Far East Journal of Electronics and Communications

Published Online: April 2016
http://dx.doi.org/10.17654/EC016020269
Volume 16, Number 2, 2016, Pages 269-281 **ISSN: 0973-7006**

CANCER DETECTION BASED ON MICROARRAY DATA CLASSIFICATION USING PCA AND MODIFIED BACK PROPAGATION

Adiyasa Nurfalah, Adiwijaya and Arie Ardiyanti Suryani

School of Computing
Telkom University
Jl. Telekomunikasi no. 1
Bandung 40257, Indonesia
e-mail: adiyasa.nurfalah@gmail.com
adiwijaya@telkomuniversity.ac.id
ardiyanti@telkomuniversity.ac.id

Abstract

According to the data from the World Health Organization (WHO) in 2012, cancer is considered as the leading cause of death in the world. About 8.2 million people died because of cancer and the number is estimated to increase each year due to an unhealthy lifestyle [15]. Deaths due to cancer could be prevented if it is detected early. In recent decades, microarray has taken an important role in cancer research. Microarray is a technology that is capable of storing thousands of gene expressions taken from several specific tissues of human at once. By analyzing microarray data, it can be known whether the tissues are affected by cancer or not. This study provides a fast and accurate framework for cancer detection based on microarray data classification using principal component analysis (PCA) and modified back propagation (MBP). MBP is a modification of standard

Received: September 3, 2015; Revised: September 20, 2015; Accepted: October 12, 2015
Keywords and phrases: cancer detection, microarray data, modified back propagation, principal component analysis.

back propagation (BP) that implements conjugate gradient algorithm on search direction in BP training. The experiment results show that the proposed system (MBPorPCA+MBP) is able to outperform BP-based system (BP or PCA+BP) in accuracy and especially in training time.

1. Introduction

Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. Cancer cells can spread to other parts of the body through the blood and lymph systems [12]. According to data reported by the World Health Organization (WHO), cancer is the leading cause of death worldwide, which is about 8.2 million deaths in 2012 and estimated to increase each year due to an unhealthy lifestyle [15].

In recent decades, microarray data takes an important role in the diagnosis of cancer and in order to improve the accuracy of cancer diagnosis compared to traditional techniques. Microarray can be used to see the level of gene expression in specific cell samples as to analyze thousands of genes simultaneously [14]. The data itself has small samples but the dimension is very large. Consequently, there was a challenge for researcher to provide solutions, i.e. to generate microarray data classification that has a high performance both in accuracy and running time. In classification study, Artificial Neural Network (ANN) is one of popular methods that gives satisfactory result. There are several algorithms to train ANN, one of the most popular algorithms is back propagation (BP). ANN trained by BP has successfully solved various classification problems that gives good accuracy.

Although BP is a suitable algorithm for ANN training, it still has several major deficiencies in BP such as [1]: First, the BP algorithm will get trapped in local minima, it can lead to failure in finding a global optimal solution. Second, the convergence rate of BP is still too slow even if learning can be achieved. Third, the convergence behavior of BP depends on choices of learning rate in advance. Many improvements have been made to improve the performance of BP, one of them is to modify BP based on conjugate gradient [3, 12]. By modifying BP with conjugate gradient, search direction

is performed not only along descent direction, as standard BP, but along conjugate directions, which produces generally faster convergence than BP [12]. The study proposes an alternative framework for cancer detection based on microarray data classification using combination of principal component analysis (PCA) and modified back propagation (MBP).

2. Research Method

After the development conducted by Patrick O. Brown, Joseph DeRisi, David Botstein, and colleagues in the mid of 1990s, DNA microarrays have become a key tool in the fight against cancer [13]. The data itself has small samples but the dimension is very large. Since then, there is a challenge for researchers to provide solutions for microarray data classification that shows a high performance both in accuracy and running time. Figure 1 below illustrates of how microarray data obtained.

