Journal of Computer Science

Original Research Paper

Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification

Adiwijaya, Untari N. Wisesty, E. Lisnawati, A. Aditsania and Dana S. Kusumo

School of Computing, Telkom University, Bandung, Indonesia

Article history
Received: 12-05-2018
Revised: 04-09-2018
Accepted: 09-10-2018

Corresponding Author:
Adiwijaya
School of Computing, Telkom University, Bandung, Indonesia
Email: adiwijaya@telkomuniversity.ac.id

Abstract: Cancer is one of the most deadly diseases in the world. The International Agency for Research on Cancer (IARC) noted 14.1 million new cancer cases and 8.2 million deaths from cancer in 2012. In the last few years, DNA microarray technology has increasingly been used to analyze and diagnose cancer. Analysis of gene expression data in the form of microarray allows medical experts to ascertain whether or not a person suffers from cancer. DNA microarray data has a large dimension that can affect the process and accuracy of cancer classification. Therefore, a classification scheme that includes dimension reduction is needed. In this research, a Principal Component Analysis (PCA) dimension reduction method that includes the calculation of variance proportion for eigenvector selection was used. For the classification method, a Support Vector Machine (SVM) and Levenberg-Marquardt Backpropagation (LMBP) algorithm were selected. Based on the tests performed, the classification method using LMBP was more stable than SVM. The LMBP method achieved an average 96.07% accuracy, while the SVM achieved 94.98% accuracy.

Keywords: Cancer Detection, Classification, Dimensional Reduction, PCA, SVM, LMBP

Introduction

Cancer is one of the major causes of human mortality in many countries. According to the WHO 2018, there were 14.1 million new cancer cases, 8.2 million deaths from cancer and 32.6 million people suffering from cancer worldwide. To address this growing concern, a new technology that can accurately analyze and detect cancer is needed so that cancer could be treated at its early stage.

For the last few years, microarray data has been used to analyze and diagnose cancer. DNA microarray is a technology used to monitor large numbers of various gene expressions at the same time. Microarray technology has been used for medical diagnosis and gene analysis, especially for analyzing the pattern changes in gene expression under certain conditions. Gene expression analysis can assure medical experts whether or not a patient suffers from cancer within a relatively shorter time than traditional methods.

The number of recorded genes for predicting cancer in one individual is significant. The number of genes is not proportional to the number of individual samples. Therefore, this large dimensional data has very high complexity. Vanitha *et al*. (2015) stated that the analysis of gene expression requires the identification of informative genes. Siang *et al*. (2015) explained that gene expression classification or cancer classification is the process of identifying informative genes that can be used to predict new sample classes. Therefore, an optimal solution is needed to ensure an efficient classification scheme for gene expression (microarray), as this will enable the handling of complex data besides yielding more accurate results within a relatively short time.

Techniques, methods, or classification processes are a field of bioinformatics used to analyze or detect cancer. Researchers have conducted numerous researches on cancer classification methods based on microarray data. For example, Vanitha *et al*. (2015) built a Mutual Information-based Gene Selection scheme (MI) as a feature selection approach with various classifier methods to deal with microarray data. In the research, accuracy results were obtained based on LOOCV mean

accuracy rate. The accuracy results obtained on colon cancer data based on the Artificial Neural Network method, SVM with linear kernel, SVM with RBF kernel and SVM with polynomial kernel were 0.5094, 0.6774, 0.6051 and 0.4683, respectively. The research concluded that SVM with linear kernel and RBF yielded better accuracy than ANN. The same result was also shown in Pirooznia *et al*. (2008), who focused on cancer microarray classification with various classifier methods. From the various scenarios and data sets used, SVM and BP showed higher accuracy compared to the family of decision tree algorithms. This was because multiple output attributes were not allowed in the decision tree and algorithms were unstable. To overcome the problem, Aydadenta and Adiwijaya (2018) used clustering k-means algorithm to group features with high similarity. Relief method was used to sort the clusters. The dimension reduction process showed increasing classification performance with Random Forest method. Meanwhile, Seeja (2011) constructed F-Score and SVM schemes as the feature selection and microarray data classifier, respectively. Using leukemia data, the scheme achieved a faster running time than ANN. Nurfalah *et al*. (2016) developed a scheme using PCA and MBP (Modified Backpropagation using Conjugate Gradient) as the dimension reduction method and microarray data classification, respectively. The scheme yielded 96% accuracy for ovarian cancer, 76.92% for colon cancer and 97.14% for leukemia data. The research showed that the combination of PCA and MBP methods resulted in faster training time than the standard Backpropagation method.

