On the classification techniques in data mining for microarray data classification

Husna Aydadenta and Adiwijaya
School of Computing, Telkom University, Jl. Telekomunikasi No.1, Bandung, 40257, Indonesia
E-mail: aydadenta@student.telkomuniversity.ac.id and adiwijaya@telkomuniversity.ac.id

Abstract. Cancer is one of the deadly diseases, according to data from WHO by 2015 there are 8.8 million more deaths caused by cancer, and this will increase every year if not resolved earlier. Microarray data has become one of the most popular cancer-identification studies in the field of health, since microarray data can be used to look at levels of gene expression in certain cell samples that serve to analyze thousands of genes simultaneously. By using data mining technique, we can classify the sample of microarray data thus it can be identified with cancer or not. In this paper we will discuss some research using some data mining techniques using microarray data, such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes, k-Nearest Neighbor (kNN), and C4.5, and simulation of Random Forest algorithm with technique of reduction dimension using Relief. The result of this paper show performance measure (accuracy) from classification algorithm (SVM, ANN, Naive Bayes, kNN, C4.5, and Random Forest). The results in this paper show the accuracy of Random Forest algorithm higher than other classification algorithms (Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes, k-Nearest Neighbor (kNN), and C4.5). It is hoped that this paper can provide some information about the speed, accuracy, performance and computational cost generated from each Data Mining Classification Technique based on microarray data.

1. Introduction

Microarray technology is one tool used by many biologists to monitor the genome level of gene expression within a particular organism [1]. Microarray is usually a glass slide where DNA molecules are installed regularly in certain locations called spots (features). Microarray contains thousands of spots and each spots contain millions of copies of identical DNA molecules that uniquely fit the genes (Figure 1).

Because the microarray contains many spots (features) to perform cancer detection using microarray data requires the help of data mining. Data mining is a series of processes to extract the added value of information that has been unknown manually from a database to manipulate data into more valuable information obtained by extracting and recognizing important or interesting patterns of data contained in the database [2].

Detecting cancer with data mining can be accomplished by using classification technique, this technique is used to predict the group membership of data instances [2]. To do the classification using microarray data is the biggest challenge in data mining. This is due to the small number of samples on microarray data while the number of features on microarray is a lot (high dimensional). The higher the dimension in the data and the number of fixed
observations then the value of accuracy on the classification at a certain point will decrease. To overcome this problem, we can use the process of reduction.

Dimensional reduction can eliminate redundancy in the data thus the features used in the classification are features that have a high correlation to the class. There are two types of dimensional reduction: feature selection and feature extraction. The feature selection works by removing irrelevant features and redundancy [3]. The purpose of feature selection is to get rid of irrelevant and noisy genes from the input data set, to speed up the processing of data by reducing the dimensionality, and to avoid over fitting of the classifier [4] While extraction features work by transforming the original data into a new representation. Feature extraction has the same goal as feature selection that eliminates irrelevant or noise features in the data and removes redundancy in the data in order to increase the value of classification accuracy.

2. Literature Survey

Kohbalan Moorthy and Mohd Saberi Mohamad [5] in their paper classified microarray data using Random Forest algorithm on microarray data. The data used were Adenocarcinoma, Brain, Breast2, Breast3, Colon, Leukemia, Lymphoma, NCI60, Prostate, SRBCT. For the classification using Random Forest they tested the OOB error rate with conditions, ie the lowest OOB error rate and the fewest subsets, the lowest OOB error rate and the largest number of subsets and standard processes or default random forest algorithm. From the research, the accuracy of the standard classification or the default random forest algorithm for gene selection was always lower than the lowest OOB error rate and the lowest number of subsets and the lowest OOB error rate and the largest number of subsets.

Devi Arockia Vanitha, Devaraj, and Venkatesulu M [6] in their research used Support Vector Machine (SVM) algorithm to classify two microarray data namely Colon cancer and Lymphoma. Then, they also used Mutual Information (MI) between genes and class labels to identify informative genes. The selected genes were used for learning using SVM and the evaluation in this study was using the Leave-one-Out Cross Validation (LOOCV) method. The proposed method of research conducted was able to reduce the feature dimension by identifying the most
informative subset of genes as well as improving classification accuracy.

