안한다. ReliefF 알고리즘은 데이터의 중요도에 따라 데이터 순위를 매기고 필터(filter) 속성 선택 알고리즘이다. SVM-RFE 알고리즘은 속성의 가중치 기반으로 데이터 순위를 매기고 데이터를 감싸는 래퍼(wrapper) 속성 선택 알고리즘이다. 이러한 두 가지 기법을 조합함으로써, 우리는 SVM-RFE는 0.3096779이고 OCFE는 0.3016138으로 에러율 평균이 좀 더 낮게 나타났다. 또한, 제안된 기법은 SVM-RFE가 69%이고 OCFE는 70%으로 좀 더 정확한 것으로 나타났다.

Key Words : Feature elimination method, OCFE, ReliefF, SVM-REF

1. Introduction

Throughout time it is now known that feature selection is holding a key role in classification problems. Including many features in the classifiers did not mean that it shows a better classification performance. With feature selection, it selects the most relevant feature subset to get better accuracy in classification. There are two different feature selection methods based on the criterion: filter models and wrapping models.

Filter models rely on the general characteristics of the training data to select features independently from any predictors, while wrapper models are optimizing a predictor as part of the selection process. The ReliefF algorithm is an exact example of feature selection filter models which can effectively provide quality of each feature in problems with dependencies among the feature spaces[1-2].

Wrapper models are involved a learning algorithm as a black box and consists of using its prediction performance to evaluate the relative usefulness of subsets of variables. One exact example of feature selection

*Corresponding Author : Department of Computer Internet, Gangwon Provincial College(skparkhg@hanmail.net)
Received September 11, 2015 Revised September 18, 2015 Accepted September 25, 2015

wrapping models is SVM-RFE. The technique using SVM-RFE to wrap the data it is first done in [3].

The idea of using SVM-RFE for filtering the selection gene data is first proposed in [3] by making chunks of data subset into features and wraps it with the SVM-RFE to select the data. However, this method have some chances like the result of the method is not the most optimum features subset data and through the method the optimum data features has been eliminated.

Those are why the proposed paper combines these two feature selection to select the best gene for classifier. With combining the filter model and wrapper model, it increases the percentage of accuracy from the feature data. The final subset is the most optimum data which have been classified with SVm through SVM-RFE.

Support Vector Machine (SVM) was already known as one of classifier for linear or non-linear data. It has been applied to many subjects as a classifier, such as pattern recognition [4], bankruptcy prediction model [5], food chemistry [6], gene selection [3]. SVM has a good record as one of classifier method than other classifiers method.

2. Related Work

2.1 ReliefF for Feature Selection

ReliefF is an extended version of Relief Algorithm. The main idea of Relief algorithm is to estimate the quality of features according to how well their values distinguish between instances that are near to each other [7]. RealiefF improves the original algorithm by estimating probabilities more reliably and extends it to deal with incomplete and multi-class data sets [8]. ReliefF is usually implied in data pre-processing for selecting feature subset. This method has low bias, includes interaction among features and may capture local dependencies.

Here is the ReliefF algorithm:

-Input the data matrix D, repeated times: n, the number of the neighbors: k
-Randomly select an instance Rj
-Find k nearest hit h and nearest misses m
-Update the estimation Wi by Equation(1) below and repeat it to all features
-Repeat the step from step 2 until n times

The output of this algorithm is vector W for feature attributes ranking.

First the ReliefF algorithm selects an instance randomly Rj from its class, and then searches for k of its nearest neighbors from the same class. The nearest neighbors k is called nearest hits h, and k nearest neighbors from each of the different classes, namely nearest misses m. After that it updates the quality measure Wi for all attributes i depending on their values like Rj, m, and h. If instances Rjandh have different values of the attribute i then the attributes I separates two instances with the same class which is not desirables owe decrease the quality estimation Wi. While if Rj and those m have different values on feature I, the estimation Wi is increased. The process is repeat edn times. The n is usually set by users and it can be safely set to 10.

