Classification of Microarray Data Involves Naïve Bayes and Dimension Reduction Using Haar Wavelet

Aniq A Rohmawati, Adiwijaya, Milah Sarmilah

Abstract: ***A general problem solving for handling microarray data is classification process with added a selection process from huge attributes. In particular, the escalated of attributes dimensionality provides a challenge to microarray handling techniques, related to microarray represents the large amount of genes expression. The multi-dependency (multicollinearity) may affect the performance when determining the parameter of classification. Many ways of solving the multicollinearity problem exists, the variable selection technique has become particularly popular. This is the method which use wavelet transformation for a few carefully selected variable and the method which regress respond variable onto a few linier combinations (components) of the original attributes. Wavelet is commonly used in image processing, spectral data using wavelet transformation have proved very successful in capturing the distinction among hyperspectral data. This paper investigates a new method of transformation data using Haar wavelet for selection processes. Our extensive study compares the selection processes using Haar wavelet transformation and Genetic Algorithm considering the selection dataset that implemented to Naïve Bayes classification. In addition, the selection-classification using Haar wavelet and Naïve Bayes describes a classification cancer and non-cancer quite well related to the accuracy of confusion matrix***

Keywords: ***Microarray, dimension reduction, Haar wavelet, Naïve Bayes.***

I. INTRODUCTION

Cancer is a disease which is caused by abnormal cell growth over human body. Based on a survey conducted by the World Health Organization (WHO) in early 2015. This is the second leading cause of death in the world and is responsible for 8.8 million deaths in 2016. Therefore, a technology to detect cancer early is required, in order to get the early handling with accurate results. However, the cancer detection is tough, as long as this detection conducted with microarray techniques. Microarray is a modern technique facilitating simulated analysis of the large amount of genes expression data required to solve complex biological problems.

Revised Manuscript Received on February 05, 2019.

Aniq A Rohmawati, Department of Computational Science, School of Computing, Telkom University, 40257, Bandung, Indonesia

Adiwijaya, Department of Computational Science, School of Computing, Telkom University, 40257, Bandung, Indonesia

Milah Sarmilah, Department of Computational Science, School of Computing, Telkom University, 40257, Bandung, Indonesia

The most problem of handling microarray is grouping a given sample into one of the subclasses of the type of disease, in which the subclasses of a disease are established. Then, microarray data classification requires more effort because of the huge of dataset dimension and complex relationships between various genes. Another important aspect of microarray processing is the phenomenon of finding information or even important information among genes (attributes). Then, the dataset processing becomes less effective and inefficient caused the unstable computational load [1]. Consequently, a scheme is required to handle unfavorable detection. Hence, a solution that presented from previous research is dimension reduction and classification processes for handling microarray dataset.

The corresponding dimension reduction for microarray data prior to the classification processes, aims to relieve the computational load of the classification given by [2].According to previous research, they proposed the Principal Component Analysis (PCA) and Modified Back Propagation (MBP) as selection – classification processes of microarray data [3]. We noted the very well result that PCA-BP produce accuracy about 83%. Then, Wavelet decomposition of Surface Electromyography (sEMG) signals based on Discrete Wavelet Transformation (DWT) and Surface Electromyography (sEMG), provided feature classification extraction with outstanding accuracy [4][5]. The essential problem related to microarray data is the multi-dependency between genes (attributes). One of the reasons is that the number of available records is often much smaller than the number of attributes. Many ways of solving the multi-dependency problem exist, the related research of wavelet estimator behavior in nonparametric regression has been performed by the previous research recently. They handled large dimension of FTIR percent transmittance data using Haar wavelet and Partial Least Square (PLS) [6]. The result of dimension reduction has significant result to estimate parameters of PLS calibration model than those of a baseline study over Principal Component Analysis (PCA) [7]. Some researchers computed the Naïve Bayes classification engaging with Aspect-Based Sentiment Analysis with high final accuracy [8]. We purpose to present intact result of the classification process, this objective to classify cancer or non-cancer data adopt the Haar wavelet selection processes and Naïve Bayes classification.

