Advances in Social Science, Education and Humanities Research, volume 218
First International Conference on Science, Mathematics, and Education, (ICoMSE 2017)

Prediction of Disease-Resistant Gene in Rice Based on Support Vector Machine and Fuzzy Kernel C-Means

1st Glori Stephani Saragih
Department of Mathematics, FMIPA
Universitas Indonesia
Depok, Indonesia
glori.stephani@sci.ui.ac.id

***Abstract*— Rice production in Indonesia is important especially for national economy. Rice is the staple food for Indonesian, so Indonesia is the biggest rice consumers in the world, averaging more than 200 kilograms per head each year. Therefore, Indonesia needs to be aware of rice production systems because there are many diseases that distract from the growth of rice. One of that is Bacterial Leaf Blight (BLB) causes severe damages in many rice cultivation regions of the world. Bacterial Leaf Blight disease control through the development of resistant varieties is one of the effective and easiest ways to apply to farmers. The computational method is a way out to solve this problem with the use of machine learning. However, the two most popular, efficient and high accuracy methods are Support Vector Machine (SVM) and Fuzzy Kernel C-Means. Therefore, we propose to compare these two methods to assess the chance of a protein being disease-resistant. In this research, we found that SVM is better than Fuzzy Kernel C-Means because SVM has represented protein information with 90.91% accuracy and Fuzzy Kernel C-Means with 80.58% accuracy in the model. However, if the researcher is stuck on finding data, based on this research Fuzzy Kernel C-Means gives the highest accuracy in smallest dataset then SVM. Fuzzy Kernel C-Means has represented protein information with 80.58% and SVM just 23.2% accuracy in 10% training data set**

Keywords*— *Fuzzy Kernel C-Means, Support Vector Machine, disease-resistant gene

I. INTRODUCTION

This research is to look for the BLB disease control, which is Xoo resistant gene in rice. Xoo stands for *Xanthomonas oryzae pv. oryzae* which causes a bacterial blight disease. BLB can significantly decrease rice production by as much as 20-80%. Currently, BLB is reported to not only damage wetland rice but also upland rice in Indonesia [1]. The one of effective and easiest way to control BLB disease through the development of BLB resistant varieties. Identification of genes is a hard work in experimental studies. There are many disadvantages to experimental method especially time and cost. With the development of the computational method, the focus is put on the machine learning method. Among the various methods in machine learning, SVM and Fuzzy Kernel C-Means are two most popular, efficient and high accuracy rate. Both of that have been widely used to analyzing gene expression data. Y. Ren et al used SVM-RFE to predict disease-resistant gene in rice and Z. Rustam and N. Maghfirah used the SVM-RFE to classify cancer with a good accuracy rate [2,3]. Not only in the microarray database but has been applied in a wide range of application, the used of SVM and Fuzzy Kernel C-Means as classifiers for intrusion detection systems [4,5].

In addition, the study of feature extraction is very common and popular, especially in bioinformatics. Global encoding is the one of feature extraction that use in this study to extract amino acid sequences to vector input to be used in machine learning. Huang et al. used global encoding to predict protein-protein interaction using weighted sparse representation model combined and have a high accuracy even in the purposed method and when SVM as a classifier. Process of global encoding is classifying 20 kinds of amino acids into 6 classes (e.g., A1 = {A, V, L, I, M, C}) and then gets 10 combinations each that contains three different classes (e.g., {A1, A2, A3} vs {A4, A5, A6}). And then, transform 10 combinations protein sequence into 10 binary characteristic sequences. Each characteristic sequences would be further divided into specific numbers of subsequences according to a partition method. Finally, two descriptors, composition and transition, would be extracted from these subsequences to depict the global composition of every protein sequence and form the final feature vectors [6].

II. METHODS

Data

The data used in this research are from National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) public database. The protein dataset is chosen based on literature reported on the *Xanthomonas* resistance [7]. Non-disease resistant amino acid sequences data took through UniPort (https://uniport.org) Protein database. The data will be classified into two classes, disease-resistant and non-disease resistant. After extraction process, the data is divided into training to build a model and testing to validate a model.

Global encoding

Feature extraction in this study started from the transformation of the protein sequence. First, classify 20 kinds of amino acids into six classes based on physicochemical characteristic [6].