Based on previous researches, the general scheme in the process of classification of microarray data for the detection of proposed cancer can be conducted via three stages, namely preprocessing, dimensional reduction and gene classification. In this study, the step for dimensional reduction was performed using a Principal Component Analysis (PCA), where feature selection (feature extraction) was performed based on the proportion of cumulative variance. Two classification methods were used: Support Vector Machine and Multilayer Neural Network with a Backpropagation learning algorithm (Levenberg-Marquardt). In the gene classification method described by Furey *et al*. (2000), SVM showed good performance with a simple kernel for microarray data analysis. Meanwhile, the Levernberg-Marquard Backpropagation algorithm isone of the methods for weight optimization in a Neural Network that also produces quite a good performance. The Levernberg-Marquard Backpropagation algorithm is known to achieve superior speed convergence due to its use of the Gauss-Newton Algorithm. Itis also stable due to the use of asteepest descent algorithm. In the research of Wisesty *et al*. (2016) and Kişi and Uncuoğlu (2005), the Levernberg-Marquard Backpropagation algorithm was shown to achieve better accuracy than the Backpropagation method with Conjugate Gradient optimization algorithm. Therefore, this research will analyze and compare the performance accuracy of the above-mentioned methods bycombining PCA-SVM and PCA-Levernberg-Marquard Backpropagation, as well as analyze the optimal selection of features and input values of cancer data used against system performance as a method for detecting cancer in microarray data.

Research Scheme

The general scheme proposed in this research involves a process of several stages. Figure 1 shows a flow diagram of these stages. The cancer data examined in this research include leukemia, colon cancer and breast, central nervous system, lung, ovarian and prostate tumors. The data was obtained from Kent-Ridge Biomedical Data Repository reported in Li (2013). In addition to large dimensions, the range of microarray data values is also quite large. Generally, a classification system that has a large data range produces low accuracy. Therefore, according to Adiwijaya *et al*. (2014), these data need to be transformed into a range of 0 to 1. The next stage involves dimension reduction aimed at reducing complexity of data and finding informative genes. The final stage is the process of classification of microarray data to determine whether or nota person is suffering from cancer. Detailed explanation on the research process is in the following subsections.

Dataset

In this study, the data used for cancer classification are colon cancer, ovarian, central nervous system, lung, prostate and leukemia data in the form of microarray data. The data sets were obtained from the Kent-Ridge Biomedical Data Repository reported by Li (2013). The specifications of the data can be seen in Table 1.

Table 1: Data set specification

Data	Number of classes	Sample	Feature
Colon	2	62 (22 Positives, 40 Negatives)	2,001
Leukemia	3	72 (24 ALL, 28 AML, 20 MLL)	12,582
Ovarian	2	253 (91 Normal, 162 Cancer)	15,155
Central Nervous System	2	60 (21 Class1, 39 Class0)	7,129
Lung Cancer	2	181 (31 Mesothelioma, 150 ADCA)	12,533
Tumor Prostate	2	136 (77 Tumors, 59 Normal)	12,600