According to what mentioned above, the quality of Wi can be update das follows:

$$Wi = Wi - \frac{\sum_{k=1}^{K} Dh(k)}{n.k} + \sum_{c=1}^{C-1} Pc.\frac{\sum_{k=1}^{K} Dm(k)}{n.k} \quad (1)$$

2.2 SVM-RFE

In [5], Guyon et al. used SVM-RFE (Recursive Feature Elimination) to select the gene of cancer classification. They proved that the filtered data sets have more reliable and good results than the unfiltered data sets. They used SVM-RFE to wrap the data and group the data into features. Once the data have been grouped, it go through weight calculation and ranked based on their weight. The least ranked feature is eliminated. The process is on repeat until it reaches the threshold.

Algorithm SVM-RFE:

-Input the training examples and class labels
-Training the Support Vector Machine
-Calculate the weight vector of features
$w = \sum_k \alpha_k y_k x_k$
-Sort the ranking of features based on the weight vector
$c_i = (w_i)^2$
-Find the least ranked feature
-Update the list ranking
-Eliminate the features with the least weight vector
-Repeat the process with the remaining subset of surviving features

It shows that in [3] the SVM-RFE eliminate chunks of data at once rather than eliminate it one by one. So there is might be a problem if the data is small data, because there is a chance the important data also be eliminated if the data is small.

3. A Design of an Optimized Classifier based on Feature Elimination

This paper proposes an optimized classifier based on feature elimination (OCFE) for gene selection. The OCFE combines ReliefF algorithm and SVM-RFE algorithm. OCFE combines the quality and importance ranking of ReliefF algorithm and the weight ranking of SVM-RFE. The combined ranking uses as the basic ranking.

The OCFE eliminates the least ranking on the based ranking continuously while there is surviving feature. The last feature that eliminated is the final feature. Then the final feature goes through the classifier. The purpose of combining between both algorithms is to optimize the final subset result for classifier to classify to get the most accurate classification result.

Guyon et al [3] proposed the use of SVM-RFE for cancer gene classification. The method that was used by grouping chunks of data of gene into features then using SVM-RFE, it eliminates the feature that has the least weight ranking. This process is recursive or repeated continuously if there is still surviving feature. The last feature that eliminated is the important feature.

However this method only ranks the data based on the weight of the feature. There is still a chance that the method can eliminate the important feature which contains some important gene for classification, because this method reduces chunks of gene per elimination.

Therefore the proposed method is combining the use of SVM-RFE with ReliefF algorithm.

ReliefF algorithm is also a selection feature algorithm. But instead of ranking the feature based just on the weight, ReliefF ranks the feature based on the quality or importance of the feature. So with combining the two algorithm between ReliefF and SVM-RFE it reduces the chance elimination of the important feature and to get more optimum feature for classification using SVM.

Optimized Classifier based on Feature Elimination Algorithm

Begin
Input the training examples and class labels
repeat
Training the Support Vector Machine
Calculate the weight vector of features
w = $\sum_k \alpha_k y_k x_k$
Input the data matrix D, repeated times: n, the number of the neighbours: k
Randomly select an instance Rj
Find k nearest hit h and nearest misses m
Update the estimation Wi by Equation(1) and repeat it to all features
Repeat the step from step 2 until n times
Calculate Ci
$c_i = (w_i)^2$
Normalized Ci and Wi with [0,1]
Sort the ranking of features based on Gi
$G_i = W_i + C_i$
Find the least ranked feature
Update the list ranking
Eliminate the features with the least weight vector
until(the process with the remaining subset of surviving features)
end

Step 1: input the training examples and class labels.

Step 2: OCFE starts train the data according to the SVM classifier.

Step 3: The calculation of the weight vector of features with equation (2).

$$w = \sum_k \alpha_k y_k x_k \qquad (2)$$

Step 4: The data that have been trained is inserted to the ReliefF algorithm. The number of repeated times or n can be safely set as 10.

Step 5: OCFE randomly selects an instance of Rj.

Step 6: When the instance Rj have been selected randomly, then it search k from the nearest neighbour in the same class. K called nearest hits h in the neighbors, and k nearest neighbors from each of the different classes are namely nearest misses m.