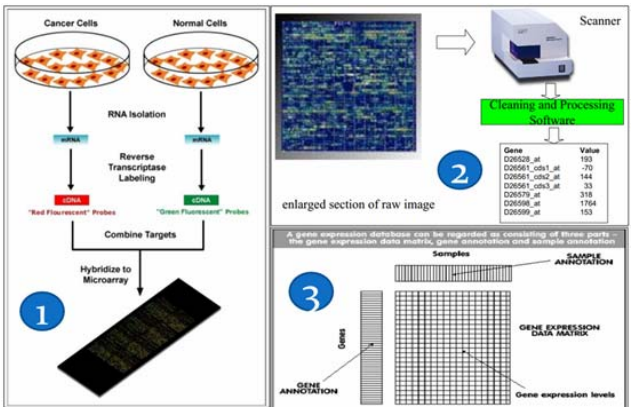

Figure 1. The processes to obtain microarray data [http://www.ebi.ac.uk].

In machine learning field, cancer detection problem can be done by classifying data into classes that have been defined. Classification is the process of determining a class of data using methods such as Artificial Neural Networks.

ANN trained by back propagation (BP) is a popular method for classification that produces good accuracy for microarray classification. The

followings are some literature review pertaining to the method. Bai [5] in her master thesis has built framework based on PCA and back propagation to classify microarray data. She used some public microarray dataset such as ovarian cancer data, colon cancer data, and leukemia data, resulted in 96% of accuracy of ovarian, 95% of accuracy for colon and 97% accuracy for leukemia. In terms of training time, Bai admitted that back propagation required long time for training, around 20-23 seconds for each data. Huynh et al. in [8] built a framework based on PCA and back propagation to classify microarray data. PCA is done by implementing singular value decomposition (SVD) in this study. Experiment is provided by compared proposed system accuracy with extreme learning method (ELM) and proposed system was able to outperform ELM with 83.6% accuracy for colon cancer data.

As has been mentioned in Introduction, although BP is a good algorithm for train ANN, BP requires a lengthy time in the training. This study tried to improve the accuracy of standard back propagation by modifying back propagation by implementing conjugate gradient (CG) algorithm on search direction in BP training. BP algorithm uses the steepest descent to calculate search direction of new weights and static learning rate as step-size of direction. The steepest descent uses the most negative gradient to be the search direction. This is the direction in which the performance function is decreasing most rapidly. It turns out that, although the function decreases most rapidly along the negative gradient, it does not necessarily produce the fastest convergence [12]. By modifying search direction of standard back propagation using conjugate gradient method, the search direction does not only decrease but also conjugates directions [2, 9].

2.1. The modified back propagation (MBP) algorithm

Conjugate gradient (CG) uses non-zero vector which is orthogonal and linearly independent [8]. Two vectors d_i and d_j are orthogonal (G-conjugate) if their inner product is zero, it can be written as follows:

$$d_i^T d_j = 0. \tag{1}$$

Before getting into CG algorithm, the objective function must be determined so that it can be optimized. If the CG algorithm is used as an artificial neural network training algorithms, then the goal is to minimize the errors that depend on the weights that connect between neurons. Therefore, the objective function is the error function:

$$f(w) = \frac{1}{2N} \sum_n \sum_j (t_{nj} - y_{nj}(w))^2, \tag{2}$$

where N is the number of patterns in training data, w is weighting matrix, t_{nj} and $y_{nj}(w)$ are target data and the output neurons for n patterns, respectively.

Conjugate gradient is a method to minimize the derivative function by calculating iteratively w_{k+1} approach based on:

$$w_{k+1} = w_k + \alpha_k d_k,$$

$$d_{k+1} = -g_{k+1} + \beta_k d_k, \tag{3}$$

where α and β are the parameters of the momentum to avoid local convergence.

Conjugate gradient algorithm for training of back propagation is as follows [3]:

1. Initializing all weights to small random numbers.
2. If the termination conditions are not fulfilled, do the steps 2-12.