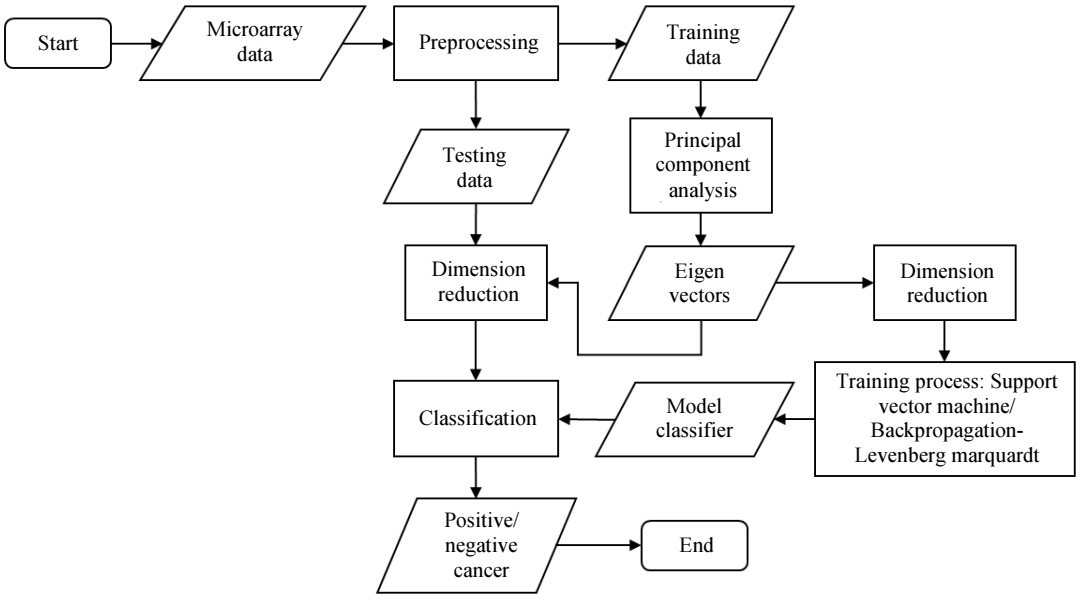

Fig. 1: General scheme of the cancer detection process based on microarray data

Preprocessing

Preprocessing is the process undertaken to make the data easier to use. In the preprocessing stage, the normalization of data was carried out-by changing the scale or range of data into a range of 0 to 1. Normalization of data is required because microarray data has a significant difference in range. The data normalization function is shown in Equation 1, where y' is the value of features in the domain of normalization, y is the value of the data before the process of normalization, while $y_{\min}$ and $y_{\max}$ respectively declare the smallest value and the largest value of all data in an attribute to be normalized:

$$y' = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \tag{1}$$

Some cancer data from Kent-Ridge Biomedical Data Repository have been divided into training and testing data. Some others are manually divided into training and testing data, with proportion 70% as testing data and 30% as testing data.

Dimension Reduction

The dimensions and complexity of microarray data are very large. Therefore, a process that can reduce the complexity of microarray data is required. Complexity reduction aims at minimizing errors in the classification process. Dimensional reduction-a form of complexity reduction-is done using a Principal Component Analysis (PCA) algorithm. Applying dimensional reduction with PCA will reduce dimensional complexity because the microarray data will extract its features using eigenvectors and eigenvalues that have been obtained. The steps for dimensional reduction algorithm using PCA, according to Astuti (2018), are described below:

1. Let X be an input matrix for PCA. X is training data composed of a n-vector with data dimension m
2. Calculate the mean data of each dimension $\left(\bar{X}\right)$ using Equation 2:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

Where:
n = Number of samples or number of observation data
X_i = Observation data

3. Calculate the covariance matrix (C_X) using Equation (3):

$$C_X = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T \tag{3}$$

Where:
n = Number of samples or number of observation data
X_i = Observation data
$\bar{X}$ = *mean* data

4. Calculate the eigenvectors (v_m) and eigenvalues (λ_m) of the covariance matrix using Equation (4):

$$C_X v_m = \lambda_m v_m \tag{4}$$

5. Sort the eigenvalues in descending order
6. Principal Component (PC) is a collection eigenvector corresponding to the sorted eigenvalues in step 5
7. PC dimension will be reduced based on the eigenvalues

There are several ways for reducing PC dimension based on eigenvalues, such as:

a. Using a scree plot. The number of eigenvectors is selected based on the point of a curve that is no longer declining sharply
b. Using the cumulative proportion of variance (eigenvalues) of the total variance (eigenvalues)

In this research, the number of eigenvectors was determined via the cumulative proportion of variance (eigenvalues). The Proportion of Variance (PPV) for each main component (eigenvector) is outlined by Equation (5):

$$PPV = \frac{\lambda_i}{\sum \lambda_i} \times 100\% \quad (5)$$

where, λ_i is eigenvalue.

After that, the number of eigenvectors was determined by comparing the threshold with the cumulative Proportion of Variance (PPV).

PC that has dimensionally been reduced $\left(\widehat{PC}\right)$ is a matrix consisting of selected k to the largest k eigenvalues so that they eigenvectors, where k eigenvectors are vectors corresponding meet Equation (6):

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \times 100\% > Threshold \quad (6)$$

Reduce testing data (Y) dimensions by multiplying testing data with $\widehat{PC}$, as per Equation (7):

$$Y' = Y \times PC \quad (7)$$

Classification

After reducing the dimensional complexity of data, the next step is the classification process. Classification is the main objective of this research. At this stage, the data was diagnosed (classified) based on whether or not they are affected by cancer. Two classification methods were used: Support Vector Machine and Multilayer Neural Network with Levenberg-Marquardt Backpropagation learning algorithm. The results of these two algorithms were then compared and analyzed based on accuracy and processing time.