Nurfalah, A., Adiwijaya, and Suryani, A.A. [7], did research which aimed to explore the use of Neural Networks (ANN) and feature extraction using PCA. Furthermore, Tsun-Chen Lin, Ru-Sheng Liu, Ya-Ting Chao, and Shu-Yuan Chen [8] did a research which aimed to explore the use of gene expression data in differentiating heterogeneous types of cancer. They used Genetic Algorithm hybrid learning methodology (GA) and Neural Networks (ANN) to find an optimal subgroup of genes for tissue cancer classification. The proposed method was tested on two microarray datasets namely NCI60 and GCM. The results of the research showed that GA and ANN methods not only excel in the prediction approach but also able to reduce the number of predictive genes required in the analysis of classification.

C.Lavanya, M.Nandihini, R.Niranjana, and C.Gunavathi [9] built a comprehensive framework that combined the features of selection and classification technique, in their research. In which classification technique used was Naive Bayes algorithm. As for the feature selection using T-Test measurement, Chi-Square Test was used to select the possibility of cancer-related genes from large microarray gene expression data. From the classification framework built, it successfully classified new samples as infected or normal.

Sampath Deegalla and Henrik Bostrom [10] conducted classification research on microarray data using KNN algorithm. The dataset used among others were Colon Tumor, Leukemia, Central Nervous, SRBCT, Lymphoma, Brain, NCI60, and Prostate. Prior classification they did feature reduction which included several method of features reduction used, namely Principal Component Analysis (PCA), Random Projection (RP), Partial Least Squares (PLS) and Information Gain (IG). The results showed that PCA and PLS achieved the best accuracy with fewer components than the other two methods. And the main conclusion of this research was that the option of the feature reduction method can be very important when classifying the microarray using the KNN algorithm.

Hong Hu, Jiuyong Li, Ashley Plank, Hua Wang, and Grant Daggard [11] in their research, did a classification using several methods such as LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest in seven data collection of Microarray cancer (Breast Cancer, Lung Cancer, Lymphoma, Leukemia, Colon, Ovarian, and Prostate). The results of the study showed that all ensemble methods outperform C4.5. The experimental results also showed that these five methods benefit from preprocessing data, including gene selection and discrimination, in classification accuracy.

3. Classification Techniques for Microarray Data
In this section includes the following subsections, namely: (1) Comparison of Classification Techniques; (2) Random Forest on the Microarray Data Classification

3.1. Comparison of classification techniques
In this subsection we provide a strength and weakness of some classification techniques for microarray data classification, namely Support Vector Machine, Artificial Neural Network, Naive Bayes, k-Nearest Neighbor, and C4.5.

3.1.1. Support Vector Machine (SVM) is a learning machine algorithm first introduced by Vapnik in 1992 [12]. The algorithm works with the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates the two classes in the input space. The basic principle of SVM is linear classifier, and further developed to work on nonlinear problems. by incorporating the kernel trick concept in high-dimensional workspaces.

3.1.2. Artificial Neural Network (ANN) is a classification algorithm that represents the behavior of human neural network. The basic principle of ANN is that a number of parameters
as inputs are processed in such a way as in the hidden layer (multiplication, addition, division, etc.), then processed again in the output layer to produce an output.

3.1.3. Naive Bayes

Naive Bayes is a classification algorithm using a simple probability that applies Bayes Theorem with high independent assumptions. Bayes theorem is a theorem used in statistics to calculate the probability of a class of each attribute group present, to determine which class is optimal.

3.1.4. k-Nearest Neighbor (k-NN)

is a supervised learning algorithm in which the results of a new instance query are classified by the majority of the categories in KNN. Working system of k-NN algorithm based on the shortest distance from query instance to training sample to determine its KNN. Training samples are projected into many-dimensional spaces, where each dimension represents a feature of the data. This space is divided into sections based on training sample classification. A point in this space is marked by class $c$ if class $c$ is the most common classification of the $k$ nearest neighbor of the point. Nearby or the distance of neighbors are usually calculated based on Euclidean Distance.

3.1.5. C4.5

is one of the most popular classification algorithms in the decision tree algorithm group. At the learning stage of the train data, the C4.5 algorithm constructs a decision tree. Then at the classification stage, the decision tree is used to predict the class of a sample whose class is not yet known.

