Chapter 4
Methods for Gene Selection and Classification of Microarray Dataset

Mekour Norreddine
Dr. Tahar Moulay University of Saida, Algeria

ABSTRACT

One of the problems that gene expression data resolved is feature selection. There is an important process for choosing which features are important for prediction; there are two general approaches for feature selection: filter approach and wrapper approach. In this chapter, the authors combine the filter approach with method ranked information gain and wrapper approach with a searching method of the genetic algorithm. The authors evaluate their approach on two data sets of gene expression data: Leukemia, and the Central Nervous System. The classifier Decision tree (C4.5) is used for improving the classification performance.

INTRODUCTION

DNA microarray technology is a revolutionary method enabling the measurement of expression levels of thousands of genes in a single experiment under diverse experimental conditions. Since its invention, this technology has proved to be a valuable tool for many biological and medical applications (Beatrice et al., 1999). Microarray data analysis can be carried out according to at least two different and complementary perspectives. In one hand, data clustering (non supervised classification) aims to identify groups of genes, or groups of experimental conditions that exhibit similar expression patterns. In such a context bi-clustering is particularly interesting since it allows the simultaneous identification of groups of genes that show similar expression patterns across specific groups of experimental conditions (samples) (Beatrice et al., 1999).

Nowadays, people can obtain the expression datasets of thousands of genes simultaneously using microarray technology. One of the important fields in using these gene expression datasets is to classify and predict the diagnostic category of a sample. Actually, precise diagnosis and classification is crucial

DOI: 10.4018/978-1-5225-3004-6.ch004

for successful treatment of illness (Zhen et al., 2009a). Knowledge Data Discovery (KDD) consists of several phases like Data selection, Data mining.

Data mining is one of the important phases of knowledge data discovery, There is a technique which is used to find new, hidden and useful patterns of knowledge from large databases. There are several data mining methods such as Prediction, Clustering and Classification (D.Lavanya et al., 2011).The problem of classification of data is identified as one of the major problems in extracting knowledge from data.

BIOLOGICAL BACKGROUND

Cells are the basic operating units of each living system. All the directions required to direct their actions are contained inside the chemical DNA or shortly deoxyribonucleic acid. A deoxyribonucleic acid molecule may be a double-stranded compound composed of 4 basic molecular units specifically nucleotides. The nitrogen bases include adenine (A), guanine (G), cytosine (C) and thymine (T).

The ordering provides a example for the synthesis of a range of ribonucleic acid molecules. the method of transcribing a gene's deoxyribonucleic acid sequence into ribonucleic acid is termed organic phenomenon. A gene's expression level indicates the approximate variety of copies that gene's ribonucleic acid created in a very cell and it's correlative with the quantity of the corresponding proteins created. This mechanism controls that genes are expressed in a very cell and acts as a "volume control" that will increase or decreases the extent of expression of explicit genes as necessary (Nagamma et al., 2013b).

MICROARAY DATA FORMAT

A gene expression data set from a microarray experiment can be represented by a real-valued.

Expression matrix = { G(i,j) | 1≤ i≤ n,1≤ j ≤ m }

where the columns G={ $\overleftarrow{g_1}$, $\overleftarrow{g_2}$,...., $\overleftarrow{g_m}$ } form the expression patterns of genes, the rows S={ $\overleftarrow{s_1}$, $\overleftarrow{s_2}$,...., $\overleftarrow{s_n}$ } . An example of a gene expression microarray dataset for Leukemia is shown (in Table 1). the table organizes data into m columns (genes) and n rows (samples) where m mostly varies from thousand to hundred thousand according to the accuracy of microarray image processing technique, while n is always less than 200 samples according to the previously collected datasets (Zeeshan et al., 2014a). Category column presents the actual class of the sample. For the shown example AML stands for acute myeloid leukemia disease and ALL represents acute lymphoblastic.

PROBLEM DEFINITION

The selection of attributes has become a very active research topic a few years in the fields of artificial learning (Jain et al.,1997; Dash et al.,1997; Kohavi et al.,1997; Blum et al.,1997; Duch et al.,2006) Data mining, image processing (Singh et al., 2002; Fukunaga et al.,1990) and data analysis in bioinformatics.