Table 1. Amino Acid Classification

Amino Acid Classification	
Aliphatic amino acid:	A1 = {A, V, L, I, M, C}
Aromatic amino acid:	A2 = {F, W, Y, H}
Polar amino acid:	A3 = {S, T, N, Q}
Positive amino acid:	A4 = {K, R}
Negative amino acid:	A5 = {D, E}
Special conformations:	A6 = {G, P}

By doing this, ten combinations will get which each of combination contains three different classes. Ten combinations can be obtained as follows: {A1, A2, A3} vs {A4, A5, A6}, {A1, A2, A4} vs {A3, A5, A6}, {A1, A2, A5} vs {A3, A4, A6}, {A1, A2, A6} vs {A3, A4, A5}, {A1, A3, A4} vs {A2, A5, A6}, {A1, A3, A5} vs {A2, A4, A6}, {A1, A3, A6} vs {A2, A4, A5}, {A1, A4, A5} vs {A2, A3, A6}, {A1, A4, A6} vs {A2, A3, A5} and {A1, A5, A6} vs {A2, A3, A4} [5]. Let's defined protein sequence as $S = s_1, s_2, \ldots, s_{10}$. Ten combinations of sequences will be transformed into ten binary sequences that symbolize as $T_1, T_2, \ldots, T_{10}$. Then, we calculate two numerical sequences $T_1(s_1)$and $T_2(s_2)$ as below [8].

$$T_1(s_i) = \begin{cases} 1, & s_i \in \{A_1, A_2, A_3\} \\ 0, & s_i \in \{A_4, A_5, A_6\} \end{cases} i = 1, 2, \ldots, n \quad (1)$$

$$T_2(s_i) = \begin{cases} 1, & s_i \in \{A_1, A_2, A_4\} \\ 0, & s_i \in \{A_3, A_5, A_6\} \end{cases} i = 1, 2, \ldots, n \quad (2)$$

T_i as the *i*-th characteristic sequence of amino acid sequence and s_i is the *i*-th amino acid of the given protein sequence.

Second, divide each of characteristics sequences to be sum of subsequences with different length and specific strategy. Each characteristics sequences will be divided into specific numbers of subsequences according to a partition method [9]. For any characteristic sequence $T_n = t_1, t_2, \ldots, t_n$of length n, given a positive integer L, T_n will be divided into L subsequences, k-th subsequence is called $SubT_k$ (k = 1, 2,..., L) and $SubT_k$ is composed of the first $\lfloor T_n/L \rfloor$ numbers of T_n.

Third, for the last step, feature vectors of composition and transition descriptors will be extracted from the subsequences. The composition descriptor is the frequencies of '0' and '1' in each subsequence. From composition descriptor, one subsequence contains two frequency values, any character sequence would be represented by a 2*L dimensional feature vector. The transition is count for the switching frequency between '0' and '1' in every subsequence. The times where '0' follows 1' and '1' follows '0' happen.

SVM

Dataset from the extraction process will be classified with SVM. SVM is a discriminative classifier formally defined by separating planes. It means training data will be labeling into '1' for disease-resistant and '0' non-disease resistant. We must suppose that data that we want to classify can be separated by a line. If the data sample is $D = \{(x_i, y_i)\}_{i=1,2,\ldots,n}$ and w is a hyperplane with equation, $wx + b = 0$, where w is a norm vector on the hyperplane. There are two conditions for the data sample:

$$if\ wx + b > 0, then\ D \in A_+$$
$$if\ wx + b < 0, then\ D \in A_-$$

This hyperplane separator called decision boundary. If x_+ and x_- showed the outer vector from each of the closest class to the decision boundary, then the scope through each vector was denoted as $H_+$and H. Range between H_+ and H_- is called margin. While d_+ and d_- showed the range between decision boundary with H_+ and H_-. So margin can be defined by $d = d_+ + d_-$. The decision function that is used in this classification issue is:

$$f(x, y) = sign(wx + b) \quad (3)$$

where w is the normal vector from hyperplane and b is constant which showed the range for hyperplane to the center point.

Two dataset is said linearly separable by a hyperplane if it can be determined as a pair (*w, b*) so the sample data will be classified as A_+ class and A_- class. For $x_i \in D$, so:

$$\begin{matrix} (wx_1 + b) \geq +1, & if\ y_1 = 1 \\ (wx_i + b) \leq +1, & if\ y_i = -1 \end{matrix} \quad (4)$$

Both of the equation can be write down as

$$y_i(< w.x_i > + b) \geq +1, \forall x_i \in D \quad (5)$$

The number of the pair (*w,b*) which can configure hyperplane not just one, so if it is can be found it's not the only one. The optimal hyperplane is the one we looking for, it can maximize the range between the two classes. If $d_+$is range from data $x_+$to hyperplane and d_- is range from data x_- to hyperplane, so the margin is defined as $d = d_+ + d_-$. The margin will be maximum if:

$$d = d_+ + d_- = \frac{1}{\|w\|}(|wx_+ + b| + |wx_- + b|) = \frac{2}{\|w\|} \quad (6)$$

because of this issue, for search the maximum margin ($d_+ = d_-$) is equal with minimizing $\|w\|^2$. So the mathematical model of the problem as follows:

$$\min \frac{1}{2}\|w\|^2$$
$$obstacles: y_i(< w.x_i > + b) \geq 1, i = 1, 2, \ldots, n \quad (7)$$

In this research, SVM is used to classify disease resistant and non-resistant gene in rice, by applying RBF as a kernel.

