Chapter 70
Gene Selection from Microarray Data for Alzheimer's Disease Using Random Forest

Kazutaka Nishiwaki
Tokyo University of Science, Chiba, Japan

Katsutoshi Kanamori
Tokyo University of Science, Chiba, Japan

Hayato Ohwada
Tokyo University of Science, Chiba, Japan

ABSTRACT

A significant amount of microarray gene expression data is available on the Internet, and researchers are allowed to analyze such data freely. However, microarray data includes thousands of genes, and analysis using conventional techniques is too difficult. Therefore, selecting informative gene(s) from high-dimensional data is very important. In this study, the authors propose a gene selection method using random forest as a machine learning technique. They applied this method to microarray data on Alzheimer's disease and conducted an experiment to rank genes. The authors' results indicated some genes that have been investigated for their relevance to Alzheimer's disease, proving that their proposed cognitive method was successful in finding disease-related genes using microarray data.

INTRODUCTION

DNA microarray is widely used in the fields of medical science and biology. Since DNA microarray is able to measure expression levels of many genes at the same time, researchers can find gene(s) related to specific phenomena such as disease. Statistical analysis such as t-test or analysis of variance (ANOVA) is used to analyze such data. However, the human has, for example, 20,000 to 25,000 protein-coding genes (The International Human Genome Sequencing Consortium, 2004), which is too many to analyze using

DOI: 10.4018/978-1-7998-2460-2.ch070

conventional techniques. In addition, small sample size and high noise of gene expression data make analysis difficult. Thus, gene selection from gene expression data using machine learning techniques is gaining interest among researchers.

Gene selection is one type of feature selection for machine learning research, and many methods have been applied in this research field. A Bayesian model was applied for variable selection to identify important genes using their expression levels, and it was successful for cancer classification via cDNA microarrays and leukemia data (Lee et al., 2003). The Support Vector Machine (SVM) performs well even with high-dimensional data and can be applied for more than classification tasks (Imai et al., 2013; Kharrat and Abid, 2014). SVM was applied to microarray data related to cancer, and the top-ranked genes by SVM Recursive Feature Elimination (RFE) are related to cancer (Guyon et al., 2002). Ensemble learning methods have also been applied to feature selection with good results. Random forest (Breiman, 2001), a well-known learning technique of the ensemble method, was successful for not only classification and regression tasks but also feature selection tasks (Genuer et al., 2010; Wang and Yang, 2011). Some researchers used random forest successfully for gene selection and classification of gene expression data (Díaz-Uriarte and Alvarez de Andrés, 2006; Moorthy and Mohamad, 2011). Feature selection has been applied in many research fields and has provided new knowledge for data scientists, biologists, and cognitive computing researchers.

Random forest is an ensemble learning technique of machine learning that builds decision trees in a forest when it learns. The core idea of this method is tree bagging or bootstrap sampling, and a random subset of features. Random forest selects a random sample and features with replacement of the training data when it builds trees. In addition, random forest indicates the importance score of each feature for classification (decreasing information gain). Many studies on feature selection with random forest focus on and use this function.

Random forest performs well for gene selection. However, previous studies used well-known microarray data to bioinformatics researchers such as Leukemia (Golub et al., 1999) or NCI 60 (Ross et al., 2000) and it is unknown that random forest can select genes from other microarray data. In the present study, we used datasets obtained from a microarray database on the Internet. The objective of this study was to select genes from DNA microarray data obtained from public datasets. More specifically, we used microarray data related to Alzheimer's disease and looked for Alzheimer's disease-related genes. This paper presents our methodology to find disease-related genes using random forest.

DATASET

We obtained datasets of DNA microarray from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (Edgar et al., 2002). The Gene Expression Omnibus contains numerous datasets of DNA microarray obtained from different species and environments of experiments. Since we had to select datasets obtained from homo sapiens, we conducted a search for such datasets using the following steps:

1. We searched the phrase "Alzheimer's disease" and checked the box of "DataSets" to set Entry type, and limited the sample organism as "homo sapiens" on the Gene Expression Omnibus;
2. As a result, we obtained 13 papers discussing Alzheimer's disease, and we downloaded datasets of DNA microarray referenced in these papers;
3. We selected datasets including samples from human brain cells.