Table 1: Strength and weakness of classification techniques for microarray data

| No | Algorithm                          | Strength                                                                 | Weakness                                                                                     |
|----|-----------------------------------|--------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| 1  | Support Vector Machine (SVM)      | - Generally generates accurate classification.                           | - Requires long training time.                                                               |
|    |                                   | - Less overfitting, robust to noise                                      | - Because it takes long learning thus it will produce expensive computing.                    |
| 2  | Artificial Neural Network (ANN)   | - Relatively easy to use.                                                | - It takes a long process time for large neural networks.                                    |
|    |                                   | - Good for complex issues like image recognition                        | - ANN architecture is different from the microprocessors architecture so it takes time to imitate it. |
| 3  | Naive Bayes                       | - It only requires a small amount of training data to determine the estimated parameters required in the classification process. | - Requires expensive computing.                                                              |
|    |                                   | - Able to handle quantitative and discrete data.                        | - Not applicable if the conditional probability is zero, if zero then the predicted probability will be zero as well. |
|    |                                   | - Handle missing values by ignoring agencies during an opportunity .    | - Assume independent variables.                                                              |
| 4  | k-Nearest Neighbor (k-NN)         | - Able to generate strong or clear data (especially if using a quadratic multiplication derivative at a distance magnitude). | - Requires a value of $k$ as a parameter.                                                    |
|    |                                   | - Effective if used for large data.                                     | - Computational price calculation is very high because this experiment requires calculating the distance from multiple queries for all experimental data. |
5. C4.5

- The area of decision making that was previously complex and very global, can be transformed into simple and specific.
- Elimination of unnecessary calculations, because when using the decision tree method the example is tested only on the basis of certain criteria or classes.
- Overlap occurs especially when the classes and criteria used are numerous.
- Difficult in optimal decision tree design.
- Decision-quality results obtained from the decision tree method depend largely on how the tree is designed.

3.2. Random forest algorithm on the microarray data classification

In the research conducted by Ramn Diaz-Uriarte and Sara Alvarez de Andrs about "Gene selection and classification of microarray data using random forest" said that if Random Forest Algorithm is suitable for use in Microarray Datasets this is because Random Forest algorithm can be used when the number of variables are more than the number of samples, as it is known that it is characteristic of microarray. In addition the Random Forest algorithm can also be used on data that has 2 classes or multiple classes, has good predictive performance even when most are predictive noise variables, and therefore RF does not require feature selection, and able to handle categorial and continuous mixtures Predictors [13].

The Random Forest algorithm is a classification algorithm developed by Leo Breiman [14] which uses an ensemble classification tree. The ensemble method is a classification method that produces more than one model, thus the random forest consists of more than 1 tree. Then, each of the classification trees is built using a boostrap sample of data at random. Classes resulting from this classification process are derived from the most classes generated by the decision trees present in the Random Forest. By voting on available decision trees making accuracy of Random Forest increases [13]. For the working process of random forest can be seen in the following Figure 2.

One of the hallmarks of the random forest algorithm is that it can be applied to a dataset that has a larger number of variables than the sample, but by applying random forest as gene selection, it is still very difficult, it is necessary to pre-process the removal dimension by sorting strong gene as a feature subset of the classification process. And for gene selection, this research used method of relief. Relief is one feature selection methods that utilizes weighting technique to measure the significance of a feature in the context of classification and features that have a weight value above the threshold used will be selected. Relief Weights are continuous values and allow features to be classified by relevance. Relief is also an interesting algorithm in Feature Selection because it has efficient computing.

As in Figure 2, the proposed method in this research consists of two processes, namely preprocessing and processing. In preprocessing, do normalization, it aims to speed up the computation process, then feature selection using relief method, using relief method will generate the rank for each feature, then selected top ranking as feature subset to be used for processing, it aims to remove redundancy and features that are no relevance. Then, the preprocessing is classified using a random forest algorithm, where the data subset used for learning is a feature subset of top ranking results in pre-processing. The random forest algorithm will produce more than one classifiers according to the number of trees in use. To make predictions using this algorithm using voting, where the most class of voting is what will be the predicted result on the model. Furthermore, after the model is obtained will be evaluated by using confusion matrix, and this process will be obtained accurately from the proposed method.
4. Results and Analysis

In this research, it was conducted Random Forest algorithm simulation using microarray data, which used data taken from [http://www.gems-system.org/](http://www.gems-system.org/), in this simulation, the data used were amounting to 6, among others: Table 3 shows the accuracy produced using Random Forest algorithm for gene selection and classification and using Relief for gene selection & Random Forest algorithm for classification.

