Expert Systems with Applications 38 (2011) 14065–14071

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A gene selection method for microarray data based on risk genes

Tzu-Tsung Wong *, Ding-Qun Chen

Institute of Information Management, National Cheng Kung University, 1, Ta-Sheuh Road, Tainan City 701, Taiwan, ROC

ARTICLE INFO

Keywords:
Cancer classification
Clustering analysis
Gene selection
Microarray data
Risk genes

ABSTRACT

Many gene selection methods have been proposed to select a subset of genes that can have a high prediction accuracy for cancer classification, and most set the same preference for all genes. However, many biological reports have pointed out that mutated or flawed genes, named as risk genes, can be one of the major causes of a specific disease. This study proposes a gene selection method based on the risk genes found in biological reports. The information provided by risk genes can reduce the time complexity for gene selection and increase the accuracy of cancer classification. This gene selection method is composed of two stages. Since all risk genes must be chosen, the first stage is to remove the genes that have similar expression levels or functions to risk genes. The next stage is to perform gene selection and gene replacement based on the results of a process that divides the remaining genes into clusters. Based on the test results from four microarray data sets, our gene selection method outperforms those proposed by previous studies, and genes that have the potential to be new risk genes are presented.

1. Introduction

In a microarray data set with thousands of genes, most items are irrelevant to the disease of interest. The occurrence of a disease is generally related to certain genes only when they act in combination with others. Many gene selection methods have been proposed to select a subset of genes that can have a high prediction accuracy for cancer classification. Several studies addressed the problem of selecting a proper gene subset for multiclass classification (Le Cao, Bonnet, & Gadat, 2009; Ooi, Chetty, & Teng, 2007; Shen, Shi, & Kong, 2009). Since the number of genes is huge with respect to the number of available microarray instances, dimension reduction techniques are developed to filter irrelevant and redundant genes for improving computational complexity and classification accuracy (Chuang, Ke, Chang, & Yang, 2009; Li, Wu, & Tan, 2008; Zeng, Li, Yang, Yang, & Wu, 2007; Zhu, Ong, & Dash, 2007). The correlation among genes can also provide useful information for gene selection in classifying microarray instances (Hu, Qiu, Glazko, Klebanov, & Yakovlev, 2009; Lee, Chang, & Chao, 2008; Song, Bedo, Borgwardt, Gretton, & Smola, 2007). All of those studies set the same preference for all genes. However, many biological reports have pointed out that mutations or flaws in genes can be one of the major causes of a specific disease.

Gormley, Dampier, Ertel, Karacali, and Tozeren (2007) used a Web tool supported by the Ingenuity Pathway Database to collect hundreds of genes that are susceptible to a specific disease. They then randomly selected some of these for classification. Their experimental results showed that the gene subset chosen by individual-gene-ranking methods outperforms the gene subset composed of randomly chosen susceptibility genes. Tai and Pan (2007) employed a similar approach to group hundreds of susceptibility genes according to their functional associations. They then proposed a gene selection method that integrates biological knowledge of gene functions for classifying microarray data. Their method achieved a better performance and improved the biological interpretation of the selected genes. In these two gene selection methods, the number of susceptibility genes is more than needed, meaning that some of the susceptibility genes may contain duplicated information for cancer classification.

A gene will be named as a risk gene if its mutation or flaw has been reported to be relevant to a disease in the biological literature. Unlike the previous two studies employing susceptibility genes, the number of risk genes will be far less than needed. The final gene subset for classification chosen by our gene selection method must include the risk genes identified for a disease. After the appropriate risk genes for a disease have been determined, we will employ gene removal, clustering analysis, gene ranking, and group gene selection to constitute a risk-gene-based (RGB) gene selection method.

This paper is organized as follows. The data mining tools that are employed in this study are briefly introduced in Section 2. Section 3 presents the risk genes found for breast and prostate cancer in the biological literature. The risk genes will be tested on four microarray data sets to show that a subset composed of only the risk genes cannot achieve a high prediction accuracy. Thus, in Section 4, we propose the RGB method for gene selection. The RGB method will then be tested on the same four microarray data sets

* Corresponding author. Tel.: +886 6 2757575x53722.
E-mail address: tzutsung@mail.ncku.edu.tw (T.-T. Wong).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.215

in Section 5 to analyze its computational efficiency and prediction accuracy. Finally, conclusions and the directions for future study of this paper are addressed in Section 6.

2. Data mining tools

The data mining tools that are employed in this study include the techniques for gene ranking, clustering analysis, and classification. We will briefly introduce them in this section.

2.1. Gene ranking methods

Lu and Han (2003) divided gene selection methods into two categories: individual gene ranking and gene subset ranking. Individual-gene-ranking methods calculate the correlation between each gene and the class value and select the genes that have correlations larger than a pre-specified threshold. This approach is usually simpler and more computationally efficient, but may exclude genes that are important for disease diagnosis only in combination with others. To overcome this deficiency, gene-subset-ranking methods employ a measure to evaluate the discernability of a gene subset on the class value. Since the number of possible gene subsets grows exponentially to the number of genes in a microarray data set, the computational cost of a gene-subset-ranking method is high; hence only limited gene subsets can be evaluated and ranked. Although individual-gene-ranking methods may also find the genes with redundant information for classification, they are still useful in identifying the genes relevant to the class value.