Also, a numerical analysis is carried out to demonstrate how the selection-classification techniques can be implemented. In this paper, we also propose comparative analysis of Naïve Bayes classification by comparing preprocessing between Haar wavelet and GA as selection methods.

II. DATA AND SYSTEM DESIGN

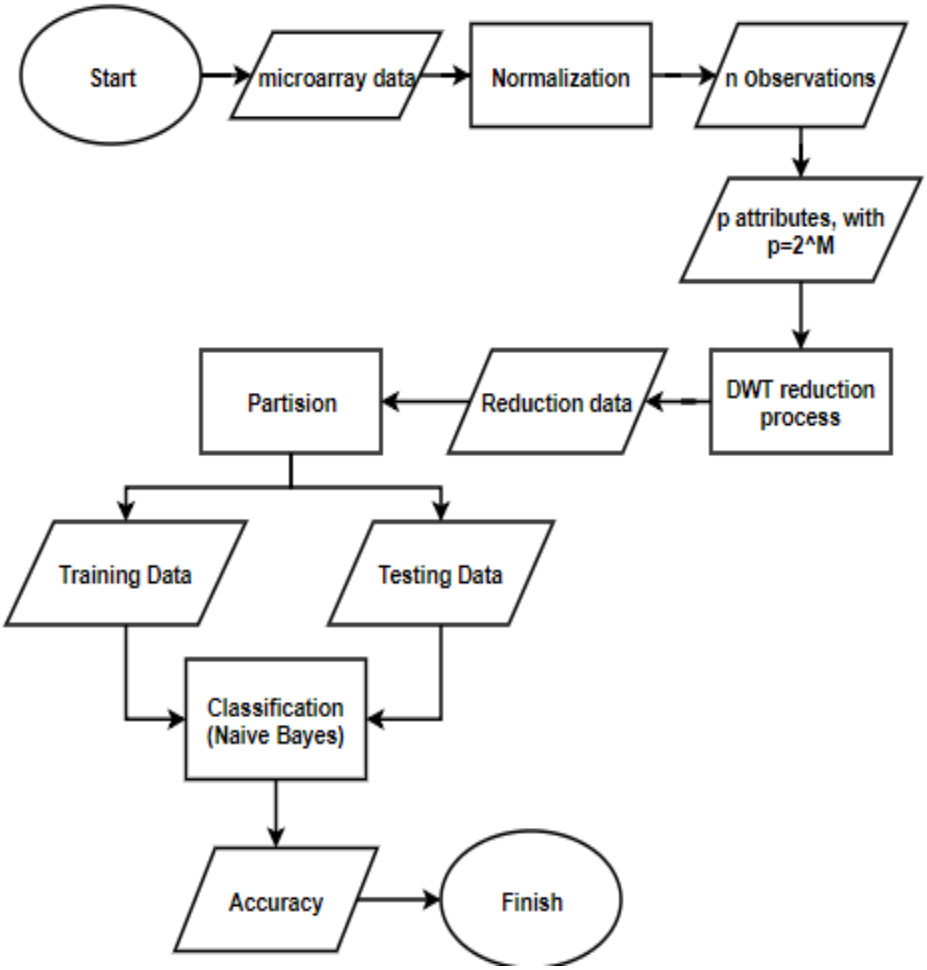

Fig. 1 System Design

Particularly, microarray data has a large dimension, the data is random variable obtained from Kent Ridge-Bio-Medical Dataset Repository without any additional information about records and attributes (classified) [9].

Table. 1 Dataset Microarray

Cancer Dataset	Number of Gen	Number of Class	Number of Sample	
Colon tumor	2000	2	40 Negative	22 Positive
Lung Cancer	12533	2	31 Mesothelioma	150 ADCA
Ovarian	15154	2	91 Negative	162 Positive

According to Table 2, the microarray dataset has a various range value of attributes, we have to normalize the dataset establishing a uniform range of values for each of attribute between 0 and 1. The general normalization formula as below,

$$y_i = \frac{x_i - min(X)}{max(X) - min(X)}$$

where y_i is normalization data and x_i as actual data (microarray). The objective of dimension reduction is finding informative genes and reducing the complexity of computing considering Haar wavelet of Discrete Wavelet Transformation (DWT). Thereafter selection dimension, we treated the selection data into testing and training dataset, with 70% and 30% proportion. The Naïve Bayes classification is implemented to determine of cancer and non-cancer class prediction. Overall, we construct two mainly processes: selection and classification, the entirely technique can figure out the design system as Fig. 1.

