Implementing binary particle swarm optimization and C4.5 decision tree for cancer detection based on microarray data classification

A C Pradana¹, Adiwijaya² and A Aditsania³
School of Computing, Telkom University, Bandung, Indonesia

E-mail:¹amalyacitra@student.telkomuniversity.ac.id,²adiwijaya@telkomuniversity.ac.id,³aaditsania@telkomuniversity.ac.id

Abstract. Cancer is one of deadly disease in the world and needed to detect the symptoms early. Cancer can be represented with microarray data with measuring the changes occurred in gene expression level. Cancer detection can be done by doing classification technique for microarray data. One of most algorithm that applied for classification is C4.5 Decision Tree. It is a linear method which is easy to interpret and included into the algorithm which has given impact in classification but it is sensitive to noise data. Microarray data has a large features (high dimensional) which is not all the features has important information (high noise) and small samples which is causing the classification is difficult and affect the accuracy. Binary Particle Swarm Optimization (BPSO) is one of search optimization algorithm that could find the optimal feature. The purpose in this research consists of implementing and analysing the influence of feature selection and classification on microarray data using Binary Particle Swarm Optimization (BPSO) as feature selection and Decision Tree C4.5 as classifier. The discretization is needed for Decision Tree rule model and applied using K-Means. System is divided into two schemes such as Information Gain (IG) – C4.5 and BPSO – C4.5. The accuracy result based on IG – C4.5 and BPSO – C4.5 both are 54% and 99%. Applying feature selection before the classification could avoid the noise data in microarray data so it could form the rule accurately. With applying BPSO and Decision Tree is able to find the most significant feature and improve the accuracy.

1. Introduction
Bioinformatics is a research area that applied biology molecular into data that contains human DNA sequences (genomic data) with computer technology. This technology helps the researcher to diagnose health problem accurately with measuring the mRNA level to get information that occurs in cell condition and can understand biological process in it, for example is cancer disease. Cancer is one of deadly disease in the world [1]. One of way to avoid cancer is to detect the symptoms early. Cancer can be represented by using DNA microarray technology.

DNA microarray is a chip made by silicon/glass such as affymetrix array or microscopic head in illumin array which is molecule sequence or oligonucleotides complimentary DNA (cDNA) are planted called feature. With comparing result from DNA microarray technology with normal cell, the researcher can measure the changes occurs in gene expression level so all of gene expression can be investigated simultaneously [2].
To detect the cancer symptoms, one of technique that can be used is applied data mining to feature in microarray data. Data mining is a sequence of process to receive an important information that is still unknown, which is a classification technique to classify some data which have same characteristic. With this technique, we can determine the patient has cancer [3].

One of algorithm applied for classifying microarray data is decision tree algorithm, for example is decision tree C4.5. Decision tree is a linear method which is easy to interpret and included in ten of data mining algorithm which have given impact in microarray data classification but it has sensitive to noise data like microarray data [2]. Microarray data has a large features (high dimensional) which is not all features have important information (high noise) and has a small samples that caused the application of classification is difficult and affected the accuracy [1]. To overcome this problem, we can reduce the dimension before the classification which is applied feature selection as pre-processing data tool.

Feature selection consists of univariate and multivariate category. One of method on multivariate category is an embedded method. Embedded method has advantages which can test the prediction well [3]. In embedded method is looking for the feature that has the best contribution and optimal feature. Meta-heuristic approaches can be used in order to obtain the optimal feature. One of algorithm in meta-heuristic approaches is Particle Swarm Optimization (PSO). Feature selection is a problem which is handling discrete data so the PSO has to be modified into Binary Particle Swarm Optimization (BPSO) [4].

From the problem that has been described, we proposed to implement the Binary Particle Swarm Optimization (BPSO) as feature selection with Decision Tree C4.5 as classifier to obtain optimal feature and high accuracy.

Limitation in this research is the data which is used in this research are Breast Cancer, Colon Tumor, Lung Cancer, Leukimia ALL-AML, dan Ovarian Cancer refers from Kent Ridge Biomedical Dataset Repository. The initialize for the parameter in BPSO which will be used refers to previous research [4]. The Decision Tree C4.5 will work if the data is in discrete form. Thus, because microarray data is in continuous form then it has to be transformed into discrete. K-Means will be applied in order to transform the data into discrete form which is refers to previous study recommendation for applying discretization on microarray data [5].