Fuzzy Kernel C-Means

Fuzzy C-Means method is one of the clustering method that similar to the K-Means method in which each group is represented by a central group or centroid. The selection of group members is determined by the distance between data with the centroid. The main difference between the two methods is that k-means are a hard clustering where a data can only be a member of a group, while Fuzzy C-Means is a soft clustering where a data may belong to multiple groups. Given a set of data $\mathrm{x_n}, \mathrm{n} = 1, \ldots, \mathrm{N}$, suppose there is k-cluster with centroid $\mu_\mathrm{k}, \mathrm{k} = 1, \ldots, \mathrm{K}$. Each of data $\mathrm{x_n}$ has an indicator variable $\mathrm{r_{nk}} \in [0,1]$ which shows association degree of $\mathrm{x_n}$ to k-group. Next, the sum of distance each data to centroid can be represented by the equation:

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2 \quad (8)$$

Then the solution of clustering problem using Fuzzy C-Means clustering method is:

$$\min_{\mu_k} J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2 \quad (9)$$

First step in Fuzzy C-Means is initialization centroid $\boldsymbol{\mu_k}$ and $m \in R, m \geq 1$ is the fuzziness degree for cluster partition. Then we calculate the membership $\boldsymbol{r_{nk}}$ by the equation:

$$r_{nk} = \frac{1}{\sum_{j=1}^{K}\left(\frac{\|x_n - \mu_k\|}{\|x_n - \mu_j\|}\right)^{\frac{2}{m-1}}} \quad (10)$$

After that, we calculate the centroid $\boldsymbol{\mu_k}$,

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = \mathbf{0} \rightarrow \mu_k = \frac{\sum_{n=1}^{N} r_{nk}x_n}{\sum_{n=1}^{N} r_{nk}} \quad (11)$$

Then, evaluate the convergence of parameter $\boldsymbol{\mu_k}$. If the criteria of convergence unfulfilled, we back to the second step, which is we calculate back from update the membership.

The accuracy of Fuzzy C-Means classification depends on the type of data. If the data is non-linear, its convergence is slow and inaccurate. We can solve this problem to transforming the dataset into another space (feature space) that the dimension is much higher than the data space. But learning is very hard in high dimensional data, it can make classification overfitting and high computational cost. In this case, we use Kernel method, with use a "connector" between data space and feature space. So, we have a better accuracy without directly working at feature space. Set a nonlinear mapping φ from input data space $\mathbb{R}^\mathrm{d}$ into feature space F. Then, find a way to measure the distance between transformed data $\varphi(\mathrm{x})$ and $\varphi(\mathrm{y}), \mathrm{x}, \mathrm{y}$ are objects at data space without knowing the explicit form of φ [5]. We calculate the kernel function K, measure distance between $\varphi(\mathrm{x})$ and $\varphi(\mathrm{y})$ by:

$$\begin{aligned} \mathrm{d}^2(\varphi(\mathrm{x}),\varphi(\mathrm{y})) &= \|\varphi(\mathrm{x}) - \varphi(\mathrm{y})\|^2 \\ &= \varphi(\mathrm{x})^\mathrm{t}\varphi(\mathrm{x}) - 2\varphi(\mathrm{x})^\mathrm{t}\varphi(\mathrm{y}) + \varphi(\mathrm{y})^\mathrm{t}\varphi(\mathrm{y}) \\ &= \mathrm{K}(\mathrm{x},\mathrm{x}) - 2\mathrm{K}(\mathrm{x},\mathrm{y}) + \mathrm{K}(\mathrm{y},\mathrm{y}) \end{aligned} \quad (12)$$

In this research, we use Fuzzy Kernel C-Means by applying kernel method into Fuzzy C-Means to clustering disease resistant and non-resistant gene in rice.

Algorithm

The following stepwise procedure is employed so as to implement the algorithm:

1. Download and prepare the amino acid sequence data.
2. Assign labels 1 or 0 to disease-resistant or non-disease resistant gene separately.
3. Extract the amino acid sequence use global encoding to convert amino acid sequence into a vector input to be used in the classification method.
4. Divide the data into training and testing dataset.
5. Run SVM and Fuzzy Kernel C-Means, then obtained the trained model which have derived the classification rules.
6. Run machine learning methods on the test data to assess the prediction.

III. RESULTS AND DISCUSSION

In this research, we used parameter L = 5 while extraction process with global encoding. It means, there are 150 features in this experiment. We used SVM and Fuzzy Kernel C-Means as classification method and we will compare the accuracy between these two classification methods with all of the features in the dataset. Both of the methods use Gaussian-RBF Kernel with parameter 0.05.