Haar Wavelet

The function $\psi(t)$ with real-value is considered a wavelet if it consists of [10],

1. Integral $\psi(t)$ *is* zero: $\int_{-\infty}^{\infty} \psi_{j,k}(t)\, dt = 0$
2. Integral $\psi^2(t)$ is one: $\int_{-\infty}^{\infty} \psi_{j,k}^2(t)\, dt = 1$

According to Sony and Notodiputro in Rohmawati, Haar wavelet is an orthogonal mother wavelet in which case the Haar-mother wavelet equation can be seen as below [6][7].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2 \\ -1, & 1/2 \le t < 1 \\ 0, & other \end{cases} \quad (1)$$

Note that a father waveletcan be written as,

$$\phi(t) = \begin{cases} 1, & 0 \le t < 1 \\ 0, & other \end{cases} \quad (2)$$

In particular, Discrete Wavelet Transform (DWT)describe a linear combination of the attributes as basic functions that called as wavelet function[6]. If vector data has size$p$, $p = 2^M$, $M$is a positive integer, then the vector may be expressed in function that rely on the interval [0.1).

$$f(t) = \sum_{k=0}^{2^M-1} x_k I_{\left\{k/2^M \le t < k+1/2^M\right\}}$$

Then, consider $f(t)$ can be decomposed into,

$$f(t) = c_{0,0}\phi(t) + \sum_{j=0}^{M-1}\sum_{k=0}^{2^j-1} d_{j,k}\,\psi_{j,k}(t) \quad (3)$$

The equation above is formally known as DWT.The $c_{0,0}$ is a smoothing coefficient function, where $d_{j,k}$ represents as wavelet coefficients, for each the value of the mother $\psi_{j,k}(t)$ and father wavelet $\phi(t)$ with t in different values, and $j$number is a level of resolution [6]. Then, we found $f(t)$representing all levels of decomposition. Those are the resolution of level 0 to $(M-1)$. According to the general equation, DWT formula can be expressas,

$$\underline{\boldsymbol{x}} = \boldsymbol{W^T}\underline{\boldsymbol{d}}$$

Let the following is Haar wavelet matrix if $\boldsymbol{p = 4}$,then we obtained $\boldsymbol{W^T}$as,

$$\boldsymbol{W^T} = \begin{bmatrix} 1/2 & 1/2 & 1/\sqrt{2} & 0 \\ 1/2 & 1/2 & -1/\sqrt{2} & 0 \\ 1/2 & -1/2 & 0 & 1/\sqrt{2} \\ 1/2 & -1/2 & 0 & -1/\sqrt{2} \end{bmatrix}$$

In particular,$\boldsymbol{W}$is orthogonal matrix, then wavelet coefficient may be calculated with,

$$\underline{\boldsymbol{d}} = \boldsymbol{W}\,\underline{\boldsymbol{x}}$$

where $\underline{\boldsymbol{d}} = \left(c_{0,0}\,, d_{0,0}, d_{1,1}, d_{1,0}, \dots, d_{n-1,0}\right)^T$.In general formula of DWT, $\boldsymbol{W}$ is called the transformation wavelet matrix, where $\boldsymbol{W}$ is the orthogonal for all types of mother wavelet that used[9]. As noted,$\boldsymbol{W}$is the matrix elements of the columns contained by$\phi(t)$and $\psi(t)$ for a variety of t ∈ [0.1), specify for Haar wavelet. The dimension reduction process is completed by taking $m < p$, by giving zero value in column $m+1$ until p [7].

$$\boldsymbol{D^*_{(nxm)}} = \boldsymbol{X_{(nxp)}}\boldsymbol{W^{*T}_{(pxm)}} \quad (4)$$

$\boldsymbol{D^*_{(nxm)}}$ is the selection result by implementing transformation Haarwavelet. Subsequently, we conducted the result of Haar wavelet to the Naïve Bayes classification process.