The purpose in this research consists of implementing feature selection and classification on microarray data using Binary Particle Swarm Optimization (BPSO) as feature selection and Decision Tree C4.5 as classifier, obtaining the influence of feature selection on microarray data classification, and obtaining the performance which is generated from the system on cancer detection.

The structure in this research consist of five sections such as introduction section, literature review section that explains related study, definition and the method which is used, proposed system section that explains how the experiment and system is built, result and analysis section that explains the result and analysis from the experiment, and conclusion section.

2. Literature review

In this section explains the related study, the definition of microarray and some brief explanation of the methods which is used in this research.

2.1. Related study

The growth of machine learning in the latest 20 years is used in bioinformatics area for analysing health problem, such as to detect the symptoms of cancer. Some previous research which are implemented those algorithm for microarray data classification such as:

a. The research which is conducted by Kun-Huang Chen et al. in 2014 is analyzing some techniques that applied for classification for ten microarray data, such as Support Vector Machine, Self-organizing map, Back Propagation Neural Network, CART Decision Tree, C4.5 Decision Tree, Artificial Immune Recognition System, and Naive Bayes. Thus, the proposed method is Binary Particle Swarm Optimization (BPSO) as feature selection and Decision Tree C4.5 as classifier.
The result which is obtained is BPSO with C4.5 can provide high accuracy around 87% rather than Support Vector Machine (83%), Self-organizing map (55%), Back Propagation Neural Network (44%), CART Decision Tree (70%), C4.5 Decision Tree (73%), Artificial Immune Recognition System (50%), and Naive Bayes (75%) [8].

b. The research which is conducted by Meng-Chang Tsai, Kun-Huang Chen, Chao-Ton Su, and Hung-Chun Lin in 2012 is analyzing some techniques that applied for classification for five microarray data, such as Support Vector Machine, Back Propagation Neural Network, Logistic Regression, and C4.5 Decision Tree. Thus, the proposed method is Binary Particle Swarm Optimization (BPSO) as feature selection and Decision Tree C4.5 as classifier. The result which is obtained is BPSO with C4.5 can provide high accuracy around 75% rather than Support Vector Machine (70%), Back Propagation Neural Network (58%), Logistic Regression (74%) and C4.5 Decision Tree (71%) [9].

c. The research which is conducted by Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, and Cheng-Hong Yang in 2007 is analyzing some techniques that applied for classification for five microarray data, such as SFS, PTA, SFFS, SGA, and HGA. Thus, the proposed method is Binary Particle Swarm Optimization (BPSO) as feature selection and Support Vector Machine as classifier. The result which is obtained is BPSO with SVM can provide accuracy above 70% [10].

d. The research which is conducted by Jian J. Dai, Linh Lieu, and David Rocke in 2006 is analyzing some techniques that applied for classification for two microarray data, such as Particle Least Square, Sliced Inverse Regression, and Principal Component Analysis. The error rate result which is obtained is PLS is not obtained around 0.025 for leukemia and around 0.136 for colon rather than SIR which is around 0.026 for leukemia and around 0.141 for colon, thus PCA which is around 0.042 for leukemia and around 0.162 for colon [11].

e. The research which is conducted by Nurfalah, A., Adiwijaya and Suryani, A.A. in 2016 is analyzing a technique that applied Neural Network and PCA. The result which is obtained from analyzing ovarian, colon and leukemia data can provide accuracy above 80% [12].

Others related research, namely microarray data classification using Levenberg-Marquardt Backpropagation (LMBP)-Principal Component Analysis (PCA) [16] and Random Forest-K-Means [17] give a good performance on accuracy around 90%.

2.2. DNA microarray

Bioinformatics is research area which is applied biology molecular into data contains human DNA sequence (genomic data) with computer technology. Genes is made into protein depends on condition and time differently. At producing the protein, the instruction are transcripted by messenger RNA(mRNA) from DNA on left gene in nucleus and the production occurs in ribosom cell. As the result, the cell condition will correlate with the changes in mRNA level. With measuring mRNA level will obtain what is happened with the cell and be able to understand biological information process that occurred in the cell. This technology helps the researcher to diagnose health problem accurately. Those health data can be represented using DNA microarray technology [2].

