CHAPTER 1

INTRODUCTION

This chapter discusses the underlying background of the research, the concept, problem statement, objective, hypotheses and lastly the importance of the study. This chapter consists of seven sub-chapters namely: (1) Background; (2) Gap of Previous Research and This Research; (3) Problem Definition; (4) Problem Limitations; (5) Objective; (6) Hypotheses; and (7) Scope of Work

1.1 Background

Cancer is one of the deadly diseases, according to WHO data by 2015 there are 8.8 million more deaths caused by cancer, and this will increase every year if it is not resolved in early [1]. There are many ways to detect cancer, one of them is by using microarray technique. Microarray takes an essential part in diagnosing a disease because microarray analysis can be used to look at the level of gene expression in a particular cell sample that serves to analyze thousands of genes simultaneously [2]. Therefore, microarray has high data dimension. The higher the size of the data and the number of fixed observations then the value of accuracy on the classification at a certain point will decrease. To overcome the problem, reduction process is conducted.

There are two types of dimensional reduction processes that are often used, namely feature selection and feature extraction. The feature selection works by removing unnecessary features and redundancy [3]. The objective of feature selection is to get rid of foreign and noisy genes from the input data set, to speed up the processing of data by reducing the dimensionality, and to avoid overfitting of the classifier [4]. While feature extraction works by transforming the original data into a new representation. Feature extraction has the same goal as feature selection that eliminates unnecessary features or noise on the data and removes redundancy in the data to increase the value of classification accuracy.

Microarray data provides information about the expression of a gene, so it is possible that some other genes have the same expression. For example, some genes in a data may have properties that are similar to other genes in other words, a gene that has a high relevancy to its class and other genes that have a high relevancy resembling the previous gene can be selected [5]. In one study, combining highly effective genes with other highly effective genes is not a good feature set [5], because combining two genes would not increase the information. This is called redundancy condition [5]. Therefore, a minimum redundancy process is required in the gene selection process.

Based on the research of Ramn Daz-Uriarte and Sara Alvarez de Andrs [6], the Random Forest classification algorithm has several advantages, such as can be used for variations

of observations (characteristics of microarray data). It can be used both for two- class and multi-class problems of more than two types. And Random Forest has good predictive performance even when most predictive variables are noisy, and does not require a pre-selection of genes, does not overfit. Also, it can handle the mixture of categorical and continuous predictors, incorporates interactions among predictor variables. The output is invariant to monotone transformations of the predictors, returns measures of variable (gene) importance, and there is little need to fine-tune parameters to achieve excellent performance. Of the several advantages of Random Forest algorithm suitable to be implemented on a microarray.

Based on the advantages of Random Forest algorithm described in previous research, this research will analyze the performance of Random Forests algorithm using several data microarray dataset by doing development dimension reduction process. To see the performance of the algorithm some scenarios are conducted, such as using parameters of algorithms, and dividing between training data and data testing.

1.2 Gap of Previous Research and This Research

Many researchers have performed cancer detection using Microarray data. Several studies have been proposed by classification. Several algorithms have been proposed, some papers examined each algorithm separately in a closed condition. Those algorithms have advantages and weaknesses.

The main problem on Microarray data has more variables than the samples. Whereas to get an accurate model with classification requires a lot of sample data and variables that have a correlation with the class on the dataset. So, the first thing to do is to look for the variables or features relevant to the type.

In the [7] research, they use the Random Forest for gene selection, and classification since Random Forest Algorithm can look for essential variables in the dataset. This made it suitable for use on datasets that have variable numbers more than the number of samples. However, its weakness is that Random Forest is only able to detect significant variable, but not redundancy variable. Redundancy variables are variables or features that have similarly. In [8] research, if using the same or similar characteristics or variables in the classification process is not right, it can decrease the accuracy of the model. So it takes an approach to remove redundancy on the data.

Some studies apply two approaches to find relevant variables and remove redundancy on the dataset, ie, using feature extraction and selection. The feature selection works by removing irrelevant features and redundancy. The purpose of the feature selection is to get rid of the irrelevant and noisy genes from the input data set, to speed up the processing of data by reducing the dimensionality, and to avoid overfitting of the classifier [4]. While extraction features work by transforming the original data into a new representation. Fea-

ture 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 [9].

Both approaches have weaknesses and advantages. Such feature selection can be preserving data characteristics for interpretability, but discriminative power, lower shorter training times and reducing overfitting. While extraction features higher discriminating power, but a loss of data interpretability and transformation maybe expensive. To avoid change perhaps costly and loss of data interpretability, this research will propose feature selection to remove redundancy features from the data.

The proposed approach is to use clustering. In the research of Dewi Pramudi Ismi, Shireen Panchoo, and Murinto [10], the clustering approach can be used to remove feature redundancy by grouping features that have similarly on the same cluster. Then, after each group clustered the same cluster will be taken one sample to represent each cluster as a subset of features that will be used in the classification process.

The ranking system is used to select a sample that will represent each cluster. The ranking process is to know which feature has the highest correlation to the dataset class. This process design is expected to produce the absolute accuracy of the classification model being constructed.

1.3 Problem Statement

The following are some problems that become the background of this thesis:

1. Microarray data has a small number of samples and features very much. To perform cancer detection in the field of data mining or machine learning requires a lot of sample data for learning a model.

2. With the condition of high-dimensional microarray data and little sample data, that is the challenge in this study. Build models with little data, many features but still produce models with high accuracy.

3. The important step is preprocessing (feature selection process), so that duplicate or similarity features are not used in the model development process. The process of identifying duplicate features is also a stage of feature selection and a sensitive process.

1.4 Problem Limitations

According to problem statement, this thesis has limitations as follow:

1. Using three microarray data (Colon cancer, Lung cancer, and Tumor prostate) from `http://www.gems-system.org/` and each datasets have different sample and feature.

2. Colon cancer have 62 samples and 2000 features.
3. Lung cancer have 181 samples and 12533 features.
4. Tumor Prostate have 136 samples and 12600 features.
5. Focus on datasets that have two class.
6. Focus on redundancy feature

1.5 Objective

The following are some objective that become the background and problem statement of this thesis:

1. Build a model of detection cancer using microarray data (data has a small number of samples and features very much)
2. Removing redundancy feature (feature selection) using clustering approach by grouping features that have similarity in the same cluster, and then the features of each cluster are calculated correlation to the class.

1.6 Hypotheses

If the Random Forest algorithm has some advantages that can be used when it is more variables than observations, it fits perfectly with the characteristics of the microarray data. And if the random forest algorithm can find the critical feature of thousands of features then improving by removing redundancy before the classification process with clustering approach will result in a feature subset that has a high correlation to the class on the data. And when the classification process uses a subset of fewer and essential features. Thus, the process of classification with Random Forest produces a more accurate model.

1.7 Scope of Work

This research analyzes the performance of clustering approach on the microarray data classification using Random Forest algorithm implemented in a real-world dataset cancers. The work contains **two part** activities;

1. To monitor the expression profile of thousands of genes simultaneously, making it a possible tool to study transcriptome evolution.
2. Gene products will be identified which will be further help in the drug discovery to combat the cancer progressions.