Support Vector Machines

SVM is a linear classification that finds the best hyperplane separating between classes. In non-linear problems, SVM uses a kernel trick in the training data so that the dimension becomes widespread. Once the dimensions are customized, SVM will seek the optimal hyperplane that can separate a class from other classes. The process for finding the best hyperplane using SVM according to Campbell (2005) is detailed below:

1. Let $x_i \in \{x_1, x_2, \ldots, x_n\}$, where x_i is data consisting of m-attributes and target class $y_i \in \{+1, -1\}$
2. Assume that classes +1 and -1 can completely be separated by a hyperplane, as defined in Equation (8):

$$w.x + b = 0 \quad (8)$$

Then, from Equation (8), Equations (9) and (10) are obtained:

$$w.x + b \geq +1, \ for\, class + 1 \quad (9)$$

$$w.x + b \leq -1, \ for\, class - 1 \quad (10)$$

where, x is the input data, w is the normal plane and b is the position relative to the middle field coordinates.

3. SVM aims to find hyperplanes that maximize margins between two classes. Maximizing margins is a quadratic programming problem that is solved by finding the minimal point of Equation (11) with the constraint equation of Equation (12):

$$\min_w \frac{1}{2} \| w \|^2 \quad (11)$$

$$y_i \left(x_i w + b \right) - 1 \geq 0 \quad (12)$$

where, x_i is the i-th input data and y_i is the i-th data target.

The problem in quadratic programming can be solved using Lagrange Multipliers outlined in Equation (13):

$$L(w, b, \alpha) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i \left(\left(w x_i + b \right) - 1 \right) \right) \quad (13)$$

Taking the derivatives with respect to b and w and substituting them again into Equation (13), Equation (14) can be formed:

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i x_j \quad (14)$$

where, α_i is the weight (parameter obtained from the Lagrangian Multipliers), x_i and x_j is the *i*-th input data and *j*-th input data, while y_i is the *i*-th data target.

4. For making decisions, Equation (15) is used for linear equations while Equation (16) is used for non-linear equations:

$$f(x_d) = sign\left(\sum_{i=1}^{n}\alpha_i y_i (x_i, x_d) + b\right) \quad (15)$$

$$f(x_d) = sign\left(\sum_{i=1}^{n}\alpha_i y_i K(x_i, x_d) + b\right) \quad (16)$$

where, n is the number of support vectors and x_d is the test data and $K(x_i,x_d)$ is the kernel function used, as per Equation (17).
Linear Kernel:

$$K(\vec{x}_i, \vec{x}_d) = (\vec{x}_i, \vec{x}_d) \quad (17)$$

The radial basis function kernel (RBF) is outlined by equation (18):

$$K(\vec{x}_i, \vec{x}_d) = \exp\left(\frac{-|\vec{x}_i - \vec{x}_d|^2}{2\sigma^2}\right) \quad (18)$$

σ is a kernel parameter option for the RBF kernel function.

Levenberg-Marquardt Backpropagation Algorithm

The Backpropagation learning algorithm is an algorithm based on a Multi-Layer Perceptron that finds the optimal weight in the data classification process. In standard backpropagation, there are three stages for finding optimal weight, which are forward propagation, backward propagation and weight update. The weight update process depends on the parameter of the learning rate, where in the standard backpropagation, the value of the learning rate is always constant at each iteration. This has an impact on the slowness of the algorithm in achieving optimal convergence. It can also sometimes become stuck at the local optimal point. To overcome these problems, the Levenberg-Marquardt algorithm is used. The learning rate parameter of the Levenberg-Marquardt algorithm is no longer constant, but it adjusts the error value for each iteration based on the decay rate. The change in weight value that occurs in each iteration is influenced by error factor, learning rate and the Jacobian Matrix. The Pseudocode for the Levenberg-Marquardt Backpropagation algorithm is outlined in more detail in Wisesty *et al*. (2016) and Suratgar *et al*. (2005).

Results and Discussion

This section discusses the results of the data classification process for cancer. The data used were obtained from Kent-Ridge Biomedical Data Repository. Two classification process schemes were used i.e., PCA and SVM methods (using linear kernel and RBF), as well as PCA and Levenberg-Marquardt Backpropagation (LMBP).