Step 7: Then it updates the quality measure Wi for all attributes i depending on their values like Rj, m, and h. The estimation Wi is calculated by equation (1). Dh(k) or Dm(k) is the sum of distance between the selected instance and its kth nearest neighbour in h or m. Pc is the prior probability of class c. These steps is repeated from step 5 until step 7 as much as n times.

Step 8: OCFE calculate the Ci with this equation (3).

$$c_i = (w_i)^2 \qquad (3)$$

Step 9: Before the Gi calculation the Wi and Ci is normalized with [0, 1]. The normalization implemented by equation (4).

$$Normalized(W_i) = \frac{W_i - I_{\min}}{I_{\max} - I_{\min}} \qquad (4)$$

Where Imin is the minimum value for variable I and Imax is the maximum value for variable I. The purpose of the normalization is to reduce the error that can be occurred if the data's dimensions are different.

Step 10: In the SVM-RFE algorithm, Ci is

usually used as the basic to sort the ranking features in SVM-RFE. But in our proposed method, instead of Ci we use Gi as the basic to sort the ranking features.

$$G_i = W_i + C_i \quad (5)$$

After all of the calculation is done, the data sort based on Gi.

Step 11: OCFE search the least ranked feature.

Step 12: After that it updates the ranking list.

Step 13: Once the ranking is done, it eliminates the least ranked feature.

Step 14: These steps are all repeated while there is a survive feature subsets which is the final subset.

The main reason combining ReliefF and SVM-RFE is to get the best classification data. The result is more optimum then using just SVM or SVM-RFE. As we know, the function of SVM is to get the best hyperplane between two input classes. With the use of feature selection it optimizes the search of the best margin and hyperplane, because the feature selection result is the most important and optimum feature.

The final result is a data that have been classified and indicate whether it +1 or -1.

4. Simulation of OCFE

The proposed method for gene selection is combining ReliefF algorithm and SVM-RFE as feature selection. The simulation of this technique is conducted using MATLAB. ReliefF algorithm is already included in MATLAB toolbox for feature selection. The Spider [9] is used for the SVM-RFE and SVM source code with some modifications for combining the ReliefF algorithm and SVM-RFE.

4.1 Description of data sets.

The data used a test and train data which available online. The data that has been retrieved go through a simple preprocessing step before input into the program. From each gene expression value, we subtracted its mean and divided the result by each gene standard deviation. The data that used is colon cancer data [10].

It is a MicroArray data set of colon cancer. There is approximately aroung 6500 genes, but only 2000 genes are selected based on the confidence in the measured expression levels. The data contains expression set of 2000 genes and 62 samples. 40 samples are from tumors and 22 samples are from normal biopsies from healthy parts of the colons of the same patients.

There are two classes in the data, normal and tumor. It also noted that the normal colon biopsy included smooth muscle tissues form the colon walls. Smooth muscle related genes exhibit high expression levels in the normal samples for this reason.

4.2 Estimate the classifier accuracy and quality

It stated above that before going through the simulation using the proposed method, the data need to be preprocessed first. The preprocessed data subtract the mean from each gene expression value and divided the result by each gene standard deviation. The result is input to the program for simulation. There are 2 methods that run and compared

in this section, the SVM-RFE and OCFE.

Here is the calculation of accuracy of both methods to compare both of them. To calculate the accuracy, the error rate of each method need to be known because the accuracy is complimented by the error rate of the classifier or method.

Error rate is stated as class_loss in the program, it calculated the means of the classifier results with the standard error.

$$Standard\ error = \frac{Standard\ deviation}{\sqrt{number\ of\ trials}} \qquad (6)$$

After that it combines with calculating the loss with equation (7).

$$L = \sum_{j=1}^{n} w_j e_j / \sum_{j=1}^{n} w_j \qquad (7)$$

Table 1. Error rate set table based on the number of feature

Number of feature	SVM-RFE	OCFE
4	0.193545	0.258061
8	0.19355	0.161293
16	0.225804	0.129033
32	0.35484	0.338711
64	0.35484	0.35484
128	0.35484	0.35484
256	0.35484	0.35484
512	0.35484	0.35484
1024	0.35484	0.35484
2000	0.35484	0.35484

