International Journal of Software Science and Computational Intelligence
Volume 13 • Issue 1 • January-March 2021

Cancer Classification From DNA Microarray Using Genetic Algorithms and Case-Based Reasoning

Lilybert Machacha, Botho University, Gaborone, Botswana
Prabir Bhattacharya, Concordia University, Canada

ABSTRACT

There are many similarities in the symptoms of several types of cancer and that makes it sometimes difficult for the physicians to do an accurate diagnosis. In addition, it is a technical challenge to classify accurately the cancer cells in order to differentiate one type of cancer from another. The DNA microarray technique (also called the DNA chip) has been used in the past for the classification of cancer but it generates a large volume of noisy data that has many features, and is difficult to analyze directly. This paper proposes a new method, combining the genetic algorithm, case-based reasoning, and the k-nearest neighbor classifier, which improves the performance of the classification considerably. The authors have also used the well-known Mahalanobis distance of multivariate statistics as a similarity measure that improves the accuracy. A case-based classifier approach together with the genetic algorithm has never been applied before for the classification of cancer, same with the application of the Mahalanobis distance. Thus, the proposed approach is a novel method for the cancer classification. Furthermore, the results from the proposed method show considerably better performance than other algorithms. Experiments were done on several benchmark datasets such as the leukemia dataset, the lymphoma dataset, ovarian cancer dataset, and breast cancer dataset.

KEYWORDS

Case-Based Reasoning, Gene Expression, Genetic Algorithm, K-Nearest Neighbor, Mahalanobis Distance, Microarray

DOI: 10.4018/IJSSCI.2021010102

1. INTRODUCTION

Different types of cancer may have similar symptoms and an accurate classification of the type of cancer is thus necessary in order to treat a patient properly. Various cancer classification techniques have been developed in the past but most of them are based on the clinical analysis of morphological symptoms (Hong & Cho, 2004) and with such methods, even a trained specialist may make diagnostic errors. In order to overcome these problems, classification techniques using human gene information have been investigated (e.g., (Ben-Dor et al., 2000; Brazma & Vilo, 2000; Park & Cho, 2003)). Gene information (usually called the "gene expression data") could be collected by the DNA microarray technique (Amaratunga et al., 2014) and it provides useful information for the classification of different kinds of cancers. Since the original format of the data is an array of numbers, it is not easy to analyze them directly and discover useful classification rules. The DNA micro-array technology (Amaratunga et al., 2014) has been used to profile the global gene expression patterns of normal and transformed human cells in several types of cancers (Alizadeh et al., 2000; Alon et al., 1999; Bittner et al., 2000; Bubendorf et al., 1999; Golub et al., 1999; Perou et al., 2000). With the increase of cancer cases and its re-occurrence in many patients, it is clear that better and faster solutions are currently needed, which is the main motivation of our paper.

Microarray data is composed of many genes but very few samples; therefore to obtain many subsets of genes that can discriminate between different classes of samples is a multidimensional search problem. The *Mahalanobis distance* (e.g., Duda et al., 2001) is widely used as a multivariate outlier statistic for examining data profiles such as the learning curves, serial position effects, and group profiles, and it has a lesser confusion percentage as compared to the Euclidean distance (Campbell, 1997). The metric essentially addresses the question of whether a particular case would be considered an outlier relative to a particular set of group data. Clinicians usually compute the "z-scores" (see e.g., Mitchell, 1997, p. 235) to determine the percentile ranks (e.g., Li et al., 2000) and then correlate the client's scores with the mean scores for a selected group. The problem with this approach is that it incorporates only the group mean-values into the computation leaving the variability within each measure, and the correlations and variability between measures are not taken into account. In effect, correlation assumes that the measures in a profile are independent of each other.

1.1. Related Methods for Cancer Classification

Several methods for selecting a subset of discriminative genes for sample classification have been proposed (e.g., Brown et al., 2000; Bubendorf et al., 1999; Campbell, 1997; Cho & Won, 2003; Dasarathy, 1991; Duda et al., 2001; Dudoit et al., 2000; Eisen et al., 1998; Fix & Hodges, 1951) and these researchers applied the neighborhood analysis methods to identify a subset of genes using a separation measure similar to the t-statistic. Several classification methods (both supervised and unsupervised) were applied including the K-NN (without gene selection), and support vector machine (SVM) after gene selection. A *boosting technique* (Freud & Schapire, 1997) was used to search for a threshold (expression level) for each gene that would maximally discriminate between two types of samples (e.g. normal versus tumor). Several machine learning techniques have been used in classifying gene expression data, including the Fisher linear discriminant (Brazma & Vilo, 2000), K-nearest neighbor (Li, Weinberg, Darde et al, 2001), decision tree, multi-layer perceptron (Duda et al., 2001; Xu, Selaru, Yin, Zou, Shustova, & Mori, 2002), support vector machine (SVM) (Brown et al., 2000; Furey et al., 2000), boosting, and the self-organizing map Golub et al., 1999; Tamayo et al., 1999. Feature selection algorithms have been used widely in building CBR classifiers in the process of removing non-formative genes (Pedersen & Moult, 1996).

1.2. Main Contributions of This Paper

We propose an approach that combines two standard classifiers – Case-Based Reasoning (CBR) and K-Nearest Neighbor (K-NN) to improve the performance of cancer classification. We use Genetic