CHAPTER 1:

THE PROBLEM

1.1. Rationale

Cancer is a term used for a type of diseases in which abnormal cells divide uncontrollably and are able to invade other tissues. Cancer cells can spread to other parts of the body through the blood and lymph systems [1]. According to data reported by the World Health Organization (WHO), Cancer is the leading cause of death worldwide, which is about 8.2 million deaths in 2012 and estimated will increase each year due to an unhealthy lifestyle [2].

Conventional methods for monitoring and diagnosing cancers are very dependent on human observation to detect certain features. A cancer diagnosis is usually performed using imaging systems such as X-ray, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasonography. Other conventional methods used in cancer diagnosis involves analysis of morphological and clinical data. In recent decades microarray take an important role in the diagnosis of cancer and improve the accuracy of cancer diagnosis compared to traditional techniques. By using microarray can be seen the level of gene expression in specific cell samples to analyze thousands of genes simultaneously [3].

The characteristic of microarray data is small sample but huge dimension. Since that, there is a challenge for researcher to provide solutions for microarray data classification with high performance both in accuracy and running time. In classification study, Artificial Neural Network (ANN) is one of popular method that gives satisfactory result. There are several algorithms to train ANN, one of the most popular algorithm is Back Propagation (BP).

Although BP is good algorithm for train ANN, but BP have several major deficiencies such as [7]: First, the BP algorithm will get trapped in local minima, this can lead to failure in finding a global optimal solution. Second, the convergence rate of BP is too slow even if learning can be achieved. Third, the convergence behavior of BP depends on the choices of learning rate in advance.

Many improvements were made to improve the performance of BP, one of them is to modify BP by implementing conjugate gradient algorithm in training BP [8][18]. By modifying BP into conjugate gradient. The search direction is not only decrease but also along conjugate directions. This is can produces faster convergence than BP generally [18]. A line search technique, such as secant method and backtracking, is performed to determine the search direction in conjugate gradient back propagation.

Later, as mentioned in [7][8][17] various conjugate gradient type were suggested to improve the efficiency of error minimization process or in other words the training efficiency. Among these types are Powel-Beale, Fletcher-Reeves, and Polak-Ribiere.

This study proposed a new framework for cancer detection based on microarray data classification using combination of Principal Component Analysis (PCA) and Conjugate Gradient Back Propagation.

1.2. Theoretical Framework

Since the mid-1990s, Patrick O. Brown, Joseph DeRisi, David Botstein, and colleagues, have developed DNA microarrays; and these microarrays have become a key tool in the fight against cancer [15]. Using machine learning, the detection of cancer can be done by classifying data into classes defined. The classification is the process of determining a class of data using methods such as Artificial Neural Networks. This study is presented an alternative framework for cancer detection based on classification microarray data. The proposed framework is developed using combination of PCA and conjugate gradient back propagation.

The dimension of microarray data is so huge, over thousands dimension, it condition can appear the problem called “The Curse of Dimensionality” if all dimension used in training of data [4]. The problem of “The Curse of Dimensionality” causes training of data requires long time and produces low accuracy of testing the framework. This study uses Principal Component Analysis (PCA) for reducing the dimensionality of the data. In its process, PCA tried to transform the high dimensional data into new coordinate system generated from linear combination of original data [4]. PCA uses eigenvector resulted by covarian matrix to represent data features. Some eigenvector with the largest eigenvalue, called principal components (PC), will be selected to be the new coordinate system of data in the database. These PC will be used for reducing dimension of data.

Conjugate gradient back propagation is modification of back propagation by implementing conjugate gradient algorithm in back propagation training [18]. BP has several major deficiencies such as [7]: First, the BP algorithm will get trapped in local minima, this can lead to failure in finding a global optimal solution. Second, the convergence rate of BP is too slow even if learning can be achieved. Third, the convergence behavior of BP depends on the choices of learning rate in advance.

The standard back propagation algorithm uses steepest descent algorithm to calculate search direction of new weights and uses static learning rate as step-size of direction. The steepest descent algorithm uses the most negative of gradient to be search direction. This is the direction in which the performance function is decreasing most rapidly. It turns out that, although the function decreases most rapidly along the negative of the gradient, this does not necessarily produce the fastest convergence [18]. By modifying search direction of standard back propagation using conjugate gradient method, the outcomes of the search direction is not only decrease but also along conjugate directions [13] [18].

1.3. Conceptual Framework

Variables that influence the result of this study are as follows.