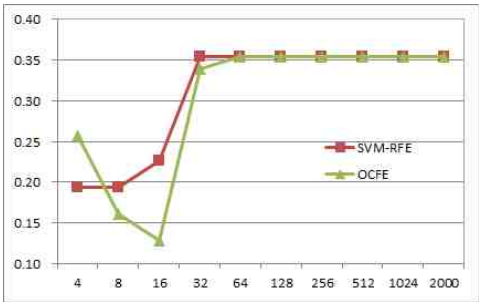

Fig. 3. Error rate based on the number of feature

Table 1 and Fig. 3 showed the error rate that occurred based on the number of feature. For SVM-RFE the error rate is at the most minimum point in the number feature eight then eventually go up until it become stabilize in the 0.35484 point from number feature 32. While for OCFE, the error rate is at the most minimum point at number feature 16 and then eventually go up until it become stabilize in the 0.35484 point from number feature 64.

Fig. 3 also showed that at number feature 4 the error rate of OCFE is higher than SVM-RFE. It proved that OCFE is not quite efficient for handling small data features but more efficient for handling larger data features compared to SVM-RFE which more efficient in handling small data features.

Table 2. Average Error rate set table

	Error Rate
SVM-RFE	0.3096779
OCFE	0.3016138

Table 2 showed the calculation result of average error rate that concluded from the error rate based on the number of feature which has been conducted before. It showed that OCFE algorithm has more low error rate than SVM-RFE. It mean that each folds of the feature data when trained with OCFE method send lesser error report than SVM-RFE with zero rejection.

Accuracy is complimented by the error rate, it calculated by equation (8).

$$Accuracy = 1 - error\ rate \times 100 \qquad (8)$$

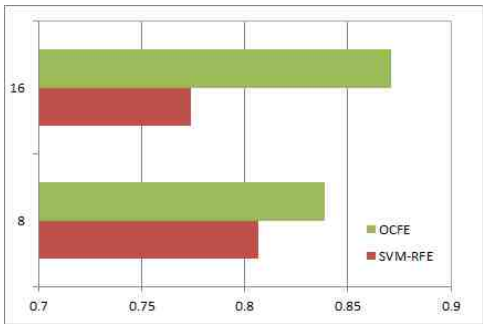

Fig. 4. Accuracy Chart in the minimum error rate point from each method

OCFE has higher accuracy than SVM-RFE even in the most minimum error rate point in the SVM-RFE. Fig. 4 showed in the most minimum error rate point in the SVM-RFE is number of feature 8 and the accuracy of OCFE is higher than SVM-RFE at the number of feature 8. And for OCFE most minimum error rate point is 16 and the accuracy of OCFE indeed is higher than SVM-RFE at the number feature of 16.

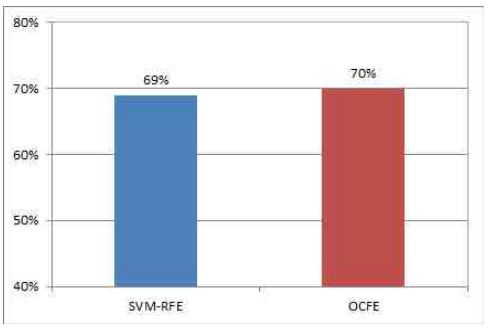

Fig. 5. Average Methods Accuracy Chart

Fig. 5 showed that SVM+ReliefF has higher average accuracy result than SVM-RFE accuracy. SVM-RFE has 69% average accuracy and OCFE has 70% average accuracy. In [3], Guyon et al calculate the classifier quality with error rate, rejection rate at fixed threshold, and classification confidence. In the paper they modified the version of these metrics. The paper use four metrics of classifier quality:

- Error is number of errors at zero rejection
- Reject is minimum number of rejected samples to obtain zero error.
- Extremal margin is difference between the smallest output of the positive class samples and the largest output of the negative class samples. The result is rescale by the largest difference between outputs.
- Median margin is difference between the median output of the positive class samples and the median output of the negative class samples. The result is rescale by the largest difference between outputs.