DNA microarray is a chip which is made from silicon or glass in affymetrix array where thousand od cDNA molecule are implanted. The mixed from mRNA is inherited from cell and allowed to do hybridization to probe chip sequence. Hybridization level is detected using fluorescence dyes with imaging application from chip fabric. Microarray has characteristic such as large features (high dimensional) which is not all of feature have important information (high noise) and small samples [3].

DNA microarray technology is developed for gaining gene expression data which has large data simultaneously. Gene is expressed from some condition and time into protein. The instruction are transcripted by mRNA from DNA in left gene in nucleus. As the result, the cell condition will be correlated with the changes in mRNA level so it can be determined what happens in the cell [1].
2.3. Binary Particle Swarm Optimization (BPSO)

Binary Particle Swarm Optimization (BPSO) is a method which is modified from Particle Swarm Optimization (PSO) for feature selection case, which is included to meta-heuristic approach based on modest social behaviour from the bird flock or fish flock. Observing from the birds, if a bird locates food, the information about the location will be spread out to their flock and they can allocate the food. Simulating this behaviour based on their scheme will generate a good optimization algorithm. Assuming those birds is a particle and the flock are the population of particle. Particle has ability to communicate with each other and tries to reach other nearby particle which has the best fitness, for example is a particle which has the position close to the origin. From the result shows that all of particle that locate in same position will have good fitness like the origin. Each particle has velocity parameter that can move other particle immigrate to the particle which has the best fitness. The velocity formula is defined as in equation (2.1) [6].

\[ v_{id}^{\text{new}} = w \times v_{id}^{\text{old}} + c_1 r_1 (p_{bi}^{\text{old}} - x_{id}^{\text{old}}) + c_2 r_2 (g_{bi}^{\text{old}} - x_{id}^{\text{old}}) \]  

where:
- \( v_{id}^{\text{new}} \): the newest velocities of particle
- \( v_{id}^{\text{old}} \): the oldest velocities of particle
- \( x_{id}^{\text{old}} \): the oldest position of particle
- \( r_1 \) and \( r_2 \): random value between 0 and 1
- \( c_1 \): cognitive learning factor
- \( c_2 \): social learning factor
- \( w \): inertia weight
- \( g_{bi} \): global best
- \( p_{bi} \): personal best

The latest position of particle before which has good fitness will be saved as personal best (pb) and the best position of particle which has the best fitness among the population will be saved as global best (gb). \( x_{id} \) is the current position starts from particle i\( ^{\text{th}} \) where d is dimension from search space [4]. In this method is required to input the parameter value such as \( v_{\text{max}}, v_{\text{min}}, r_1, r_2, c_1, c_2 \) and \( w \), so it will obtain optimal search result [13].

In PSO, position of particle is generated by random number only. But in this case the representation of position could not be just a random number because it could not represent the feature, so it has to be modified into binary value. In BPSO, position of particle is modelled into bit of string to limit the velocity at interval \([0,1]\). The velocity is defined as probability consists of bit \( X_{ij} \) (i\( ^{\text{th}} \) particles and j\( ^{\text{th}} \) bit) to obtain number value of one. To limit the velocity can use limiting transformation formula where is applied sigmoid function proposed from previous research and is conducted by Kennedy Eberhart [13], which is defined in equation (2.2) [6].

\[ x_{id}^{\text{new}} = \begin{cases} 1, & \text{sigmoid}(v_{id}^{\text{new}}) > \text{rand} \\ 0, & \text{sigmoid}(v_{id}^{\text{new}}) < \text{rand} \end{cases} \]  

where \( x_{id}^{\text{new}} \) is new position and rand is a random value at interval \([0,1]\) and the formula of sigmoid is defined in equation (2.3).

\[ \text{sigmoid}(v_{id}^{\text{new}}) = \frac{1}{1 + e^{-v_{id}^{\text{new}}}} \]  

2.4. Decision Tree C4.5

Decision Tree C4.5 is one of top-down algorithm which builds decision tree model using recursive (divides and conquer strategy). This algorithm is improved from Decision Tree ID3, where is the calculation of gene selection and tree construction are improved. C4.5 has been used for solving real
world problem, exclusively for making decision in medical problem because it provides classification accuracy and simply can be represented. Using gain ratio as splitting criteria, the attribute which has important information is computed on training data, then selects the informative attribute, and so on [9].