Dimension Reduction Results Using PCA

The dimensional reduction process in PCA is based on the eigenvalues and eigenvectors obtained from the covariance matrix. Only some eigenvectors with the largest eigenvalues were selected. The number of eigenvectors was determined by comparing the threshold with the cumulative proportion of variance (PPV). Therefore, the threshold plays an important role in determining PPV.

Table 2 shows that the highest eigenvalue is 8.97. This value-when compared to the total eigenvalues-yields aproportion of eigenvalues of 17.38%. Of the 2001 attributes in the colon cancer datasets, the first 47 largest eigenvalues yielded 100% cumulative PPV ($\pm 10^{-5}$). This means that the eigenvectors corresponding to the 48th-2001st eigenvalues did not have sufficient information to determine the class.

The threshold value greatly affects the selection of eigenvectors to be used during data transformation. Transformed data are used in the classification process. In this research, an empirical test was conducted on the effect of threshold on the final result (accuracy). Kişi and Uncuoğlu (2005) revealed that the threshold on the proportion of cumulative variance for the selection of minimal eigenvectors was 80%. In this research, the threshold value used was 60%, 70%, 80%, 90% and 95%. This test was used to determine the extent to which the threshold influences the classification system performance.

The number of eigenvectors for the specified threshold can be seen in Table 3. Table 3 shows that 10 eigenvectors corresponding to the 10 largest eigenvalues yielded 60% cumulative PPV. Meanwhile, to achieve 95% cumulative PPV on the central nervous system data, 35 eigenvectors corresponding to the 35 largest eigenvalues were required.

Table 2: Results of PPV calculations on colon cancer data

Number of eigenvectors (PC)	Eigenvalues	Cumulative PPV (%)	Number of eigenvectors (PC)	Eigenvalues	Cumulative PPV (%)
1	8.973402	17.38383	25	0.354675	91.84916
2	6.566147	30.10418	26	0.345146	92.51780
3	5.959143	41.64861	27	0.300977	93.10087
4	4.130259	49.65000	28	0.289972	93.66262
5	3.293468	56.03031	29	0.283838	94.21249
6	2.242024	60.37370	30	0.27634	94.74783
7	2.187057	64.61061	31	0.271611	95.27401
8	1.721803	67.94619	32	0.249645	95.75764
9	1.480780	70.81485	33	0.225333	96.19417
10	1.207234	73.15358	34	0.218001	96.61650
11	1.113624	75.31096	35	0.200622	97.00515
12	1.015076	77.27743	36	0.187881	97.36913
13	0.877387	78.97716	37	0.184288	97.72614
14	0.790535	80.50863	38	0.169082	98.05370
15	0.721548	81.90646	39	0.154449	98.35291
16	0.702840	83.26805	40	0.142972	98.62988
17	0.631187	84.49082	41	0.129604	98.88096
18	0.579217	85.61292	42	0.127175	99.12733
19	0.569173	86.71556	43	0.120994	99.36173
20	0.561455	87.80324	44	0.110364	99.57553
21	0.502843	88.77738	45	0.091832	99.75344
22	0.443881	89.63729	46	0.074693	99.89814
23	0.395773	90.40401	47	0.052581	100
24	0.391299	91.16206			

Table 3: Results of PPV calculations on colon cancer data

	Number of eigenvectors (PC)				
Data	60%	70%	80%	90%	95%
Central Nervous System	10	14	19	28	35
Colon	6	9	14	23	31
Leukemia	12	19	27	39	46
Ovarian	3	4	7	16	31
Lung	30	46	67	98	118
Prostate	4	9	17	32	52

Table 4: Result of tests for PCA threshold value on system accuracy using the SVM classifier

	Accuracy (%)					
Threshold (%)	Leukemia	Ovarian	Central nervous system	Colon	Lung	Prostate
60	100.00	94.74	93.33	85.71	90.63	88.000
70	100.00	94.74	80.00	85.71	90.63	97.060
80	93.33	98.25	66.67	85.71	93.75	47.960
90	100.00	100.00	73.33	85.71	84.38	41.180
95	100.00	100.00	80.00	85.71	84.38	76.470
Average accuracy	98.67	97.55	78.67	85.71	88.75	70.134

Table 5: Result of tests for PCA threshold value on system accuracy using the LMBP classifier

	Accuracy (%)					
Threshold (%)	Leukemia	Ovarian	Central nervous system	Colon	Lung	Prostate
60	93.330	98.250	80.000	78.570	96.880	97.060
70	93.330	96.490	73.330	78.570	100.000	97.060
80	93.330	100.000	80.000	92.860	93.750	94.120
90	100.000	100.000	86.670	78.570	96.880	100.000
95	86.670	100.000	86.670	78.570	93.750	100.000
Average accuracy	93.332	98.948	81.334	81.428	96.252	97.648