1. Parameters of PCA which is number of principal components (PC).
2. Parameters of ANN which is number of hidden neurons
3. Additional parameters which are the type of conjugate gradient algorithm, Powell-Beale, Fletcher-Reeves, and Polak-Ribiere, and the line search technique, backtracking and secant method.

Figure 1 is schematic diagram of relationship of each parameter in the system.

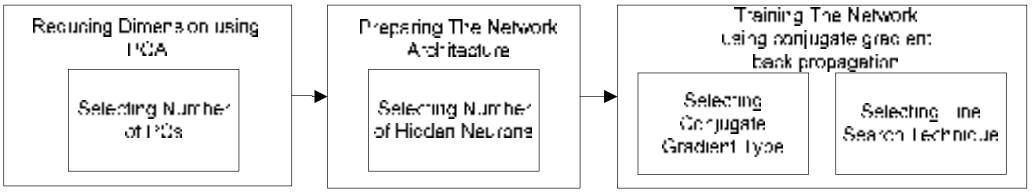

Figure 1 Schematic Diagram of Relationship of Each Parameter in The System

Number of PCs determined of how many samples of microarray data remaining after dimension reduction conducted. It means also become number of input of ANN training.

Furthermore, multi layer perceptron used as network architecture and one hidden layer was specified in this study. Number of hidden neurons were means number of neurons in hidden layer.

The last in training the network using back propagation modified by conjugate gradient algorithm, there are several type of conjugate gradient algorithm which were Powell-Beale, Fletcher-Reeves, and Polak-Ribiere, and several type of line search technique used for conjugate algorithm which are backtracking and secant method. This study was investigated the type of conjugate gradient algorithm and the type of line search technique.

1.4. Statement of the Problem

The characteristic of microarray data is small sample but huge dimension. Given data with these limitations, faced to the problem called "The Curse of Dimensionality" that can cause require long time for training a system and produces low accuracy. Therefore, it is a challenge for researchers to provide solutions for microarray data classification with high performance both in accuracy and training time.

Some literature about cancer detection based on microarray data have been reviewed, such as Bai [5] in his master thesis build a framework based on PCA and back propagation to classify microarray data. He was used some public microarray dataset, ovarian cancer dataset, leukemia dataset, and colon cancer dataset. He was straight divided the data into training data and testing data with ratio 2:1 and resulting 96.2% accuracy for ovarian data, 97.3% accuracy for leukemia data, and 95.02% accuracy for colon data. In terms of training time, Bai was admitted back propagation required long time for training, around 20-23 seconds for each data.

According to Bai's work, this study was presented an alternative framework for cancer detection based on microarray data classification. The proposed framework still use PCA as dimension reduction method but improving back propagation performance by implementing conjugate gradient algorithm in back propagation training.

The state of the art of this study is to improve back propagation training by implementing conjugate gradient algorithm. Conjugate gradient algorithm is a search algorithm in which the direction of the search is based on conjugate direction. In general, this algorithm is faster than the steepest descent method that used in back propagation [13].

The combination of PCA and conjugate gradient back propagation performance hopes able to outperform back propagation performance, both in training time and accuracy.

1.5. Hypothesis

The hypothesis of this study were proposed system require shorter time for training than back propagation system and the accuracy of proposed system is able to outperform back propagation system. To prove the hypothesis will be performed some experiments which are included proposed system and back propagation system.

1.6. Assumption

Raw microarrays data are glass microscope slides onto which genes are attached at fixed and ordered locations. Raw microarrays data have to go through a series of processes such as scanning, cleaning, and processing by special software to become matrix form data that easily to be analyzed using machine learning techniques. In this study is used public microarray data in matrix form taken from Kent-Ridge Biomedical Data Repository [14].

1.7. Scope and Delimitation

Some scopes and delimitations from the study are described below:

1. The conjugate gradient types used in this study were Powell-Beale, Fletcher-Reeves, and Polak-Ribiere.
2. The line search techniques used in this study were secant method and back-tracking.
3. Computer used to build and test system in this study is a personal computer.
4. The system is developed and simulated using MATLAB.
5. The performance measurements used in this study are accuracy and training time (ignoring big-o calculation of algorithm).

1.8. Importance of the Study

The research in the field of microarray data analysis is still very useful to do, it because can help doctors or pathologist to detect cancer. One of the research main focus of research in the field of microarray is to produce a fast and accurate cancer detection system. The system developed in this study are hopefully useful for similar studies to be developed in the future. In addition, the system can be the right choice for others researchers who will build fast and accurate cancer detection application.