Table 2: Microarray Data

| Data              | Number of class | Sample                              | Feature |
|-------------------|-----------------|-------------------------------------|---------|
| Colon             | 2               | 62 (22 Positif, 40 Negatif)         | 2001    |
| Ovarian           | 2               | 253 (91 Normal, 162 Cancer)         | 15155   |
| Central Nervous   | 2               | 60 (21 Class1, 39 Class0)           | 7129    |
| Lung Cancer       | 2               | 181 (31 Mesotheioma, 150 ADCA)      | 12533   |
| Tumor Prostate    | 2               | 136 (77Tumor, 59 Normal)            | 12600   |

The simulation was done using two scenarios, using Random Forest for classification and gene selection, then before using Random Forest for classification, the data was done preprocessing i.e. gene selection using Relief method. When the random forest algorithm is used as a gene selection as in research from Kohbalan Moorthy and Mohd Saberi Mohamad [5], it is performed by bootstrap by selecting as much as N at random with the replacement of all N available in training data. Bootstrap is a training data used to form a tree. Bootstrap consists of randomly selected data with returns. The data selection was done as much as the original data set, and the bootstrap was formed according to the number of trees to be built on the Random Forest method. Unselected data on bootstrap for each tree is used as Out-Of-Bag (OOB) data, this is...
Table 3: Accuracy of Random Forest algorithm simulation

| Data       | Tree | Random Forest | k | Random Forest & Relief |
|------------|------|---------------|---|------------------------|
| Colon      | 20   | 75%           | 10| 91.67%                 |
|            | 50   | 66.67%        | 10| 83.33%                 |
|            | 100  | 75%           | 10| 83.33%                 |
| Leukemia   | 20   | 85.71%        | 10| 92.86%                 |
|            | 50   | 85.71%        | 10| 92.86%                 |
|            | 100  | 80.00%        | 10| 92.86%                 |
| Ovarian    | 20   | 80%           | 10| 86%                    |
|            | 50   | 78%           | 10| 98%                    |
|            | 100  | 78%           | 10| 98%                    |

because the data selection is done randomly with a return. OOB data will be used in internal model testing using OOB error rate estimation. In the research Kohbalan Moorthy and Mohd Saberi Mohamad [5], the development of the OOB process, by comparing three scenarios, is a selection of the smallest subset of genes with lowest OOB error rates, selection of the biggest subset of genes with lowest OOB error rates and setting the minimum number of genes to be selected.

However, in this research tried to do the gene selection by using the method of relief. The relief method works by ranking for each gene, then selected top 10, 30, and 50 rankings and is used as a feature subset of the random forest process for the Bootstrap process, and tree building.

Then, on the simulation of Random Forest algorithm it was also tested with three parameters of the number of trees ie 20 trees, 30 trees, and 50 trees. In addition to the parameter of the number of trees it was also used three parameters \( k \), \( k \) value namely the number of variables used in the process of classification, the number of variables in total of three, ie 10, 30, and 50.

From the results in Table 3, it can be seen the accuracy of classification using Random Forest algorithm which before the classification process has been in dimension reduction using Relief method always higher than just using Random Forest classification algorithm. From the simulation results we can conclude that Random Forest algorithm is good for microarray data, but Random Forest also has some weaknesses, such as yielding unstable accuracy, this is influenced by input parameter in random forest, and random process of taking variable of each development trees, and large number of trees can make this algorithm slow for real-time predictions. Furthermore, this study also compared the performance of several classification algorithms on Table 4. Central Nervous has a sample size of 60 and a total of 7129 features,
and has 2 values in the target class ie class1 and class0. The results of comparing some of the classification algorithms can be seen in Table 4.