Table 4 and 5 show the results of empirical tests on the threshold value of each cancer testing data on system accuracy using SVM and LMBP classification methods. Accuracy was used as performance measure in this research. It was because, the researched data in general had balanced number of data for each class. The test result shows that there were changes in the result (accuracy) for each threshold value (60-95%) used for each cancer data. Therefore, to measure the performance of the classification method in each dataset, we also measured the average accuracy produced for each specified threshold. The analysis of each data based on the threshold, according to Table 4, showed that:

1. Larger threshold values do not always improve the result (accuracy), e.g., in leukemia data. The selection of threshold value depends on the characteristics of the data used. This is because useless data (data that is not informative) is combined in the data generated from the dimensional reduction process. The useless data has a relatively small eigenvalue (variance), so it may interfere with other informative data and affect the classification result
2. Classification using the SVM method yielded higher accuracy than that of the LMBP method. However, the accuracy produced by SVM is unstable, as seen in the accuracy results of prostate cancer, which had the smallest accuracy (up to 41.18%). This is different from the LMBP method, where the classification result was rather stable with the lowest accuracy of 73.33% for the Central Nervous System cancer data. In addition, the average accuracy of most LMBP methods achieved higher accuracy than the SVM method. This is because the LMBP method produced a separator function to classify data in general and avoid a local optimum. Therefore, if presented with new data (testing data), the LMBP method would be able to classify the new data better than the SVM method

Results of Testing Different Types of Kernel SVMs

SVM is a linear classification method that finds the best hyperplane that separates classes. In non-linear problems, SVM is combined with a kernel trick in the training data so that the dimension becomes wider. Campbell (2005) stated that once the dimensions were customized, SVM would seek an optimal hyperplane that could separate a class from the others. This research used three different kernel functions: Linear kernel, polynomial and Radial Basis Function (RBF). Based on observations in Table 6, it can be concluded that for some types of cancer (except lung cancer), the highest accuracy was achieved when using the RBF kernel (reaching 100% for leukemia and ovarian cancer data). Meanwhile, for lung cancer data, the highest accuracy (93.75%) was obtained when using a linear kernel. This is because the data from the lung cancer class-after going through dimensional reduction-could be separated linearly. However, polynomial kernels yielded accuracy results that were not as high as the linear kernels and RBF. After going through the dimension reduction process, microarray data were largely linearly separated with an average accuracy of 92.26%. If the transformation were done using the RBF kernel, the accuracy was found to increase to up to 94.46%. However, if the dimension was changed using the polynomial kernel, the data became difficult to classify by its class.

Test Results of Kernel Option Parameters in SVM

SVM is an advantageous algorithm for solving problems with high dimensions. However, it also has weaknesses, for example, in introduces complexity in selecting optimal parameter values. Therefore, SVM is very sensitive to the selection of parameters to be used. Some studies have performed trial-and-error methodsfor optimal parameter selection. For example, Harafani and Wahono (2015) tried a combination of different parameter values and then tested these on validated data to generate optimal parameters. The SVM parameters used in this research are listed below:

1. Parameter *C* controls the relationship between slack and margin variables. The larger the value of *C*, the greater the offense imposed on each classification. In this research, the parameter value *C* is 1000. The selection of this value was based on experiments conducted by Jia *et al*. (2014)
2. Parameter σ as a kernel input option in RBF kernel function

Table 6: Result of testing various types of SVM kernels on system accuracy

	Accuracy (%)						
SVM Kernel	Leukemia	Ovarian	Central nervous system	Colon	Lung	Prostate	Average accuracy (%)
Linear	100.00	100.00	80.00	85.71	93.75	94.11	92.26
Polynomial	73.33	92.98	80.00	85.71	75.00	91.18	83.03
RBF	100.00	100.00	93.33	85.71	90.63	97.06	94.46

In this research, kernel option parameter tested as an input parameter in the RBF kernel function on SVM method was done via empirical testing. The values of kernel option parameters tested were 9, 18 and 27. Empirical testing aims at finding the effect of these parameters on the cancer classification results. In Table 7, it can be seen that the greater the value of the kernel option used, the majority of accuracy of classification also increased (reaching 100% in Ovarian cancer data using kernel value option 27).

