Multiclass Molecular Classification

Chia Huey Ooi
Duke-NUS Graduate Medical School Singapore, Singapore

INTRODUCTION

Molecular classification involves the classification of samples into groups of biological phenotypes. Studies on molecular classification generally focus on cancer for the following reason: Molecular classification of tumor samples from patients into different molecular types or subtypes is vital for diagnosis, prognosis, and effective treatment of cancer (Slonim, Tamayo, Mesirov, Golub, and Lander, 2000). Traditionally, such classification relies on observations regarding the location (Slonim et al., 2000) and microscopic appearance of the cancerous cells (Garber et al., 2001). These methods have proven to be slow and ineffective; there is no way of predicting with reliable accuracy the progress of the disease, since tumors of similar appearance have been known to take different paths in the course of time.

With the advent of the microarray technology, data regarding the gene expression levels in each tumor sample may now prove to be a useful tool in molecular classification. This is because gene expression data provide snapshots of the activities within the cells and thus, the profile of the state of the cells in the tissue. The use of microarrays for gene expression profiling was first published in 1995 (Schena, Shalon, Davis, and Brown, 1995). In a typical microarray experiment, the expression levels of up to 10,000 or more genes are measured in each sample. The high-dimensionality of the data means that feature selection (FS) plays a crucial role in aiding the classification process by reducing the dimensionality of the input to the classification process. In the context of FS, the terms *gene* and *feature* will be used interchangeably in the context of gene expression data.

BACKGROUND

The objective of FS is to find from an overall set of N features, the subset of features, S, that gives the best classification accuracy. This feature subset is also known as the *predictor set*. There are two major types of FS techniques, filter-based and wrapper techniques. Filter-based techniques have several advantages over wrapper techniques:

a. Filter-based techniques are computationally less expensive than wrapper techniques.
b. Filter-based techniques are not classifier-specific; they can be used with any classifier of choice to predict the class of a new sample, whereas with wrapper-based techniques, the same classifier which has been used to form the predictor set must also be used to predict the class of a new sample. For instance, if a GA/SVM (wrapper) technique is used to form the predictor set, the SVM classifier (with the same classifier parameters, e.g., the same type of kernel) must then be used to predict the class of a new sample.
c. More importantly, unlike the typical 'black-box' trait of wrapper techniques, filter-based techniques provide a clear picture of why a certain feature subset is chosen as the predictor set through the use of scoring methods in which the inherent characteristics of the predictor set (and not just its prediction ability) are optimized.

Currently, filter-based FS techniques can be grouped into two categories: *rank-based selection* (Dudoit, Fridlyand, and Speed, 2002; Golub et al., 1999; Slonim et al., 2000; Su, Murali, Pavlovic, Schaffer, and Kasif, 2003; Takahashi & Honda, 2006; Tusher, Tibshirani, and Chu, 2001) and state-of-the-art *equal-priorities scoring methods* (Ding & Peng, 2005; Hall & Smith, 1998; Yu & Liu, 2004). This categorization is closely related to the two existing criteria used in filter-based FS techniques. The first criterion is called relevance – it indicates the ability of a gene in distinguishing among samples of different classes. The second criterion is called redundancy – it indicates the similarity between

pairs of genes in the predictor set. The aim of FS is to maximize the relevance of the genes in the predictor set and to minimize the redundancy between genes in the predictor set.

Rank-based selection methods use only relevance as the criterion when forming the predictor set. Each of the *N* genes in the dataset is first ranked based on a score which indicates how relevant the gene is (i.e., its ability to distinguish among different classes). The *P* top-ranked genes are then chosen as the members of the predictor set. The choice of the value *P* is often based on experience or some heuristics (Dudoit et al., 2002; Li, Zhang, and Ogihara, 2004).

Due to the need for fast and simple reduction of dimensionality for gene expression datasets, the most ample instances of existing filter-based FS techniques for molecular classification are those of the rank-based category. This is because rank-based techniques consider only one criterion in forming the predictor set, resulting in lower computational cost than more complex techniques where two criteria are considered. Compared to rank-based techniques, there are considerably fewer existing instances of equal-priorities scoring methods, which use two criteria in forming the predictor set: relevance and redundancy (Ding & Peng, 2005; Hall & Smith, 1998; Yu & Liu, 2004). More importantly, in these methods *equal priority* is assigned to each of the two criteria (relevance and redundancy), hence the term 'equal-priorities scoring methods'.

MAIN FOCUS

There are two areas of focus in this article. The first is the area of FS for gene expression data. The second is the area of molecular classification based on gene expression data.

In existing studies on filter-based FS, at most two criteria are considered in choosing the members of the predictor set: relevance and redundancy. Furthermore, even in studies where both relevance and redundancy are considered (Ding & Peng, 2005; Hall & Smith, 1998; Yu & Liu, 2004), both criteria are given *equal* weights or priorities. Based on a two-class example used in another study (Guyon & Elisseeff, 2003), we begin to ask the question if the two criteria should *always* be given equal priorities regardless of dataset characteristics, namely the number of classes. To find the answer, Ooi, Chetty, and Gondal (2004) introduced the concept of differential prioritization as a third criterion to be used in FS along with the two existing criteria of relevance and redundancy. The concept was then tested on various gene expression datasets (Ooi, Chetty, and Teng, 2006; Ooi, Chetty, and Teng, 2007b). Differential prioritization works better than existing criteria in FS by forming a predictor set which is most optimal for the particular number of classes in the FS problem (Ooi, Chetty, and Teng, 2007a).

In the area of molecular classification, there is a lack of formal approach for systematically combining the twin problems of FS and classification based on the decomposition paradigm used in each problem. A multiclass problem is a problem in which there are three or more classes. It can be *decomposed* into several two-class sub-problems. The number of derived two-class sub-problems will depend on the type of the *decomposition paradigm* used. The rationale for doing this is that the two-class problem is the most basic, and thus, the easiest of classification problems (divide-and-conquer strategy). Furthermore, many classifiers such as Support Vector Machine (SVM) (Vapnik, 1998) are originally devised for the two-class problem.

Predictor Set Scoring Method

A FS technique is made of two components: the predictor set scoring method (which evaluates the goodness of a candidate predictor set) and the search method (which searches the gene subset space for the predictor set based on the scoring method). The FS technique is wrapper-based when classifiers are invoked in the predictor set scoring method. Filter-based FS techniques, on the other hand, uses criteria which are not classifier-based in order to evaluate the goodness of the predictor set. The criteria are listed below:

1. **Relevance:** The relevance of a predictor set tells us how well the predictor set is able to distinguish among different classes. It is summarized in the form of the average of the correlation between a member of the predictor set and the target class vector, which, in turn, represents the relevance of the particular feature (Hall & Smith, 1998). The target class vector (consisting of class labels of the training samples) represents the target class concept. Relevance is to be maximized in the search for the predictor set.