Before extracting the data into tree model, the process that should be done is to determine attribute which will be the root based on the greatest gain ratio, then determines internal node to each branch from parent node, makes a decision node when selected attribute cannot be used, and prunes the tree if it is not significant differently between each child node [9].

a. Entropy Value Calculation

Entropy is used to measure heterogeneity from data sample. The formula for calculating the entropy is defined in equation (2.4).

\[ \text{Entropy}(S) = \sum_{i}^{C} - p_i \log_2 p_i \]  

(2.4)

where:
- \( C \) = value in targeted attribute
- \( p_i \) = sample proportion in class \( i^{th} \)

b. Information Gain Calculation

The formula for calculating information gain is defined in Equation (2.5).

\[ \text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v} \frac{|s_v|}{S} \text{entropy}(s_v) \]  

(2.5)

where:
- \( V \) = probable value for attribute \( A \)
- \( |s_v| \) = sample value for value \( V \)
- \( S \) = value from all data sample
- \( \text{Entropy}(S) = \text{entropy for sample with value } V \)

c. Gain ratio

The formula for calculation gain ratio is defined in Equation (2.6).

\[ \text{gain ratio} = \frac{\text{Gain}(S, A)}{\text{split information } (S, A)} \]  

(2.6)

where the information is defined in Equation (2.7).

\[ \text{split information} = \sum_{i} \frac{s_i}{S} \log_2 \frac{s_i}{S} \]  

(2.7)

where the value of \( s_i \) is subset \( C \) made from splitting \( S \) using attribute \( A \) with variant \( C \) value.

2.5. K-Fold Cross Validation

K-Fold Cross Validation is one of statistical method to ensure the difference classification result between data and avoid the random result. This method divides the data into two schemes which are training data model scheme and evaluation scheme. Then the data is divided as k value and doing the cross validation [8].

The first step is determining the k value. For each iteration, divides data into training data and testing data, then determines which is the data will be assigned into training data and testing data based on k values. The performance of each case is considered from the accuracy that is obtained from case. After all cases are done, then we will obtain the average of accuracy [8].
2.6. *K*-Means

*K*-Means is one of clustering method applied in most algorithm in data mining. This method is grouping the data as k value with making initialization centroid value randomly and making label as much as k value, then calculating the distance from each data to centroid. If the data which has nearest distance to any of centroid then it will be given a label from the cluster. After that, value of centroid is updated and evaluated to minimize the error rate so it will obtain significant result using Sum of Square (SSE) [14]. The formula for calculating centroid value is defined in equation (2.8).

\[
\text{centroid} = (\max(p_{i,n}) - \min(p_{i,n})) \times \text{rand} + \min(p_{i,n})
\] (2.8)

where:
- \( n \) = total data
- \( p_{i,n} \) = \( i \)th data until \( n \)th data
- \( \text{rand} \) = random value between [0,1]
- \( \max \) = the largest value of data
- \( \min \) = the smallest value of data

The formula for calculating distance between data and centroid using euclidean distance is defined in equation (2.9).

\[
\text{euclidean distance} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{k} (p_i - c_j)^2}
\] (2.9)

where:
- \( n \) = total data
- \( k \) = total cluster
- \( p_i \) = \( i \)th data
- \( c_j \) = \( j \)th centroid

The formula for calculating error rate using SSE is defined in equation (2.10).

\[
\text{SSE} = \sum_{i=1}^{n} \sum_{j=1}^{k} ||p_i - c_j||^2
\] (2.10)

where:
- \( n \) = total data
- \( k \) = total cluster
- \( p_i \) = \( i \)th data
- \( c_j \) = \( j \)th centroid

3. Proposed system

Generally, system begins with dividing the data into training data and testing data. Data is divided based on K-Fold Cross Validation scheme. After data has divided, there are two schemes in the system, they are classification with Information Gain feature selection scheme and classification with BPSO as feature selection scheme.

