Multiclass classification of leukemia cancer data using Fuzzy Support Vector Machine (FSVM) with feature selection using Principal Component Analysis (PCA)

I R Fauzi, Z Rustam and A Wibowo
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Corresponding author’s email: rustam@sci.ui.ac.id

Abstract. Cancer is the second leading cause of death globally. According to WHO prediction (2015) cases of cancer deaths will increase to 21.6 million cases by 2030. Therefore, early detection of cancer is necessary to avoid the spread of cancer and machine learning is required to increase performance in the detection of cancer. In general, microarray cancer data consist of many features. However, there are several features in cancer data that did not have important information in classification cancer. Therefore, these features will be summarized from several features under some common underlying factors into fewer components using the Principal Component Analysis (PCA) method. Then, we select the most features who have important information for classification cancer. This paper focuses on the comparison of using and without the PCA method on cancer data coupled with the Fuzzy Support Vectors Machines (FSVM) method for cancer classification. The experimental results, without the PCA method on cancer data coupled with the FSVM method for cancer classification the accuracy is 87.69 % and by using the PCA method on cancer data coupled with the FSVM method for cancer classification the accuracy is 96.92 % (obtained by using 60 features).

Keywords: Cancer, classification, fuzzy support vector machines, principal component analysis

1. Introduction
Cancer is the second leading cause of death globally. Leukemia cancer is one type of cancer that attacks the blood and blood-forming tissues. Leukemia cancer most commonly affects children under 15 years of age. One attempt to reduce the development and spread of cancer using a learning machine is by the method of early detection. The early detection method of cancer is by utilizing information contained in DNA structure (deoxyribonucleic acid) in the body. This type of data is represented in a matrix called microarray data. In Visintin [1], early detection methods in cancer patients can significantly reduce cancer death rates.

The classification method used in this study is Fuzzy Support Vector Machines (FSVM) using the leukemia cancer data. This method is the result of a composite of the SVM method and fuzzy membership function of each data. In Congqin-Yi [2], they used FSVM for breast cancer data classification without feature selection. Their results showed FSVM produces high accuracy to classification on their data experiment.
Cancer data has many attributes (usually called features). Some of these features have slight information for classification process cancer data. Therefore, a selection of features will be performed prior to the classification of leukemia cancer data in Manning [3]. In this paper, we will use Principal Component Analysis as a method of to summarize from several features under some common underlying factors into fewer components. Principal Component Analysis often used in many studies to feature selection as in Engelbrecht [4].

This paper is organized as follows. Section 2 describes Principal Component Analysis as a feature selection. Sections 3 describes FSVM as a classification method. Section 4 presents the experimental result and section 5 gives a conclusion.

2. Principal component analysis

One attempt to reduce the features of high dimension is summarized the features from several features under some common underlying factors into fewer components, which is the preprocessing stage of the classification process. In this paper, the authors used the selection of features from the extraction of leukemia cancer data using Principal Component Analysis method (hereinafter referred to as PCA). PCA is a feature extraction method, which extracts high-dimensional data to new data with dimensions lower than the initial dimension in Engelbrecht [4]. The selection of a subset is a selection technique that uses a feature set that is considered an optimal feature, which is the features who have more important information in Saeys [5].

Principal Component Analysis is an unlabeled feature extraction method that implements the transformation of data. Features will be projected onto a new feature space with lower dimensions in Engelbrecht [4]. New features obtained from the PCA extraction results are features with the most significant information. The main components are obtained by maximizing data variance. The number of new dimensions (number of features) is smaller than the number of original features, so the data can be visualized in a low-dimensional Principal Component space in Chapman [6].

The PCA algorithms can be described as follows in Aburomman [7].

Step 1: Compute mean from each feature as follow:

$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, \quad i = 1,2, ..., n \quad j = 1,2, ..., m$  \hspace{1cm} (1)

here, $\bar{x}_j$ is the data $j$-feature, and $x_{ij}$ is the data $i$-sample with $j$-feature, $m$ is the number of features, and $n$ is the number data of sample.

Step 2: Compute $\Phi$ as follows:

$\Phi = [\Phi_{ij}] = [x_{ij} - \bar{x}_j]$  \hspace{1cm} (2)

$\Phi$ is a matrix of size $n \times m$.