**Table 4: The results of comparing some of the classification algorithms**

| Method                          | Accuracy  |
|---------------------------------|-----------|
| Random Forest                   | 75%       |
| Support Vector Machine (SVM)    | 66.67%    |
| Artificial Neural Network (ANN) | 66.67%    |
| Naive Bayes                     | 41.67%    |
| k-Nearest Neighbor (k-NN)       | 58.33%    |
| C4.5                            | 58.33%    |

From Table 4 we compared some [15] classification algorithms and used Relief method for gene selection process with parameter $k$ for Relief method of 50. Parameter $k$ is number of features taken in gene selection to be subset feature. Then, other required parameters depending on the classification algorithm used. In this research, the parameter of tree number of random forest algorithm used was 30 trees, the number of neutrons in ANN 5 algorithm, and the number $k$ on k-NN was 5. From the 6 algorithms the Random Forest algorithm produced higher accuracy than the other five classification algorithms.

5. Conclusion

In this research it was done survey on some data mining classification techniques using microarray data. This study analyzed several classification algorithms, such as Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes, k-Nearest Neighbor (kNN), and C4.5. Some classification algorithms provided results based on the speed, accuracy, performance and resulting computational costs. From the research that has been done by using some of the classification algorithms produced good accuracy, and RF algorithm gave higher accuracy than other algorithms, but the high or low accuracy on the classification on microarray data can also be influenced by dimensional reduction used, such as in the simulation that has been done in this research using Relief method. And in this paper, it provides some surveys about data mining classification techniques using microarray data as well as the performance of the algorithm with the result of accuracy obtained. And for further research, it may be possible to add the compare time complexity required for some classification algorithms using microarray data.

References

[1] A. Sanchez and M. C. R. de Villa, “A tutorial review of microarray data analysis,” Departament d’Estadstica. Universitat de Barcelona., 2008.
[2] P. ning Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Pearson, 2005.
[3] Z. M. Hira and D. F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data,” in Advances in Bioinformatics, 2015.
[4] A. P. K and P. V, “Review on feature selection techniques of dna microarray data,” Computer Applications, 2013.
[5] K. Moorthy and M. S. Mohamad, “Random forest for gene selection and microarray data classification,” in Springer-Verlag Berlin Heidelberg, 2012.
[6] D. Arockia, Vanitha.C, Devaraj.D, and Venkatesulu.M. “Gene expression data classification using support vector machine and mutual information-based gene selection,” in Procedia Computer Science (ScienceDirect), 47, p. 13 21, 2015.
[7] A. Nurfalah, Adiwijaya, and A. Suryani, “Cancer detection based on microarray data classification using pca and modified back propagation.,” Far East Journal of Electronics and Communications 16(2), p. p.269, 2016.

[8] T.-C. Lin, R.-S. Liu, Y.-T. Chao, and S.-Y. Chen, “Multiclass microarray data classification using ga/ann method,” in Springer-Verlag Berlin Heidelberg, p. 1037–1041, 2006.

[9] C. Lavanya, M. Nandihini, R. Niranjana, and C. Gunavathi, “Classification of microarray data based on feature selection method,” in International Journal of Innovative Research in Science, Engineering and Technology, 3, February 2014.

[10] S. Deegalla and H. Bostrom, “Classification of microarrays with knn: Comparison of dimensionality reduction methods,” in Springer-Verlag Berlin Heidelberg, pp. 800–809, 2007.

[11] H. Hu, J. Li, A. Plank, H. Wang, and G. Daggard, “A comparative study of classification methods for microarray data analysis,” in Conferences in Research and Practice in Information Technology (CRPIT), 61, 2006.

[12] L. Auria and R. A. Moro, Support Vector Machines (SVM) as a Technique for Solvency Analysis, German Institute for Economic Research, 2008.

[13] R. Daz-Uriarte and S. A. de Andrs, “Gene selection and classification of microarray data using random forest,” BMC Bioinformatics, 2006.

[14] L. Breiman, “Random forest–random feature,” in Statistics Department University of California Berkeley, 1999.

[15] C. Sekhar, Minal, and E. Madhu, “Multimodal choice expression modelling using random forest decision trees,” in ijtte, 2016.