Forward propagation:

3. Calculating all the outputs in the hidden units y_j $(j = 1, 2, ..., p)$.

$$y\_net_j = v_{j0} \sum_{i=1}^{n} x_i v_{ji},$$

$$y_j = f(y\_net_j) \frac{1}{1 + e^{-y\_net_j}}. \tag{4}$$

4. Calculating all the outputs in the output units z_k $(k = 1, 2, ..., m)$.

$$z\_net_k = w_{k0} \sum_{j=1}^{p} y_j w_{kj},$$

$$z_k = f(z\_net_k) \frac{1}{1 + e^{-z\_net_k}}. \quad (5)$$

Backward Propagation:

5. Calculating the error factor in the output units based on difference (error) value of actual and factual values (output from the output units).

$$\delta_k = (t_k - z_k) f'(z_{net_k}) = (t_k - z_k) z_k (1 - z_k). \quad (6)$$

6. Calculating the error factor in the hidden units based on the error factor in the unit above it.

$$\delta\_net_j = \sum_{k=1}^{m} \delta_k w_{kj},$$

$$\delta_j = \delta_{net_j} f'(y_{net_j}) = \delta\_net_j y_j (1 - y_j). \quad (7)$$

7. Calculating the gradient at the unit output of the objective function.

$$g_{k+1} = \frac{1}{N} \sum_{n=1}^{m} \delta_{nk} z_{nk}. \quad (8)$$

8. Calculating the gradient at the hidden units.

$$g_{j+1} = \frac{1}{N} \sum_{n=1}^{p} \delta_{nj} y_{nj}. \quad (9)$$

9. Calculating the parameter β for all neurons in the hidden units and output units as follows:

a. Powell-Beale: $\beta_{k+1} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{d_k^T (g_{k+1} - g_k)}$,

b. Fletcher-Reeves: $\beta_{k+1} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$,

c. Polak-Ribiere: $\beta_{k+1} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T g_k}$, (10)

where: $\beta_{k+1} = \beta$ in the recent iteration.

g_{k+1} = gradient in the recent iteration.

g_k = gradient in the prior iteration.

10. Calculating the direction for all neurons in the hidden units and outputs units.

$$d_{t+1} = -g_{t+1} + \beta_t d_t. \quad (11)$$

For the initial direction: $d_1 = -g_1$.

11. Calculating the parameter α for all neurons in hidden units and output units, i.e. how big the steps taken for each direction. This parameter can be searched by line search technique [2].

Weights update:

12. Weight updating is carried out using the following equation:

$$w_{t+1} = w_t + \alpha_{t+1} d_{t+1}. \quad (12)$$

2.2. Using PCA for reducing dimension

Microarray data has very high dimension data that may cause “the curse of dimensionality” problem if we classify it without preprocessing. In this research PCA dimension reduction process is added before classification. PCA can transform the high dimensional data into a new coordinate system that is generated from linear combination of original data. PCA is obtained from the calculation of eigenvector and eigenvalue of covariance matrix. Only a number of the eigenvector with the largest eigenvalue is selected as a dimension reduction model (called principal component/PC). After some PC

is obtained, original data is projected into these PC to get transformed data with reduced dimension. PCA is used for reduction dimension.

Let data be matrix $X(m, n)$, each m-sample in the matrix X is represented as a vector sized N. Matrix X is the input of PCA. Below are the steps to reduce dimension using PCA [7].

1. Calculating mean of data

$$\mu = \frac{1}{N}\sum_{k=1}^{N} X_k, \tag{13}$$

where N = number of sample and X = number of dimension.

2. Calculating covariance matrix

$$C = \sum_{k=1}^{N}(X_k - \mu)(X_k - \mu)^T. \tag{14}$$

3. Calculating eigenvalue and eigenvector

$$CU_n = \lambda_n U_n, \tag{15}$$

where U = eigenvector and λ = eigenvalue.

4. Selecting set of eigenvector with the largest eigenvalue.

The simple way to do this; eigenvector is sorted descending by its eigenvalue and then some most-left eigenvector is selected randomly.

5. Data transformation

$$Y = U^T * (X - \mu). \tag{16}$$

2.3. Design of PCA+MBP

The proposed system is divided into two stages, that are training stage and testing stage. Training stage is done first to build a system model. After training stage is done, the system can be used to classify data in testing stage. Detailed process is described in Figure 2.