Step 3: Compute the covariance matrix as follows:

$C = \frac{1}{n-1} \Phi^T \Phi$  \hspace{1cm} (3)

$C$ is a matrix of size $m \times m$.

Step 4: Compute eigenvalues of $C$ matrix by solving the following equation

$\det (\lambda I - C) = 0$  \hspace{1cm} (4)

$I$ is the identity matrix, and $C$ is a covariance matrix.
Then, compute eigenvectors $\mathbf{x}$ corresponding to the eigenvalues $\lambda$ by solving the following equation:

$$(\lambda I - C)\mathbf{x} = 0$$  \hspace{1cm} (5)

**Step 5** Sort the eigenvectors based on the eigenvalues from the largest to the smallest and form matrix $\mathbf{x}'$ with the corresponding eigenvectors.

**Step 6** Compute the principal component as below:

$$\mathbf{PC} = \Phi \mathbf{x}'$$  \hspace{1cm} (6)

3. Classification using fuzzy support vector machine

Classification is a process of grouping objects into predefined categories. It is different from the clustering that classifies objects without targeting the classification, objects in the classification have been divided into several categories. In this process, a model that describe the rule of classifying subjects into several classes will be determined. The model could be used to assign a new subject into a class that already has a label [8].

According to Honeine [9], the multiclass classification is defined as the development of classification problems with classes on each of more than two data. The most common way of solving multiclass classification problems is to modify the multiclass problem into a two-class problem. After modify the multiclass problem into a two-class problem, the results are combined. Two common approaches in doing this are one versus rest and one versus one. The main difference between the two approaches lies in the process of dividing the samples into two-class problems and how to combine predictions of two-class problems into a final prediction of multiclass problems. In the one versus rest approach, there are as many $l$ problems of two classes, where each problem of two classes consists of a class containing all the data of the class $((k = 1, 2, ..., l))$ and a class containing the combined data from other classes. In the one versus one approach, it is done as much as $\frac{l(l-1)}{2}$ two-class problem, where each two-class problem consists of a class containing all data from the to-$k$ th class ($k = 1, 2, ..., l$) and a class containing all data from the to-$m$ th class ($m = 1, 2, ..., l, k \neq m$).

In Abe [10], Fuzzy Support Vector Machine developed by Shigeo Abe and Takuya Inoue, Abe proposed a fuzzy membership to address the problem of unclassified areas in the SVM method, and FSVM method can also reduce the effect of the outliers in the data classification process. Each data is given a degree of membership that states the contribution rate of that data to each class. In this method degree of membership of each data will be calculated, then the degree of membership that has been obtained will be used in the data to find the best hyperplane in the Support Vector Machines classification model.

Support Vector Machines (SVM) developed by Vapnik [11]. Instead of minimizing the traditional empirical risk (the error on the training data), SVMs minimize an upper bound on the expected risk derived from the capacity of hypotheses. This is made possible by constructing a classifier that separates training samples and maximizes the margin or the minimum distance between the decision surface and training samples. Theoretically, SVM is able to correctly classify any linearly separable data [12-14] which can be describe as a set of data points $(x_i, y_i)$, $i = 1, 2, ..., N$, with $x_i = [x_{i1}, x_{i2}, ..., x_{im}]$ is the set of input $m$ dimension (number of features), and $y_i \in \{-1, +1\}$ is the class label for the corresponding $x_i$, then the hyperplane that separated the two classes of the data is defined by the equation $w^* \cdot x + b = 0$. The following defines the decision function [11]:

$$f(x_i) = w^* \cdot x_i + b^*$$  \hspace{1cm} (7)
\[ f(x_t) = \text{sgn} \left( \sum_{i=1}^{N} a_i^+ y_i K(x_t, x_i) + b^+ \right) \]  

(8)

with \( w \) representing the weight parameter vector showing the normal vector of hyperplane, \( b \) is the bias parameter, with \( (\sum_{i=1}^{N} a_i^+ y_i K(x_t, x_i) + b^+) > 0 \) representing that the class label of \( x_t \) is \( y_t = 1 \) and \( (\sum_{i=1}^{N} a_i^- y_i K(x_t, x_i) + b^-) < 0 \) indicates that the class label of \( x_t \) is \( y_t = -1 \) [11].