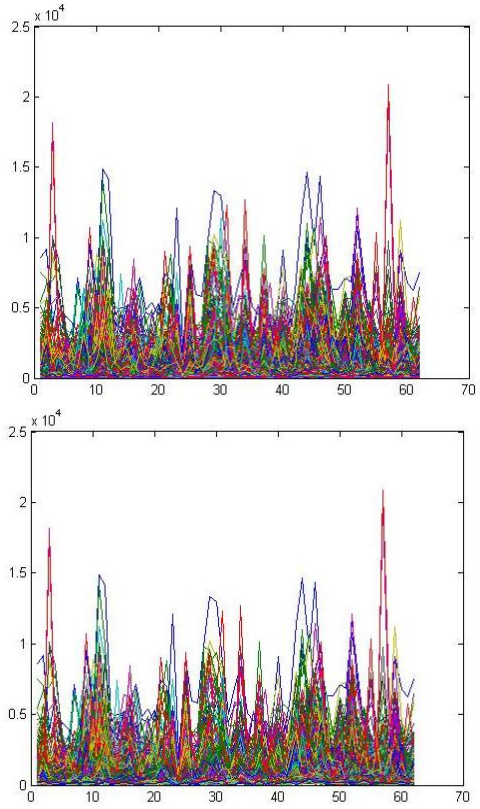

Fig. 2 Colon tumor 2000 (left) and 500 attributes after selection process (right)

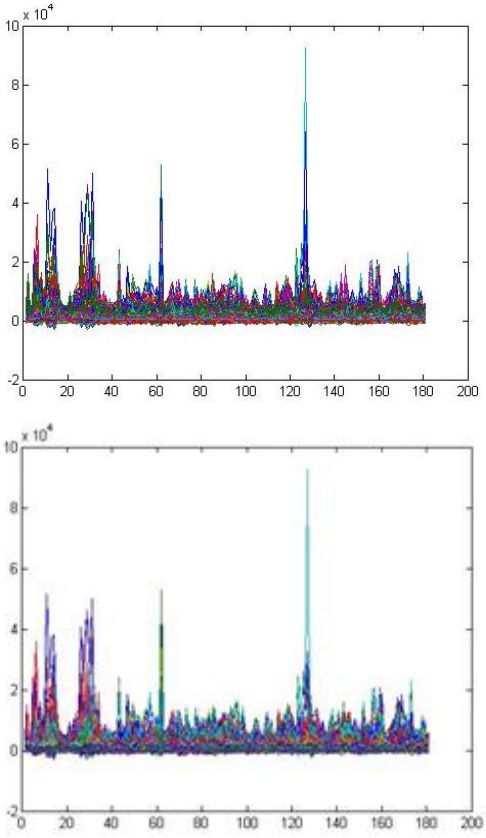

Fig. 3 Colon tumor 2000 (left) and 500 attributes after selection process (right)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

From Table 1, we are tested using colon tumor data with 2000 attributes and lung cancer with 12533 attributes. The wavelet method requires that the number of predictor variables must satisfy 2^M , the data may be reduced to 1024 attributes for colon tumor and 8192 attributes for lung cancer. From Figure 2 and 3, an *x*-axis and *y*-ordinate show the sample records and the value of each attribute. We execute selection dimension process trough by implemented Eq. 3 and Eq. 4, we can see in Fig. 2 and Fig. 3 that the selection process did not change trend dataset significantly from the actual dataset.