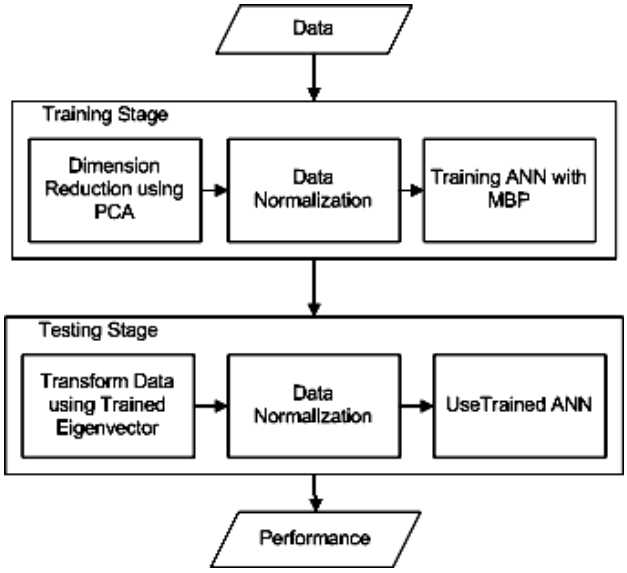

Figure 2. Block diagram of PCA+MBP system.

The results of training stage are PCA model and trained ANN. In the testing stage PCA model is used to transform data and get data with reduced dimension and then propagation is conducted by using trained ANN. Data normalization is aimed at making uniform input of ANN in range value (0, 1).

3. Result and Analysis

Three public microarray data were in use, which were leukemia data, colon cancer data, and ovarian cancer data, taken from Kent-Ridge Bio-medical Data Set Repository [10] for developing and evaluating model. Ovarian data contains proteomic patterns intensity in serum that distinguishes ovarian cancer from non-cancer. Colon data was collected from colon cancer patients that suffer from tumor biopsies and normal biopsies of colon cancer. Leukemia data was collected from the bone marrow samples of leukemia patients to classify the types of leukemia which were acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Table 1 is the specification of all data.

Table 1. Data specification

Data	Number of Classes	Number of Samples	Number of Features
Colon	2	65 (25 cancer vs 40 normal)	2000
Ovarian	2	250 (160 cancer vs 90 normal)	15154
Leukemia	2	70 (45 ALL vs 25 AML)	7129

Some systems were also run, which were: MBP (without PCA), PCA+BP, and BP to compare the results with proposed system result. 5-folds cross validation was used for model selection and evaluation. Training time (running time of all processes for building 1 model in *k*-fold cross validation) was investigated in second and accuracy in percent (%). Table 2 is the best result of all systems, bold values specify the best value values of measurement for each data.

Table 2. Testing results

	System	Accuracy	Training Time
Ovarian	PCA+BP	98	45.55
	MBP	**100**	7.87
	PCA+MBP	96	**5.35**
	BP	81.2	373.2
Colon	PCA+BP	**83.08**	15.74
	MBP	**83.08**	3.07
	PCA+MBP	76.92	**2.28**
	BP	75.38	60.9
Leukemia	PCA+BP	94.29	4.07
	MBP	69.05	5.44
	PCA+MBP	**97.14**	**1.81**
	BP	64.29	600.52

Table 2 shows that PCA+MBP (proposed system) requires the shortest training time compared to other systems, around 1-5 second for building model. In terms of accuracy, PCA+MBP only wins for leukemia data. Another finding in experiments is that the accuracy of MBP is superior for ovarian data and colon data compared with other systems. The superiority of MBP is unexpected considering that there is no dimension reduction in MBP.

4. Conclusion

This paper proposes a new framework for detecting cancer based on microarray data using the combination of principal component analysis (PCA) and conjugate gradient back propagation (MBP) method. The MBP is the modification of standard back propagation by implementing conjugate gradient algorithm in back propagation training. Some experiments were conducted by implementing 5-fold cross validation using public data from Kent-Ridge Bio-medical Data Set Repository, consisting of ovarian cancer data, colon cancer data, and leukemia data. The accuracy of testing stage and time of training stage were determined as measurement tools for measuring performance of PCA+MBP (proposed system), MBP, PCA+BP, and BP.