The data which is used in this experiment is taken from Kent Ridge Repository which can be downloaded in http://leo.ugr.es/elvira/DBCRepository/. Dataset consists of Breast Cancer, Colon Tumor, Lung Cancer, Leukimia ALL-AML, and Ovarian Cancer. Class of each record will be represented as class 1 and class 0. Dataset distribution is shown in table 1.

| Dataset       | Total Records | Total Features | Class Representation |
|---------------|---------------|----------------|----------------------|
| Breast Cancer | 97            | 24.188         | Class 1: Relapse     |
|               |               |                | Class 0: Nonrelapse   |
When the scheme is done, the accuracy will be accumulated and receive the average of accuracy. After getting the average of accuracy, this system is done. Overview of system is illustrated in figure 1.

| Cancer Type         | Count | Average | Classifications          |
|---------------------|-------|---------|--------------------------|
| Colon Tumor         | 62    | 2.000   | Positive                 |
| Lung Cancer         | 181   | 12.533  | ADCA, Mesothelioma       |
| Leukemia ALL-AML    | 72    | 7.129   | AML, ALL                 |
| Ovarian Cancer      | 253   | 15.154  | Cancer, Normal           |

**Figure 1.** System overview.
3.1. Data normalization
In normalization, data will be converted into interval [0,1]. First, the maximum and minimum value from each data will be searched. Next, original data is subtracted with the minimum value, then divided with difference between the maximum and minimum value. Thus, those step is repeated until all data has been converted.

3.2. Data discretization using K-Means
In discretization, the data that has been normalized will be converted into discrete in range one to k value. First, k value is determined, then the maximum and minimum value from each feature will be searched, and maximum of iteration and maximum of convergence is determined. In this research, k value is assigned to 5, maximum of iteration is assigned to 100, and maximum of convergence is assigned to 3.

Afterwards, centroid is initialized by random and each centroid is represented by cluster label. Next, as iteration and convergence value are still not reach their maximum, calculates the distance between data and centroid. If the distance between certain centroid and the data is the smallest than to other centroid, then assigns the data with the cluster label. The value of selected centroid is updated and evaluated by using SSE until the changes is stabilized significantly. Thus, those steps is repeated until all data has been converted.

3.3. Information gain as feature selection
In this scheme, the feature which is used will be selected by Information Gain. Training data will be used for modelling the rule. First, threshold value need to be initialize for choosing the informative attribute. Then for each feature calculates information gain value, and check if it has information gain value more than the threshold value then it will be taken for classification. After the classification, the accuracy will be obtained. The threshold values which will be used are 0.1, 0.2, and 0.3.

3.4. BPSO as feature selection
In this scheme, the feature which is used will be selected by BPSO. Training data will be used for modelling the rule. First, the parameter in BPSO is need to be initialize such as maximum of iteration value, maximum of convergence value, cognitive learning ($c_1$), social learning ($c_2$), lower bound velocity ($v_{\text{min}}$), upper bound velocity ($v_{\text{max}}$), inertia weight ($w$), $r_1$ (random value), $r_2$ (random value), total particle in population, and total feature in particle. In this research, maximum of iteration value is assigned to 100, maximum of convergence value is assigned to 25, $r_1$ (random value) is assigned randomly to interval [0,1], $r_2$ (random value) is assigned randomly to interval [0,1], total particle in population is assigned to 10, and total feature in particle is assigned as many as total feature in data. Meanwhile the other parameter such as cognitive learning ($c_1$), social learning ($c_2$), lower bound velocity ($v_{\text{min}}$), upper bound velocity ($v_{\text{max}}$), inertia weight ($w$) are using observation scheme which is described in table 2.

| Observation Number | Parameter Value |
|--------------------|-----------------|
|                    | cognitive learning ($c_1$) | social learning ($c_2$) | lower bound velocity ($v_{\text{min}}$) | upper bound velocity ($v_{\text{max}}$) | inertia weight ($w$) |
| 1                  | 1                | 1                   | -1                           | 1                        | 0.1                 |
| 2                  | 2                | 2                   | -2                           | 2                        | 0.2                 |
| 3                  | 3                | 3                   | -3                           | 3                        | 0.3                 |

Afterwards, population is initialized which consists of some particle. Each of particle has a fitness value (accuracy which is generated from classification) and a content. The content consists of real
position which has real value in interval \([0,1]\), velocity which has real, and discret position which has discret value (0 or 1) as a representation explains that the feature is being used or not. Particle length is assigned to total feature in data. Each particle has a personalbest which is the best particle among the individual or local, and a globalbest which is the best particle among the set or global.