A microarray instance i with n genes and one class value can be represented by $(x_{i1}, x_{i2}, \ldots, x_{in}, y_i)$. Suppose that the class value is one if an instance comes from an abnormal tissue, and zero otherwise. Let N_i be the number of training instances with class c for $c = 0, 1$, and let $\bar{x}_{jc}$ and s^2_{jc} be the mean and the variance of gene j calculated from the training instances with class c, respectively. According to Nguyen and Rocke (2002), the t value of gene j is $t_j = (\bar{x}_{j0} - \bar{x}_{j1})/\sqrt{s^2_{j0}/N_0 + s^2_{j1}/N_1}$. A gene with a larger t value is more relevant to the class value.

Dudoit, Fridlyand, and Speed (2002) defined the BW ratio (the ratio of between-group sums of squares to within-group sums of squares) of gene j as follows:

$$BW_j = \frac{\Sigma_i \Sigma_c I(y_i = c)(\bar{x}_{jc} - \bar{x}_j)^2}{\Sigma_i \Sigma_c I(y_i = c)(x_{ij} - \bar{x}_{jc})^2},$$

where $\bar{x}_j$ is the global mean of gene j and $I(\cdot)$ is an indicator function with a value of 1 if the condition within I holds and 0 otherwise. A gene with a larger BW ratio can interpret a larger proportion of the variance of the class, which implies the gene is more relevant to the class value.

Su, Murali, Pavlovic, Schaffer, and Kasif (2003) presented a measure called information gain to determine the relevancy between a gene and the class value. The whole range of the expression data of gene j is first divided into two regions by the average of its maximum and minimum expression values. The region with expression values less than the average is marked as region 1, and the other as region 2. Let d_m be the number of expression values of gene j in region m, and let e_{km} be the number of instances with class c in region m. Then the information gain of gene j is calculated as

$$Gain_j = \sum_{c=0}^{1} \sum_{m=1}^{2} \frac{e_{cm}}{N} \log_2 \frac{e_{cm}}{d_m} - \sum_{k=0}^{1} \frac{e_{c1} + e_{c2}}{N} \log_2 \frac{e_{c1} + e_{c2}}{N}.$$

A gene with a larger information gain implies that it can provide more information to reduce the uncertainty in classification.

The gene-subset-ranking method proposed by Li, Darden, Weinberg, and Pedersen (2001a) is a combination of the genetic algorithm and the k-nearest neighbors. Every chromosome in the genetic algorithm contains a fixed number of genes, and the fitness of a chromosome is evaluated by its classification accuracy resulting from the k-nearest neighbors. A chromosome with a fitness larger than the pre-specified threshold is stored in a list. The genetic algorithm stops when the number of chromosomes in the list reaches a pre-specified number. The frequency of a gene is defined as the number of chromosomes in the list containing the gene. A gene with a larger frequency is considered to be more relevant to the occurrence of a disease.

2.2. Clustering methods

Clustering methods have been applied to analyze the similarity among genes in recent years. D'haeseleer (2005) summarized the frequently used clustering methods for gene expression data and pointed out the issues that may be encountered in the process. In this research, since the genes in a cluster will have only one representative chosen, and the other genes in this cluster will be removed for the sake of computational efficiency, the quality threshold (QT) method proposed by Heyer, Kruglyak, and Yooseph (1999) will be adopted for gene clustering.

There are two parameters for the QT method. The first one defines how similar two genes can be in the same cluster, and the other specifies the minimal number of genes in a cluster. Based on the first parameter, the QT method finds the clustering members for every gene, which means a gene can be in several clusters. Only the cluster with the largest number of genes that is not less the second parameter will be accepted as a cluster. The QT method is then iteratively applied on the genes that have not been assigned to a cluster yet to find the next largest cluster until either the cluster membership of every gene is determined or the numbers of genes in the remaining clusters are all less than the second parameter.

2.3. Classification methods

Most classification tools favor processing discrete attributes, such as naïve Bayesian classifiers and decision trees. Since the gene expression data are numeric, classification tools that favor continuous attributes will be more appropriate for applying to the microarray data. We therefore pick the k-nearest neighbors as the classification tool in this study.

The k-nearest neighbors will use the genes chosen by a gene selection method to find k training instances that are the closest to a new instance. When the k nearest training instances have the same class value, this value will be the predicted class of the new instance. Otherwise, the new instance will be unclassifiable, and it will not be counted in calculating the classification accuracy. This is called the consensus approach, as seen in Li, Weinberg, Darden, and Pedersen (2001b).

3. Risk genes

Since not the expression data of every human gene can be found in a microarray data set, we searched the Cancer GeneticWeb and used the Google search engine to filter candidate risk genes. After the survey, the risk genes for breast and prostate cancer are summarized in Table 1. The most well-known risk genes for breast cancer are BRCA1 (breast cancer 1, early onset) and BRCA2 (breast cancer type2 risk protein). Although HPC1 (hereditary prostate cancer 1) and HPC2 (hereditary prostate cancer, X-linked) are both well-known risk genes for prostate cancer, the expression data of gene HPC1 can be missing in a microarray data set. We therefore omitted it from the risk genes of prostate cancer.