The important result of the experiments is that PCA+MBP requires around 1-5 seconds for training of each data. It is the best result of all systems. Furthermore, the MBP also still generates shorter training time than PCA+BP and BP, it proves that MBP is able to reduce training time in neural network. In terms of accuracy of testing, generally PCA+MBP produces around 76%-97% for all data. This result is not able to outperform PCA+BP that produces accuracy around 83%-98%. The outstanding result is showed by MBP, the accuracy of MBP is higher than PCA+MBP, PCA+BP, and BP in ovarian cancer data and colon cancer data that are 100% and 83.03%. Testing results show that PCA+MBP requires very short time for training, but in terms of accuracy, PCA+MBP must be improved. In the next research, we would like to propose a classification technique based on *f*-coloring [4] or implementation on mobile device using evolving neural network as illustrated in [6].

Acknowledgments

Authors would like to thank Setia Pramana, Ph.D. from Medical Epidemiology and Biostatistics Department, Karolinska Institutet, and Putri Wiki Novianti, Ph.D. from Utrech University, for discussion and their presentation in indo data mining mailing list.

Authors also thank the anonymous referees for their valuable suggestions which led to the improvement of the manuscript.

References

[1] Abbas Y. Al Bayati, Najmaddin A. Sulaiman and Gulnar W. Sadiq, A modified conjugate gradient formula for back propagation neural network algorithm, J. Comp. Sci. 5(11) (2009), 849-856.

[2] Adiwijaya, U. N. Wisesty and F. Nhita, Study of line search techniques on the modified backpropagation for forecasting of weather data in Indonesia, Far East J. Math. Sci. (FJMS) 86(2) (2014), 139-148.

[3] Adiwijaya, U. N. Wisesty, T. A. B. Wirayuda, Z. K. A. Baizal and U. Haryoko, An improvement of backpropagation performance by using conjugate gradient on forecasting of air temperature and humidity in Indonesia, Far East J. Math. Sci. (FJMS) Special Volume, Part I (2013), 57-67.

[4] Adiwijaya, A. N. M. Salman, O. Serra, D. Suprijanto and E. T. Baskoro, Some graphs in $Cf2$ based on *f*-coloring, Inter. J. Pure Appl. Math 102(2) (2015), 201-207

[5] Anita Bai, Classification and clustering using intelligent techniques: application to microarray cancer data, Master Thesis, Department of Computer Science and Engineering, National Institute of Technology Rourkela, India, 2013.

[6] F. Nhita, Adiwijaya, U. N. Wisesty and I. Ummah, Planting calendar forecasting system using evolving neural network, Far East J. Electron. Comm. 14(2) (2015), 81-92.

[7] F. R. Umbara, A. Nurfalah and T. H. Liong, Colorectal cancer classification using PCA and fisherface feature extraction data from pathology microscopic image, Information Systems International Conference (ISICO), 2013.

[8] Hieu Trung Huynh, Jung-Ja Kim and Yonggwan Won, Classification study on DNA microarray with feedforward neural network trained by singular value decomposition, Inter. J. Bio-Sci. Bio-Tech. 1(1) (2009), 17-24.

[9] J. Wang, X. Chi and T. X. Gu, Nonlinear conjugate gradient methods and their implementations by TAO on dawning 2000-II+, Inter. Conf. Parallel Algorithms and Computing Environments (ICPACE), 2003.

[10] Jinyan Li, Kent-Ridge Bio-medical Data Set Repository, School of Computer Engineering, Nanyang Technological University, Singapore.
Downloaded in January 2013 from:
URL: http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html.

[11] Kisi Ozgur and Uncuoglu Erdal, Comparison of three back propagation training algorithms for two case studies, Indian J. Engin. Mater. Sci. 12 (2005), 434-442.

[12] National Cancer Institute, What is cancer?
http://www.cancer.gov/cancertopics/-cancerlibrary/what-is-cancer.
Accessed on March 22nd, 2013.

[13] Nature Scitable, Genetic diagnosis: DNA microarrays and cancer.
http://www.nature.com/scitable/topicpage/genetic-diagnosis-dna-microarrays-and-cancer-1017. Accessed on February 26th, 2014.

[14] Ahmad M. Sarhan, Cancer classification based on microarray gene expression data using DCT and ANN, J. Theo. Appl. Inform. Tech. 6(2) (2009), 208-216.

[15] World Health Organization, Cancer fact sheet.
http://www.who.int/mediacentre/factsheets/fs297/en/index.html.
Accessed on March 22nd, 2013.