Next, as iteration and convergence value are still not reach their maximum, fitness evaluation is conducted for each particle using training data with Decision Tree C4.5. Feature that will be used in training data is selected based on discret position of the particle. If a fitness value of certain particle is better than fitness of personalbest and globalbest, then personalbest and globalbest are updated. Next, for each particle calculates the new value velocity. If the new value velocity is not between interval \(v_{\text{max}}\) and \(v_{\text{min}}\) then the new value velocity is updated. After that, sigmoid is calculated based on velocity. If sigmoid value is higher than random value generated in interval \([0,1]\) then the discret position is assigned to 1, else the discret position is assigned to 0. Those steps are repeated until the iteration and convergence value is maximum and generated a particle which has the best fitness.

3.5. Classification using C4.5 Decision Tree
In this scheme, all features in data will be used. Training data will be used for modelling the rule. First, for each feature calculates entropy, information gain, split info, and gain ratio value. Next, all features will be sorted based on their gain ratio value. The feature which has the largest gain ratio among all features is selected and chosen as a root. Afterwards, rule is constructed based on the attribute in each feature. If a certain attribute in feature cannot significantly find the class, then it has to construct another node and select another feature based on information in the root. Thus, those steps are repeated until all rule has been constructed and met the criteria by obtaining class prediction significantly.

3.6. Accuracy calculation
Accuracy calculation is used to measure the classification performance. The measurement form is represented in confusion matrix, where the true positive value represents the prediction is true for class 1, false positive value represents the prediction is false for class, true negative value represents the prediction is true for class 0, false negative value represents the prediction is false for class 0. The representation is formed into a matrix and shown in table 3 [15].

| Actual Class | Prediction Class |
|--------------|-----------------|
|              | Class 1         | Class 0         |
| Class 1      | True Class 1    | False Class 1   |
| Class 0      | False Class 0   | True Class 0    |

The formula of calculating accuracy is defined in equation (3.1).

\[
\text{Accuracy} = \frac{\text{True Class 1} + \text{True Class 0}}{\text{True Class 1} + \text{False Class 1} + \text{True Class 0} + \text{False Class 0}} \quad (3.1)
\]

3.7. Validation using K-Fold Cross Validation
K-Fold Cross Validation is used to validate the classification result to all data. In this research, k value is assigned to 5. For dividing the training data and testing data is used 4:1 comparison. Cross validation scheme is begun with the first part as the testing data and the other part as the training data, and continued alternately until all data is being used.
4. Result and analysis

The experiment in this research is done as follows the observation scenario. On data discretization, the k value is assigned with 3, 5 and 7. On information gain as feature selection, the threshold value is assigned with 0.1, 0.2, and 0.3. On BPSO as feature selection, the parameter value is assigned with value on scenario based on table 2. The result from both scheme are described in table 4, table 5 and table 6.

| Dataset                  | The Average of Accuracy (%) | Information Gain | BPSO          |
|--------------------------|----------------------------|------------------|---------------|
|                          |                            | Threshold > 0.1  | Scenario 1    |
|                          |                            | Threshold > 0.2  | Scenario 2    |
|                          |                            | Threshold > 0.3  | Scenario 3    |
|AML-ALL Leukemia Breast   | 48,05                      | 64,84            | 67,16         |
|Breast Cancer Colon Tumor | 51                         | 72,67            | 46            |
|Lung Cancer Ovarian Cancer| 83,86                      | 73,33            | 70,88         |
|Lung Cancer Ovarian Cancer| 45,84                      | 51,98            | 52,16         |
|Average                   | 55,47                      | 56,51            | 47,24         |