Naïve Bayes Classification

We propose Naïve Bayes for classification, Bayes classification assumes that a feature does not affect to other features. IfB is the input containing the feature and A is the class label, then Naïve Bayes is well written as$P\ (A|B)$ [12]. The notation means that the probability of class A label obtained after the observable Bfeatures. The Naïve Bayes equations for classification may be expressed as,

$$P(A|B_i) = \frac{\prod_{i=1}^{q} P(B_i|A)\ P(A)}{P(B_i)}$$

According to a Gaussian distribution are presented as a conditional continuous probability of features in a class $P(B_i|A)$. The Gaussian distribution is characterized by two parameters: mean (μ) and variance (σ^2), then a probability density function of Gaussian distribution is given by,

$$g_X(x_i;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Genetic Algorithm

One of the advanced algorithms for feature preference is the genetic algorithm. This is a stochastic approach based on the mechanism of natural genetics and biological evolution. A genetic algorithm can be implemented to optimize the performance of a prediction process, by selecting the most relevant features. We propose to compare some method for selection (reduction) dimension between Genetic Algorithm (GA) and Haar wavelet. The following steps for selection process using GA as below [11],

1. Each chromosome isthe attributes of dataset that consist of binary codes. Then, each of individualsin chromosome is represented by the binary number 0 or 1. Then, initialize a population based on the binary number 0 and 1 which randomly generated as many as the number of features and the size of the population. If a bit is equal to 1, then the feature is selected. We consider 500 individuals with 5 maximum generation and 100 population size,
2. The second, evaluation of fitness is obtained by the accuracy of algorithm Naïve Bayes used to evaluate individuals. The highest fitness values in training process is selected, to measure the classification accuracy considering the confusion matrix. After we keep the fitness evaluation of each individual, then thechromosome with best fitness value iskept, this process is known as elitism. Elitism leads us to the best solution considering fitness candidate,
3. Parents for each individual is selected by Roulette Wheel method, calculating each individual chromosome at the Roulette Wheel in accordance with the proportion of the fitness of each individual. The larger of fitness chromosome, the greater fitness proportion in the Roulette Wheel. Then, the individual has agreat chance to be the parents. Crossover of parent chromosome produce offspring (children) chromosome. This study conducted 0.8 single point crossover (Pc),
4. Mutation adoption offspring chromosome at random in a binary number with a particular mutation probability. A binary number has been raised to randomly and the criteria is less than the mutation probability (Pm), then that gene would be changed with a binary number instead (0 is converted into 1, and conversely). We consider 0.1 as Pm,
5. The laststep is survivor selection, there are two methods that may be used in the selection process of survivor i.e. Steady State and Generational Replacement. The methods used in this process are the Generational Replacement at a chromosome, generations are updated or replaced together with a new chromosome form crossover and mutation criteria, as well as the best chromosome is stored in the Elitisme.

III. RESULT AND ANALYSIS

We devote the accuracy of microarray databy doing comparative analysis between two selection methodologies: Haar wavelet and GA. The performance of our proposed selection process is measured by accuracy of confusion matrix. According to selection processes using Haar wavelet and GA, we have an optimal dimension reduction reported for300 and 500 attributes. Also, 500 individuals for GA and5 maximum generation and 100 population size.

Table. 2 Accuracy of microarray data by Naïve Bayes (NB) with 500 attributes

Dataset	Method	Accuracy (%)
Colon Tumor	Haar wavelet-NB	75.294
	GA-NB	73.235
Lung Cancer	Haar wavelet-NB	93.213
	GA-NB	97.568
Ovarian	Haar wavelet-NB	100
	GA-NB	95.389

Table. 3 Accuracy of microarray data by Naïve Bayes (NB) with 300 attributes

Dataset	Method	Accuracy (%)
Colon Tumor	Haar wavelet-NB	75.294
	GA-NB	73.235
Lung Cancer	Haar wavelet-NB	95.136
	GA-NB	97.568
Ovarian	Haar wavelet-NB	100
	GA-NB	93.497