In our paper, first we find the fuzzy membership value for each data, after that we will be formed new data (the result of the multiplication value of fuzzy membership with each feature) for representative data. The membership value of each data will be classified using Support Vector Machines.

4. Results

Leukemia cancer data set from URL http://www.gems-system.org/ was used to apply the method. The dataset containing expression levels of 5327 features and 72 samples with 3 class: AML/Acute Myeloid Leukemia 39 object, ALL T-Cell/Acute Lymphocytic Leukemia that affect T-cell lymphocytes 9 object, and ALL B-Cell/Acute Lymphocytic Leukemia that affect B-cell lymphocytes 27 object. Training data consist of 10 %, 20 %, to 90 % of dataset. The testing data is used to measure the accuracy and running time of the classification results using MATLAB R2015b (https://www.mathworks.com/products/new_products/release2015b.html). The following table shows accuracy of multiclass classification and running time which used PCA as feature selection methods with RBF kernel \( \sigma = 0.05 \).

In general, in table 1 we can see that by using 80 % training data, obtained the highest accuracy results than the percentage of other training data. At 50 best features, the highest accuracy is 92.858 % with 50 % training data. While the selection of 40 best features produces an accuracy of 91 % with 70 % of training data.

Based on table 1, the lowest accuracy was obtained by 51.976 % with running time 4.1568 seconds on 10 number of features and 10 % training data. While the greatest accuracy obtained in table 1 using Principal Component Analysis method reached 96.924 % with running time 8.2132 seconds on 60 number of features and 80 % of training data.

Table 1. Accuracy multi-class classification of Leukemia Cancer Data using FSVM with PCA as feature selection using RBF kernel with \( \sigma = 0.05 \).

| Training data (%) | Number of features |
|-------------------|--------------------|
|                   | 10  | 20  | 30  | 40  | 50  | 60  | 70  | 80  | 90  | 100 |
| 10                | 51.876 | 65.316 | 72.190 | 67.814 | 65.626 | 76.566 | 70.000 | 61.544 | 52.188 | 64.378 |
| 20                | 74.982 | 79.648 | 85.262 | 75.438 | 78.594 | 88.066 | 78.246 | 69.124 | 78.244 | 79.298 |
| 30                | 67.756 | 74.694 | 83.674 | 79.592 | 54.488 | 85.714 | 82.858 | 80.410 | 81.632 | 82.856 |
| 40                | 72.856 | 81.428 | 78.096 | 89.524 | 92.858 | 88.028 | 91.430 | 87.887 | 85.236 | 82.383 |
| 50                | 77.144 | 87.144 | 85.714 | 90.286 | 90.000 | 90.000 | 91.430 | 89.144 | 85.142 | 82.858 |
| 60                | 70.000 | 84.000 | 89.286 | 80.000 | 86.156 | 89.286 | 89.288 | 87.858 | 86.430 |
| 70                | 80.000 | 84.000 | 88.000 | 91.000 | 90.000 | 92.000 | 87.000 | 87.000 | 84.000 | 89.000 |
| 80                | 83.080 | 89.232 | 92.130 | 89.234 | 90.772 | 96.924 | 92.310 | 93.848 | 90.772 |
| 90                | 82.500 | 87.500 | 85.000 | 82.500 | 85.000 | 95.000 | 90.000 | 85.000 | 86.000 | 87.500 |
Table 2. Comparison multiclass classification of leukemia data using FSVM using RBF Kernel $\sigma = 0.05$ without feature selection and with 60 feature extracting used PCA

| Training data (%) | Without feature selection | 60 Feature after extracting used PCA |
|-------------------|--------------------------|------------------------------------|
|                   | Accuracy (%)  | Running time (s) | Accuracy (%)  | Running time (s) |
| 10                | 40.626         | 4.8948            | 76.566         | 3.8724            |
| 20                | 66.316         | 5.8456            | 88.066         | 4.3038            |
| 30                | 70.614         | 6.5318            | 85.714         | 5.2594            |
| 40                | 80             | 6.5416            | 88.028         | 5.934             |
| 50                | 73.096         | 5.9896            | 90.858         | 6.4308            |
| 60                | 74.284         | 6.3528            | 89.286         | 7.2514            |
| 70                | 85             | 6.826             | 92             | 7.5816            |
| 80                | **87.694**     | **7.1202**        | **96.924**     | **8.2132**        |
| 90                | 80             | 10.6658           | 95             | 8.7132            |
| Average           | **73.07**      | **6.752**         | **89.16**      | **6.395**         |