| Table 5. Result using K=5 in data discretization.|
| Dataset                  | The Average of Accuracy (%) | Information Gain | BPSO          |
|--------------------------|----------------------------|------------------|---------------|
|                          |                            | Threshold > 0.1  | Scenario 1    |
|                          |                            | Threshold > 0.2  | Scenario 2    |
|                          |                            | Threshold > 0.3  | Scenario 3    |
|AML-ALL Leukemia Breast   | 34,48                      | 66,35            | 73,72         |
|Breast Cancer Colon Tumor | 78,19                      | 58,01            | 0             |
|Lung Cancer Ovarian Cancer| 77,78                      | 71,56            | 44,28         |
|Lung Cancer Ovarian Cancer| 42,17                      | 36,36            | 86,77         |
|Average                   | 58,61                      | 58,47            | 52,97         |

| Table 6. Result using K=7 in data discretization.|
| Dataset                  | The Average of Accuracy (%) | Information Gain | BPSO          |
|--------------------------|----------------------------|------------------|---------------|
|                          |                            | Threshold > 0.1  | Scenario 1    |
|                          |                            | Threshold > 0.2  | Scenario 2    |
|                          |                            | Threshold > 0.3  | Scenario 3    |
|AML-ALL Leukemia Breast   | 50,86                      | 54,68            | 72,87         |

The summary of accuracy based on scheme and k value in data discretization is described in table 7.

| Dataset          | Information Gain | BPSO          |
|------------------|------------------|---------------|
|                  | Threshold > 0.1| Threshold > 0.2 | Threshold > 0.3 | Scenario 1 | Scenario 2 | Scenario 3 |
| ALL              |                  |               |               |            |            |            |
| Leukemia         | 67.25            | 34.57         | 10.47         | 100        | 100        | 100        |
| Breast Cancer    | 49.78            | 44.16         | 55.15         | 96.36      | 93.8       | 93.86      |
| Colon Tumor      | 57.41            | 34.84         | 35.97         | 100        | 100        | 100        |
| Lung Cancer      | 86.18            | 86.6          | 66.32         | 100        | 100        | 100        |
| Ovarian Cancer   | Average          | 62.29         | 50.97         | 43.37      | 99.27      | 98.77      | 98.72      |

From the summary, it is obtained that the accuracy of the Information Gain + Decision Tree C4.5 and BPSO + Decision Tree C4.5 scheme are unstable as the k value is used increasingly. With the discretization data using k = 3 obtains the accuracy of Information Gain + Decision Tree C4.5 is the most better than the other k value, whereas with the discretization data using k = 5 the accuracy of BPSO + Decision Tree C4.5 is the most better than the other k values.

In the discretization data using k = 3, in the Information Gain + Decision Tree C4.5 scheme the accuracy increases significantly with the increase of the threshold value and obtains the best accuracy with the threshold value = 0.3. In the BPSO + Decision Tree C4.5 scheme the accuracy is unstable with increasing scenarios and getting the best accuracy with the 2nd scenario.

In the discretization data using k = 5, in the Information Gain + Decision Tree C4.5 scheme the accuracy decreases significantly with increasing threshold value and obtains the best accuracy with threshold value = 0.1. In the BPSO + Decision Tree C4.5 scheme the accuracy is unstable with increasing scenarios and getting the best accuracy with the 2nd scenario.

In the discretization data using k = 7, in the Information Gain + Decision Tree C4.5 scheme the accuracy increases significantly with increasing threshold value and obtains the best accuracy with threshold value = 0.1. In the BPSO + Decision Tree C4.5 scheme the accuracy is unstable with increasing scenarios and getting the best accuracy with the 2nd scenario.

From these observation, it is found that the discretization k value significantly affects the accuracy of the Information Gain + Decision Tree C4.5 scheme and does not significantly affect the accuracy of the BPSO + Decision Tree C4.5 scheme. The 2nd scenario gets a stable accuracy result on the BPSO + Decision Tree C4.5 scheme with cognitive learning parameter \((c_1)\) is assigned to 2, social learning \((c_2)\)
is assigned to 2, lower bound velocity \( (v_{\text{min}}) \) is assigned to -2, upper bound velocity \( (v_{\text{max}}) \) is assigned to 2, and inertia weight \( (w) \) is assigned to 0.2.

The accuracy of the Information Gain + Decision Tree C4.5 scheme decreases significantly with the increase in the number of k on the discrete used due to the larger cluster division of data so that the classification performance decreases as the threshold value selection filters the most informative feature. Implementation of Information Gain feature selection capable of selecting informative and influential data in the contribution of modeling rule also minimizes noise data. The accuracy of the BPSO + Decision Tree C4.5 scheme is stable on the use of parameters in the second scenario and minimizes the noise data for obtaining better accuracy results.

