Bayesian Classification of Tumors Using Gene Expression Data

Bani Mallick ¹, Debasis Ghosh ² and Malay Ghosh ³

¹ Department of Statistics, Texas A&M University
² Department of Statistics, University of Michigan
³ Department of Statistics, University of Florida

1 Introduction

Precise classification of tumors is critical for cancer diagnosis and treatment. Diagnostic pathology has traditionally relied on macro and microscopic histology and tumor morphology as the basis for tumor classification. Current classification frameworks, however, are unable to discriminate among tumors with similar histopathologic features, which vary in clinical course and in response to treatment. In recent years, there has been a move towards the use of cDNA microarrays for tumor classification. These high-throughput assays provide relative mRNA expression measurements simultaneously for thousands of genes. A key statistical task is to perform classification via different expression patterns. Gene expression profiles may offer more information than classical morphology and may potentially provide an alternative to classical tumor diagnosis schemes.

This paper considers several Bayesian classification methods for the analysis of microarray data based on reproducing kernel Hilbert spaces. We consider the logistic likelihood as well as likelihoods related to the Support Vector Machine (SVM) models. It is shown through simulation and examples that SVM models with multiple shrinkage parameters produce the least amount of misclassification errors in comparison to several existing classical methods as well as Bayesian methods based on the logistic likelihood or those involving only one shrinkage parameter.

2 Methods

Targeting specific therapies to pathogenetically distinct tumor types is important for cancer treatment because it maximizes efficacy and minimizes toxicity (Golub et al., 1999). Precise classification of tumors is of critical importance to cancer diagnosis and treatment. Diagnostic pathology has traditionally relied on macro and microscopic histology and tumor morphology as the basis for tumor classification. Current frameworks, however, are unable to discriminate among tumors with similar histopathologic features, which vary in clinical course and in response to treatment. Recently, there is increasing interest in changing the basis of tumor classification from morphologic to molecular using microarrays which provide expression measurements for thousands of genes simultaneously (Schena et al., 1995; DeRisi et al., 1997). A
The key goal for the use of microarray data is to perform classification via different expression patterns. Several studies have used microarrays to profile colon, breast and other tumors and have demonstrated the potential power of expression profiling for classification (Alon et al., 1999; Hedenfalk et al., 2001). Gene expression profiles may offer more information than classical morphology and provide an alternative to morphology-based tumor classification systems. We will be focusing on the classification of microarray data.

The main difficulty with microarray data analysis is that the sample size $n$ is very small when compared to the dimension of the problem (the number of genes) $p$. The number of genes on a single array are usually in the thousands, so the number of regressors $p$ easily exceeds the number of observations $n$. This is known as the “large $p$, small $n$” problem (West, 2002). In this situation, dimension reduction is needed to reduce the high-dimensional gene space. Most of the existing approaches perform a preliminary selection of genes based on some criterion and use only 5% to 10% of the total genes for classification purposes. In our classification approach we can utilize all the genes rather than eliminating most of them based on a crude criterion.

In this talk we will construct Bayesian binary classification models for prediction based on the theory of reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Parzen, 1970). The methods are quite general and, in particular, can be used for tumor classification. One of the nice properties of RKHS methods is that they project the prediction problem into a data space which is of dimension $n$, which is much smaller than $p$. Usually RKHS has been used in a decision theory framework as an optimization problem and there is usually no explicit underlying probabilistic model. Consequently, it is not possible to assess uncertainty either with the classifier itself or with predictions based on this classifier. Our goal is to present a full probabilistic model-based approach to RKHS based classification. First we will introduce the logistic classifier in this framework, and then extend it to support vector machine (SVM) classifiers (Christianini and Shawe-Taylor, 2000; Sollich, 2001). We will propose probabilistic version of support vector machine stated as complete SVM.

As with other regularization methods, there are smoothing or regularization parameter(s) which need to be tuned for efficient classification. One popular approach is to use generalized approximate cross validation (GACV) (Wahba et al., 2002) to tune the smoothing parameters. In this article we will develop a hierarchical model where the unknown smoothing parameter will be interpreted as a shrinkage parameter (Denison et al., 2002). We will assign a prior distribution to it and obtain its posterior distribution via the Bayesian paradigm. In this way, we not only obtain the point predictors but find as well the associated measures of uncertainty. Furthermore, we will extend the model to incorporate multiple smoothing parameters; this leads to a significant improvement in prediction for the examples considered.

One of the basic features of RVM is to introduce sparseness in the model by considering heavy-tailed priors, for example, double exponential, for the regression coefficients (Bishop and Tipping, 2000; Figueiredo, 2002). This opportunity exists also for the SVM as considered in this paper, even though the binary probabilities are then modeled differently. In fact, in our examples, with a Bayesian hierarchical setup the SVM shows more sparseness than the logistic models. Several authors exploited this sparseness property to select significant genes (Roth, 2002; Lee et al., 2002). Our main emphasis, however, is to obtain predictive distributions for future observations to be used for classification rather than direct estimation of the parameters. For more details about the paper look at Mallick et al., 2002.
3 Conclusions

We have proposed a reproducing kernel Hilbert space based classification method for microarray data. It is shown that these models in a Bayesian hierarchical setup with priors over the shrinkage (smoothing) parameters performed better than other popular classification methods. Also, multiple shrinkage models always appear to be superior to single parameter shrinkage models. With multiple shrinkage parameters, the regular Bayesian SVM model emerges as the winner in all the examples with the complete SVM finishing a close second all the time. However, the complete SVM provides a more formal probabilistic motivation for the use of SVM’s, and are more satisfactory from a Bayesian angle.

References


