Supervised harvesting of expression trees

ABSTRACT

Tree harvesting is a potentially useful tool for exploration of gene expression data and identification of interesting clusters of genes worthy of further investigation.

INTRODUCTION

Background In this paper we introduce 'tree harvesting' - a general method for supervised learning from gene expression data. The scenario is as follows. We have real-valued expression measurements for thousands of genes, measured over a set of samples. The number of samples is typically 50 or 100, but will be larger in the future. An outcome measurement is available for each sample, such as a survival time or cancer class. Our objective is to understand how the genes relate to the outcome. The generic problem of predicting an outcome measure from a set of features is called 'supervised learning'. If the outcome is quantitative, the term 'regression' is used; for a categorical outcome, 'classification'. There are many techniques available for supervised learning: for example, linear regression, discriminant analysis, neural networks, support vector machines, and boosting. However, these are not likely to work 'off the shelf', as expression data present special challenges. The difficulty is that the number of inputs (genes) is large compared with the number of samples, and they tend to be highly correlated. Hastie et al. describe one simple approach to this problem. Here we build a more ambitious model that includes gene interactions. Our strategy is first to cluster the genes via hierarchical clustering, and then to consider the average expression profiles from all of the clusters in the resulting dendrogram as potential inputs into our prediction model. This has two advantages. First, hierarchical clustering has become a standard descriptive tool for expression data (see, for example,), so by 'harvesting' its clusters, the components of our prediction model will be convenient for interpretation. Second, by using clusters as inputs, we bias the inputs towards correlated sets of genes. This reduces the rate of overfitting of the model. In fact we go further, and give preference to larger clusters, as detailed below. The basic method is described in the next section for a quantitative output and squared error. We then generalize it to cover other settings such as survival data and qualitative responses. Tree harvesting is illustrated in two real examples and a simulation study is described to investigate the performance of the method. Finally, we generalize tree harvesting further, allowing nonlinear expression effects.

CONCLUSION

Conclusions The tree harvest procedure is a promising, general method for supervised learning from gene expression data. It aims to find additive and interaction structure among clusters of genes, in their relation to an outcome measure. This procedure, and probably any procedure with similar aims, requires a large number of samples to uncover successfully such structure. In the real data examples, the method was somewhat hampered by the paucity of available samples. We plan to try tree harvesting on larger gene expression datasets, as they become available. We used a forward stepwise strategy involving sum and products of the average gene expression of chosen clusters. We chose this strategy because it produces interpretable, biologically plausible models. Other models could be built from the average gene expression of clusters, including tree-based models or boosting methods (see, for example, Friedman et al.). Additional data Additional data available with the online version of this article include clusters from the harvest model applied to lymphoma data.