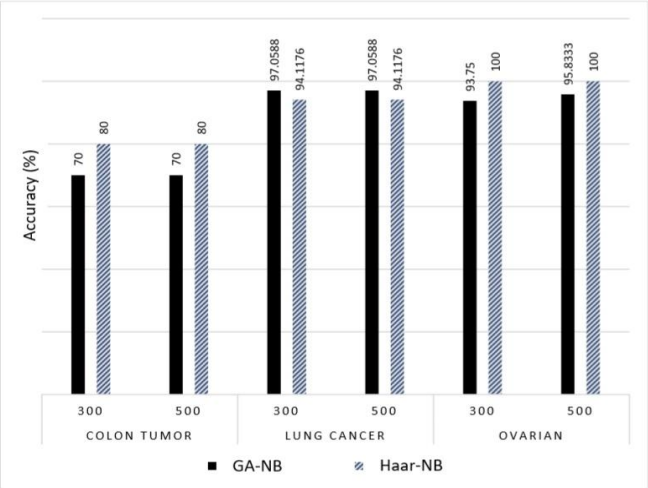

Fig. 4 An average accuracy

From Table 2, we show that Haar Wavelet-NB has better accuracy than GA-NB for colon tumor dataset, as same as ovarian cancer dataset, with mean of accuracy about 89,50%. Generally, Haar wavelet outperforms than GA-NB due to the significant accuracy performance. Then, we simulate for 300 dimension reduction using Haar wavelet, we can see Table 3. In Fig.4, an average accuracy of microarray dataset with Haar wavelet performed very well than GA. Considering 80% an 20% proportion of training and testing dataset, an accuracy Haar wavelet-NB for each colon tumor, lung and ovarian cancers are: 80%, 96,154% and 100%

IV. CONCLUSIONS

Microarray is a modern technique that facilitates simulation analysis of a large amount of gene expression data required to resolve complex biological problems. Microarray data classification processes require more effort because of the large dimensions and complex relationships between various genes. This paper proposes a new framework for detecting cancer based on microarray data using Haar wavelet dimension reduction and Naïve Bayes (NB) classification. Moreover, NB classification has a very strong assumption of independence of each attribute. Haar wavelet is the simplest wavelet function and has relatively efficient computing. The Haar wavelet-NB is able to predict quite well for colon tumor, lung and ovarian cancer dataset related to the accuracy of confusion matrix. The classification results through by Haar wavelet-NB outperforms than GA-NB due to an accuracy about 90% and 88% from actual (observable) microarray dataset

REFERENCES

1. Kumar, M., Singh, S. and Rath, S. (2015). "Classification of microarray data using functional link neural network. *Procedia Computer Science* 57", page 727–737.
2. Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. (2009)."Naïve bayes vs. decision trees vs. neural networks in the classification of training web pages".*International Journal of Computer Science Issues* 4(1), page 16-23.
3. Nurfalah, A., Adiwijaya, and Suryani, A. A. (2016)."Cancer detection based on microarray data classification using PCA and modified back propagation. *Far East Journal of Electronics and Communications* 16(2), page 269-281.
4. Morettin, P.A. (2004). *Waves and Wavelets: From Fourier to Wavelet Analysis of Time Series*. Institute of Mathematics and Statistics of University of São Paulo.
5. Phinyomar, A., Nuidod, P., Phukpattaranont, P. and Limsakul, C. (2012). "Feature extraction and selection of wavelet transform coefficients for EMG pattern classification". *Elektronika Ir Elektrotechnika* 122(6), page 28-32.
6. Rohmawati, A. A. and Adiwijaya. "A daubechies wavelet transformation to optimize modeling calibration of active compound on drug plants". In5rd*International Conference on Information and Communication Technology*, page 1-4. 2017.
7. Sunaryo, S. (2005). "Calibration model with wavelet transformation as pre-processing method". Bogor: Sekolah Pascasarjana, Institut Pertanian Bogor [PhD Thesis].
8. Mubarok, M.S., Adiwijaya, and Aldhi, M.D.(2017). "Aspect-Based Sentiment Analysis To Review Products Using Naïve Bayes". In AIP Conference Proceedings 1867(1).
9. Li, J. *Kent-ridge bio-medical data set repository*. School of Computer Engineering, Nanyang Technological University, Singapore. Downloaded on April 2017.
10. Antoniadis, A.(2003). *An Introduction to Wavelets and some Applications*. University Joseph Fourier, Laboratoire IMAG-LMC, France.
11. Suyanto, S. M. (2008). *Soft Computing*. Bandung: Informatika.