Test Result of Hidden Neurons in LMBP

The Levenberg-Marquardt algorithm is a Backpropagation optimization algorithm for finding optimal weights in a Neural Network. The Neural Network Architecture used in this research is a Multi-Layer Perceptron with one hidden layer. In the Multi-Layer Perceptron, it is difficult to determine the number of neurons in a hidden layer that can produce optimal performance. Therefore, in this research, empirical tests were conducted to obtain the optimal number of hidden neurons. The number of hidden neurons tested was 5, 10, 15, 20 and 25. Based on the test results in Table 8, it can be concluded that high accuracy (average 91.59% accuracy) could be achieved using hidden neurons of many as 5 hidden neurons. This suggests that classification of microarray data that has gone through a dimensional reduction process using PCA does not require a complicated separator function. However, unfortunately, the same cannot be said for the colon cancer data. In the colon cancer data, the best accuracy (92.86%) was only achieved when the number of neurons reached 25. Based on the overall results, it can be concluded that the more the number of neurons used, the more stable the resulting neural network model in classifying microarray data. However, the more number of hidden neurons used, the more time spent on the training process. The best average accuracy was therefore 92.68% with 15 neurons.

Comparison of Accuracy between the SVM and LMBP Methods

Several tests have been conducted in this research, including the PCA threshold parameters, kernel option and kernel option parameters in SVM and the hidden neuron number parameter in LMBP. Based on the accuracy results in Table 9, it can be observed that the result of classification using the LMBP method was more stable than SVM. The LMBP method achieved the best accuracy rate of 96.07%, while SVM achieved 94.98%. This is because LMBP can better generalize new data using the model obtained in the testing process compared to SVM on microarray data.

Table 7: Results of kernel option testing on SVM on system accuracy

Kernel option	Accuracy (%)						Average accuracy (%)
	Leukemia	Ovarian	Central nervous system	Colon	Lung	Prostate	
9	26.67	96.49	66.67	78.57	50.00	26.47	57.48
18	80.00	100.00	80.00	78.57	90.63	38.23	77.90
27	86.67	100.00	80.00	78.57	84.37	76.47	84.35

Table 8: Test result of neuron number on system accuracy

Number of hidden neurons	Accuracy (%)						Average accuracy (%)
	Leukemia	Ovarian	Central nervous system	Colon	Lung	Prostate	
5	93.33	100.00	86.67	78.57	96.88	94.12	91.59
10	93.33	100.00	80.00	78.57	90.63	94.12	89.44
15	100.00	100.00	86.67	78.57	93.75	97.06	92.68
20	86.67	100.00	80.00	78.57	90.63	100.00	89.31
25	93.33	100.00	80.00	92.86	93.75	94.12	92.34

Table 9: Comparison of accuracy of classification result of SVM and LMBP methods

Classification Method	Accuracy (%)						Average accuracy (%)
	Leukemia	Ovarian	Central nervous system	Colon	Lung	Prostate	
SVM	100.00	100.00	93.33	85.71	93.75	97.06	94.98
LMBP	100.00	100.00	86.67	92.86	96.88	100.00	96.07

Conclusion

This paper proposed a cancer detection scheme based on microarray data classification using a Principal Component Analysis (PCA) dimension reduction method and a comparison of two classification methods i.e., Support Vector Machine (SVM) and Levenberg-Marquardt Backpropagation (LMBP) algorithm.

In the dimensional reduction process using PCA, the selection of the number of eigenvectors to be used was based on the calculation of the Proportion of Variance (PPV) for each eigenvector. Based on the tests that were done, the system shows relatively little error although there is a reduction in the data dimension; the resulting data of PCA reduction is not more than 2% of the actual data. Besides, a greater threshold value does not guarantee improved system accuracy, due to the increase in the number of PC that can decrease the performance of classification system.

In testing the SVM classification method, the most influential parameter is the type of kernel used. In most of the results of cancer data classification, the use of linear kernel produced fairly high accuracy with an average accuracy of 92.26%. The average accuracy increased to 94.46%, when the reduced data was transformed again using the RBF kernel.

Based on the results of testing the parameters of the number of neurons in the LMBP method, it can be concluded that the more the number of neurons used, the more stable the generated neural network model when classifying microarray data. However, the more hidden neurons used, the longer the time spent on the training process. The best average accuracy was 92.68% when the number of neurons was 15. The result of classification using the LMBP method was found to be more stable than SVM. The LMBP method achieved the best accuracy rate of 96.07%, while SVM achieved 94.98% accuracy. This is because the LMBP algorithm can generalize new data using the model obtained in the testing process better than SVM on microarray data.