From table 2, it can be seen that the greatest accuracy result of classification on leukemia cancer data using 60 features of PCA extraction result is 96.924 % with running time 8.2132 seconds, while the greatest accuracy result of classification on leukemia cancer data without feature selection is 87.694 % with running time 7.1202 seconds. In addition, the average result of classification accuracy on leukemia cancer data using 60 features of PCA extraction result is 89.16 % with an average running time 6.395 seconds higher than the average result of classification accuracy on leukemia cancer data without feature selection is 73.07 % with average running time 6.752 seconds. Thus, the accuracy of multiclass cancer classification data using FSVM with feature selection of PCA extraction results is overall better than the multi-class cancer classification data using FSVM without feature selection.

5. Conclusion
In this paper, we propose to compare the results of accuracy multi-class classification of leukemia cancer data using FSVM method without and by using PCA as feature extraction. We use the PCA as feature selection to increase accuracy in classification of leukemia cancer data. The result is the accuracy in the multi-class classification of leukemia cancer data using FSVM with PCA as feature selection (60 feature extraction) is 96.924 % and the accuracy in the multiclass classification of leukemia cancer data using FSVM without feature selection is 87.694 %. Therefore, multi-class classification of leukemia cancer data using FSVM with PCA as feature selection gives better accuracy than a multiclass classification of leukemia cancer data using FSVM without feature selection. The running time when we use multiclass classification of leukemia cancer data using FSVM with PCA as feature selection was faster than that if we use multiclass classification using FSVM without feature selection. Moreover, multiclass classification of leukemia cancer data using FSVM gives better accuracy than the multiclass classification of leukemia cancer data using SVM. For future work, this research could be continued to find a better method that can give high accuracy in a short time to solve the classification of leukemia cancer data.
References

[1] Visintin I, Feng Z, Longton G, Ward D C, Alvero A B, Lai Y, Tenthorey J, Leiser A, Flores-Saaib R, Yu H, Azori M, Rutherford T, Schwartz P E and Mor G 2008 Clin. Cancer Res. 14 1065-72

[2] Congqin-Yi, Zhou R and Hu K 2017 2nd International Conference on Image, Vision and Computing (Chengdu) (United States: IEEE)

[3] Manning C D, Raghavan P and Schütze H 2008 Introduction to Information Retrieval (New York: Cambridge University Press, United States)

[4] Engelbrecht A P 2007 Principal Component Learning (Chichester: John Willey & Sons Ltd, England)

[5] Saets Y, Inza I, and Larranaga P 2007 Bioinformatics 23 2507-17

[6] Chapman K W, Lawless H T and Boor K J 2001 J. Dairy Sci. 84 12-20

[7] Aburomman A A and Reaz M M I 2016 Proceedings of 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (Xi’an) (New York: Institute of Electrical and Electronics Engineers Inc.) p 636-40

[8] Han J, Kamber M, Pei J 2012 Data Mining: Concepts and Techniques (Waltham: Morgan Kaufmann, United States)

[9] Honeine P, Noumir Z and Richard C 2013 Signal Process. 93 1013-26

[10] Abe S and Inoue T 2002 European Symposium on Artificial Neural Networks (Bruges) (Bruges: ESANN, Belgium)

[11] Vapnik V N 1998 Statistical Learning Theory (New York: John Wiley & Sons Inc, United States)

[12] Kecman V 2001 Learning and Soft Computing: Support Vector Machines, Neural Network and Fuzzy Logic Models (United States of America: Library of Congress Cataloging in Publication Data)

[13] Smola A J and Schölkopf B 2000 Sparse Greedy Matrix Approximation for Machine Learning available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.3153

[14] Christianini N and Taylor J S 2000 An Introduction to Support Vector Machines and Other Kernel-based Learning Methods (London: Cambridge University Press)