5. Conclusion
In this research, the performance result of the system for the five datas used obtained an average of 54\% accuracy for the classification scheme with the Information Gain as feature selection and 99\% accuracy for the BPSO classification scheme as feature selection. The accuracy of the classification scheme with BPSO as feature selection is better than the accuracy of the classification scheme with the Information Gain as feature selection. The use of the number of features that have a lot of noise affect the classification is done. The effect of feature selection on classification plays an important role in avoiding noise data to accurately model accurate prediction of test data. We can conclude that applying BPSO + Decision Tree C4.5 scheme which is using cognitive learning parameter \( (c_1) \) is assigned to 2, social learning \( (c_2) \) is assigned to 2, lower bound velocity \( (v_{\text{min}}) \) is assigned to -2, upper bound velocity \( (v_{\text{max}}) \) is assigned to 2, and inertia weight \( (w) \) is assigned to 0.2 is able to find the most significant feature and could improve the accuracy.

Other factors that affect the accuracy results are obtained from determining discretization methods, the k values in data discretization and K-Fold Cross Validation testing schemes, the iteration limit and constant weight factor, and initialization observation of the parameters performed. For further research, it may be possible to analyze in the discrete method, the number of k for discretization using K-Means, the iteration limit, the constant value of weight factor as well as the initialization of parameters with other observation with other data so it could improve the accuracy result.

References
[1] Wai-Ki Y, Amin S B, and Li C 2011 A Survey of Classification Techniques for Microarray Data Analysis Springer Handbooks of Computational Statistics chapter 10 (Berlin: Springer-Verlag)
[2] NCBI Microarrays 2007 [online] Available at https://www.ncbi.nlm.nih.gov/probe/docs/techmicroarray/ Accessed: 1 May 2018
[3] Speed T 2003 Statistical Analysis of Gene Expression Microarray Data (New York: Chapman & Hall/CRC) chapter 3
[4] Ulfrenborg B, Klinga-Levan K and Olsson B 2013 Classification of Tumor Samples from Expression Data Using Decision Trunks (Auckland: Libertas Academica).
[5] Ammu K P and Preeja V. 2013 Review on feature selection techniques of DNA microarray data 2013 Int. J. of Computer Applications 61 No 12
[6] Pashaei E O, Mustafa and Aydin N 2015 A novel gene selection algorithm for cancer identification based on Random Forest and Particle Swarm Optimization 2015 IEEE
[7] Yong L 2010 Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks 2010 BMC Bioinformatics
[8] Kun-Huang C 2014 Gene selection for cancer identification: a decision tree model empowered by Particle Swarm Optimization algorithm 2014 BMC Bioinformatics
[9] Meng-Chang T, Kun-Huang C, Chao-Ton S and L Hung-Chun 2012 An application of PSO algorithm and decision tree for medical problem 2012 Bali 2nd ICS pp 13-14
[10] Chung-Jui T, Li-Yeh C, Jun-Yang C and Cheng H 2007 Feature selection using PSO-SVM 2007
IAENG Int. J. of Computer Science

[11] Da J J, Linh L and David R 2006 Dimension reduction for classification with gene expression microarray data statistical application in genetics and molecular biology 2006 5 Issue 1

[12] Nurfalah A, Adiwijaya and Suryani A 2016 Cancer detection based on microarray data classification using PCA and Modified back propagation 2016 Far East J. of Electronics and Communication 16 Issue 2 pp 269

[13] Russell E and James K 1995 A New Optimizer using Particle Swarm Theory Sixth International Symp. on Micro Machine and Human Science IEEE

[14] Jianpeng Q 2016 K-Means: an effective and efficient K-Means clustering algorithm 2016 Int. Conf. on Big Data and Cloud Computing IEEE

[15] Fabrice G and Hamilton H J 2007 Quality Measures in Data Mining (Berlin: Springer)

[16] Adiwijaya, Wisesty U N, Lisnawati E, Aditsania A, Kusumo D S 2018 Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification Journal of Computer Science 14(11).

[17] Aydadenta H and Adiwijaya 2018 A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest Journal of Information Processing Systems 14(5).