In this research, various LMBP neural network architectures are used to find out the best system performance. Therefore, selecting architecture with more structured method still needs to be done.

Acknowledgement

The authors would like to thank the Ministry of Research, Technology and Higher Education, Republic of Indonesia for financially supporting this research.

Author's Contributions

Adiwijaya: Design the research plan (formulation of overarching research goals and aims), validation of algorithm and preparation of article.

Untari N. Wisesty: Design and develop of methodology. Verification of the overall of experiments outputs. Conduct a research and investigation process, specifically performing the experiment.

E. Lisnawati: Implement the computer code, support algorithm and test of existing code components.

A. Aditsania: Synthesize study data statistically. Prepare the published work, specifically visualization experiment outputs

Dana S. Kusumo: Conduct the research and investigation process, specifically performing the experiment and verification of the overall experiment outputs.

Ethics

This paper is original and has not been published elsewhere. The authors assure that there are no ethical issues that may arise after the publication of this manuscript.

References

Adiwijaya, U.N. Wisesty and F. Nhita, 2014. Study of line search techniques on the modified back propagation for forecasting of weather data in Indonesia. Far East J. Math. Sci., 86: 139.

Aydadenta, H. and Adiwijaya, 2018. A clustering approach for feature selection in microarray data classification using random forest. J. Inform. Process. Syst.

Campbell, C., 2005. Support vector machine and kernel methods. Note Lecture of Bristol University.

Furey, T.S., N. Cristianini, N. Duffy, D.W. Bednarski and M. Schummer *et al*., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16: 906-914. DOI: 10.1093/bioinformatics/16.10.906

Harafani, H. and R.S. Wahono, 2015. Optimasi parameter pada metode support vector machine berbasis algoritma genetikauntuk estimasi kebakaran hutan. J. Intell. Syst., 1: 82-90.

Jia, D., D. Zhang and N. Li, 2014. Pulse waveform classification using support vector machine with Gaussian time warp edit distance kernel. Comput. Math. Methods Med., 2014: 947254-947254. DOI: 10.1155/2014/947254

Kişi, Ö. and E. Uncuoğlu, 2005. Comparison of three back-propagation training algorithms for two case studies. Ind. J. Eng. Mater. Sci., 12; 434-442.

Li, J., 2013. Kent-ridge bio-medical data set repository.

Nurfalah, A., Adiwijaya and A.A. Suryani, 2016. Cancer detection based on microarray data classification using PCA and modified back propagation. Far East J. Electr. Commun., 16: 269-281. DOI: 10.17654/EC016020269

Pirooznia, M., J.Y. Yang, M.Q. Yang and Y. Deng, 2008. A comparative study of different machine learning methods on microarray gene expression data. BMC Genom., 9: S13-S13. DOI: 10.1186/1471-2164-9-S1-S13

Seeja, K.R., 2011. Microarray data classification using support vector machine. Int. J. Biometr. Bioinform., 5: 10-10.

Siang, T.C., T.W. Soon, S. Kasim, M.S. Mohamad and C.W. Howe *et al*., 2015. A review of cancer classification software for gene expression data. Int. J. Bio. Sci. Bio. Technol., 7: 89-108. DOI: 10.14257/ijbsbt.2015.7.4.10

Suratgar, A.A., M.B. Tavakoli and A. Hoseinabadi, 2005. Modified levenberg-marquardt method for neural networks training. World Acad. Sci. Eng. Technol., 6: 46-48.

Astuti, W., Adiwijaya, 2018. Support vector machine and principal component analysis for microarray data classification. J. Physics: Conference Series, 971: 012003.

Vanitha, C.D.A., D. Devaraj and M. Venkatesulu, 2015. Gene expression data classification using support vector machine and mutual information-based gene selection. Proc. Comput. Sci., 47: 13-21. DOI: 10.1016/j.procs.2015.03.178

WHO, 2018. Cancer. World Health Organization. http://www.who.int/mediacentre/factsheets/fs297/en/

Wisesty, U.N., Adiwijaya and J. Nasri, 2016. Modified backpropagation algorithm for polycystic ovary syndrome detection based on ultrasound images. Proceedings of the 2nd International Conference on Soft Computing and Data Mining, Aug. 18-20, Springer, Bandung, Indonesia, pp: 141-151. DOI: 10.1007/978-3-319-51281-5_15