Table 2. SVM Results in Predicting Disease-Resistant Gene in Rice

%Training Data	%Accuracy [a]	Running Time
10	23.30	0.13
20	44.57	0.20
30	58.75	0.28
40	65.22	0.34
50	85.96	0.48
60	84.78	0.48
70	88.24	0.55
80	86.96	0.59
90	90.91	0.67

a Accuracy = (number of the true prediction / number of testing data) * 100%

Table 2 shows that SVM has the best accuracy at 90.91% to predict disease-resistant genes in rice with training data used 90%.

Table 3. Fuzzy Kernel C-Means Results in Predicting Disease-Resistant Gene in Rice

%Training Data	%Accuracy [a]	Running Time
10	80.58	0.031
20	79.35	0.063
30	43.75	0.141
40	43.48	0.172
50	33.33	0.188
60	30.43	0.219
70	29.41	0.266
80	47.83	0.281
90	36.36	0.344

a Accuracy = (number of the true prediction / number of testing data) * 100%

Table 3 shows that Fuzzy Kernel C-Means has the best accuracy at 80.58% to predict disease-resistant genes in rice with training data 10%. From Table 2 and Table 3 we can conclude that the accuracy using all features in a dataset with classification using SVM will give the highest accuracy than classification using Fuzzy Kernel C-Means.

IV. CONCLUSION

The method used in this research is to combine global encoding as a feature extraction method to extract amino acids into vector inputs and then create models with SVM and Fuzzy Kernel C-Means methods. In the rice genetic data classification, the highest accuracy was 90.91% using the SVM based RBF kernel and 80.58% using the RBF kernel based on Fuzzy Kernel C-Means. The Fuzzy C-Means kernel shows that with a high percentage of training data it makes a low accuracy percentage and this is the opposite of SVM. From that result, it can be concluded that SVM is the best method than Fuzzy Kernel C-Means, but SVM is not necessarily the best in other analysis, because if we have trouble finding data then we can use Fuzzy Kernel C-Means as machine learning method because Table 3 shows with 10% training data, Fuzzy Kernel C-Means already gives highest accuracy result than SVM. Fuzzy Kernel C-Means accuracy was 80.58% but SVM just 23.2%.

REFERENCES

[1] Y. Suryadi, T. S. Kadir, A. Ruskandar, "Bacterial blight of upland rice", Proc. 3rd Asian Conf. Plant Pathol. The Role of Plant Pathology in Rapidly Globalizing Economies of Asia (Faculty of Agriculture, ICPP, Universitas Gadjah Mada, Yogyakarta), April 2016.

[2] Y. Ren, D. Wang, Y. Wang, J. Zhou, H. Zhang, Y. Zhou, and Y. Liang, "Prediction of Disease-Resistent Gene in Rice Based on SVM-RFE", in 3rd International Conference on Biomedical Enginering and Informatics, 2010.

[3] Z. Rustam and N. Maghfirah, "Correlated based SVM-RFE and feature selection for cancer classification using microarray databases", in The 3rd International Symposium on Current Progress (Mathematics and Sciences, Universitas Indonesia, 2017), unpublished.

[4] Z. Rustam and N. P. A. Audia, "Comparison between support vector machine and fuzzy kernel c-means as classifiers for intrusion detection system using chi-square feature selection", in The 3rd International Symposium on Current Progress (Mathematics and Sciences, Universitas Indonesia, 2017), unpublished.

[5] Z. Rustam and A. S. Talita, "Fuzzy kernel c-means algorithm for intrusion detection systems", Journal of Theoretical and Applied Information Technology, vol. 81(1), 2015.

[6] Y. Huang, Z. You, X. Chen, K. Chan, X. Luo, "Sequence-based prediction of protein-protein interaction using weighted sparse representation model combined with global encoding", BMC Bioinformatics, pp. 74-184, 2016.

[7] X. Jingbo, H. Xuehai, S. Feng, N. Xiaohui, Z. Silan, "Prediction of disease-resistant gene by using artificial neural network", in International Conference on Research Challenges (Computer Science, College of Science, Huazhong Agricultural University,Wuhan), 2009.

[8] P. He and J. Wang, "Numerical characterization of DNA primary sequence", Internet Elec J Mol Des, vol. 1, pp. 668-674, 2002.

[9] X. Li, B. Liao, Y. Shu, Q. Zeng, J. Luo, "Protein functional class prediction using global encoding of amino acid sequence", in Journal of theoretical biology, pp. 290-293, 2009.

[10] V. Panca and Z. Rustam, "Application of machine learning on brain cancer multiclass classification", in The American Institute of Physics (AIP) Conference, 2017.