In the Table 3 and 4 used success rate and acceptance rate instead of error rate and rejection rate. Success rate complement error rate, which error rate is the fraction of examples that are misclassified. Acceptance rate complement the rejection rate, which rejection rate is the fraction of examples that are rejected.

Table 3. Feature Genes trained with SVM-RFE

Number of feature	Training Set				Test Set			
	Vsuc	Vacc	Vext	Vmed	Tsuc	Tacc	Text	Tmed
4	0.81	0.99	0.01	0.70	0.66	0.54	-0.68	-0.02
8	0.81	0.98	-0.07	0.61	0.66	0.59	-0.59	0.06
16	0.87	0.97	-0.08	0.60	0.60	0.53	-0.59	0.06
32	0.81	0.90	-0.15	0.52	0.73	0.60	-0.51	0.15
64	0.68	0.84	-0.21	0.47	0.74	0.70	-0.46	0.19
128	0.62	0.82	-0.23	0.45	0.80	0.76	-0.43	0.22
256	0.63	0.75	-0.29	0.39	0.80	0.80	-0.37	0.27
512	0.63	0.73	-0.31	0.37	0.80	0.81	-0.35	0.29
1024	0.63	0.72	-0.31	0.36	0.80	0.81	-0.35	0.30
2000	0.63	0.70	-0.34	0.34	0.80	0.83	-0.33	0.33

Table 4. Feature Genes trained with OCFE

Number of feature	Training Set				Test Set			
	Vsuc	Vacc	Vext	Vmed	Tsuc	Tacc	Text	Tmed
4	0.93	0.99	0.01	0.69	0.66	0.54	-0.68	-0.02
8	0.94	0.98	-0.08	0.61	0.60	0.59	-0.59	0.06
16	1.00	0.97	-0.08	0.61	0.60	0.53	-0.59	0.06
32	0.93	0.90	-0.16	0.52	0.73	0.60	-0.51	0.15
64	0.81	0.84	-0.21	0.47	0.74	0.70	-0.46	0.19
128	0.62	0.82	-0.23	0.45	0.80	0.76	-0.43	0.22
256	0.63	0.75	-0.29	0.39	0.80	0.80	-0.37	0.27
512	0.63	0.73	-0.31	0.37	0.80	0.81	-0.35	0.29
1024	0.63	0.72	-0.31	0.36	0.80	0.81	-0.35	0.30
2000	0.63	0.70	-0.34	0.34	0.80	0.83	-0.33	0.33

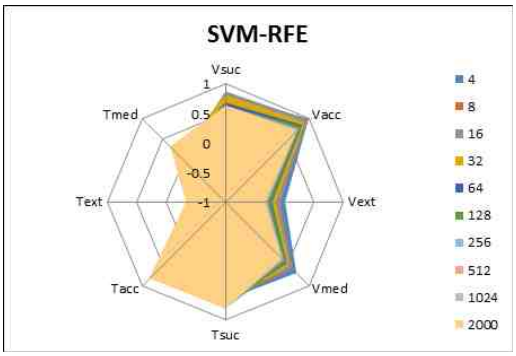

Fig. 6. Feature Genes trained with SVM-RFE

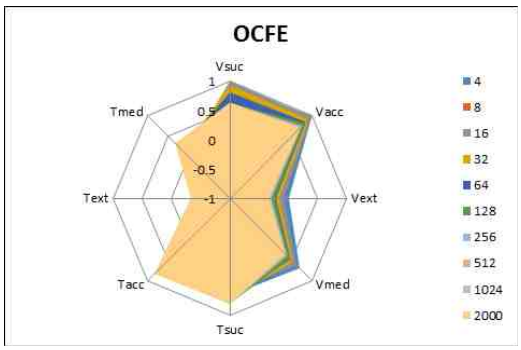

Fig. 7. Feature Genes trained with OCFE

Table 4 and 5 including Fig. 6 and Fig. 7 showed the metrics results of selected feature number with SVM-RFE and SVM-RFE. V indicated the result from training set and T indicated the result from test set. Suc is success rate. Acc is acceptance rate. Ext is extremal margin. Med is median margin.

Here we can conclude that SVM+ReliefF algorithm has more optimum and reliable result than SVM-RFE. It because SVM+ReliefF method combine the ReliefF algorithm that rank the features according to the feature importance and SVM-RFE which rank the features based on the weight of the features vector

Table 5. Methods Elapsed Time

	Elapsed Time
SVM-RFE	11.792127 seconds
OCFE	14.784528 seconds

Table 5 showed the elapsed time each of method take to get the final subset data. This simulation is conducted with ASUS and Intel core i7 processor. The proposed method is taking 2.992401 seconds than SVM-RFE. It indicates that OCFE method is slightly more computationally expensive than SVM-RFE. The difference is only 20% longer than SVM-RFE. However, OCFE has proved the result of the method is more reliable and accurate than SVM-RFE method result.

Hence, even it takes a little bit longer to get the final subset data than SVM-RFE it still prove that the proposed method is efficient because the difference of time elapsed is not that big. It only further prove the proposed method OCFE is reliable, accurate and also efficient.

5. Conclusion

The paper proposed and applied OCFE method, combining SVM-RFE and ReliefF algorithm as feature selection to classify the subset data. ReliefF is a filter feature selection, and SVM-RFE is a wrapper feature selection. ReliefF rank the features data

according to the data importance, while SVM-RFE rank the features data based on the data weight vector.

With combining these two methods, we get more reliable, accurate and efficient method. It showed in the section 4 that the error rate of our proposed method is lower with 0.096771 than SVM-RFE with 0.129031, yet our method accuracy percentage is higher with 90% than SVM-RFE with 80%. Even though our proposed method is more computationally expensive than SVM-RFE, it only 2.992401 seconds longer. So it is still efficient.

REFERENCES

[1] V. B. Canedo, N.S. Marono, A.A. Betanoz, Distributed feature selection: An application to microarray data classification, Applied soft computing, vol.30, pp.136-150, May 2015.

[2] X. Zhou, J. Wang, Feature selection for image classification based on a new ranking criterion, Journal of Computer and Communications, vol.3, pp. 74-79. March 2015.

[3] I. Guyon, J. Wetson, S. Barnhill, M. D. and V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, vol. 46, pp.389 - 422, 2002

[4] Y. Guerbai, Y. Chibani, B. Hadjadji, The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters, Pattern Recognition, vol. 48, no.1, pp.103-113, January 2015.

[5] K.S. Shin, T.S. Lee, H. J. Kim, An application of support vector machines in bankruptcy prediction model, Expert Systems with Applications vol.28, no.1, pp.127-135, January 2005.

[6] O. Devos, G. Downey, L. Dupochel, Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils, Food Chemistry, vol. 148, pp.124-130, April 2014.

[7] M.R. Sikonja, I. Kononenko, Theoritical and Empirical Analysis of ReliefF and RReliefF, Machine Learning, vol.53, no.1, pp.23-69, October 2003.

[8] I. Kononenko, M.R. Sikonja, U. Pompe, ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems, pp.1-15, 1996

[9] J. Weston, A. Eliseff, G. BakIr, F. Sinz, The Spider. Available: http://people.kyb.tuebingen.mpg.de/spider/main.html

[10] Alon et al, Package 'ColonCA'. http://microarray.princeton.edu/oncology/affydata/index.htmle

Author Biography

Byung-Kwan Lee **[Regular member]**

- Feb. 1990 : Chung-Ang Univ., Dept. of Computer Enineering, PhD
- Mar. 1988 ~ current : Catholic Kwandong Univ., Dept. of Computer Engineering, Professor

<Research Interests> Network Security, Big Data, Data mining, IoT

Seok-Gyu Park **[Regular member]**

- Feb. 2005 : Gyeongsang National Univ., Dept. of Computer Science, PhD
- Mar. 2002 ~ current : Gangwon Provincial College, Dept. of Computer Internet, Assistant professor

<Research Interests> System analysis, Software Confidence

Yusrina-Tifani **[Regular member]**

- Feb. 2014 : Bina Nusantara Univ., BS
- Mar. 2015 ~ current : Catholic Kwandong Univ., Dept. of Computer Engineering, the master's course

<Research Interests> Big Data, Data minig, IoT