Chapter 4.3
Biomedical Data Mining Using RBF Neural Networks

Feng Chu
Nanyang Technological University, Singapore

Lipo Wang
Nanyang Technological University, Singapore

INTRODUCTION

Accurate diagnosis of cancers is of great importance for doctors to choose a proper treatment. Furthermore, it also plays a key role in the searching for the pathology of cancers and drug discovery. Recently, this problem attracts great attention in the context of microarray technology. Here, we apply radial basis function (RBF) neural networks to this pattern recognition problem. Our experimental results in some well-known microarray data sets indicate that our method can obtain very high accuracy with a small number of genes.

BACKGROUND

Microarray is also called gene chip or DNA chip. It is a newly appeared biotechnology that allows biomedical researchers monitor thousands of genes simultaneously (Schena, Shalon, Davis, & Brown, 1995). Before the appearance of microarrays, a traditional molecular biology experiment usually works on only one gene or several genes, which makes it difficult to have a "whole picture" of an entire genome. With the help of microarrays, researchers are able to monitor, analyze and compare expression profiles of thousands of genes in one experiment.

On account of their features, microarrays have been used in various tasks such as gene discovery, disease diagnosis, and drug discovery. Since the end of the last century, cancer classification based on gene expression profiles has attracted great attention in both the biological and the engineering fields. Compared with traditional cancer diagnostic methods based mainly on the morphological appearances of tumors, the method using gene expression profiles is more objective, accurate, and reliable. More importantly, some types of cancers have subtypes with very similar appearances that are very hard to be classified by traditional methods. It has been proven that gene expression has a good capability to clarify this previously muddy problem.

Thus, to develop accurate and efficient classifiers based on gene expression becomes a problem of both theoretical and practical importance. Recent approaches on this problem include artificial neural networks (Khan *et al.,* 2001), support vector machines (Guyon, Weston, Barnhill, & Vapnik, 2002), k-nearest neighbor (Olshen & Jain, 2002), nearest shrunken centroids (Tibshirani, Hastie, Narashiman, & Chu, 2002), and so on.

A solution to this problem is to find out a group of important genes that contribute most to differentiate cancer subtypes. In the meantime, we should also provide proper algorithms that are able to make correct prediction based on the expression profiles of those genes. Such work will benefit early diagnosis of cancers. In addition, it will help doctors choose proper treatment. Furthermore, it also throws light on the relationship between the cancers and those important genes.

From the point of view of machine learning and statistical learning, cancer classification using gene expression profiles is a challenging problem. The reason lies in the following two points. First, typical gene expression data sets usually contain very few samples (from several to several tens for each type of cancers). In other words, the training data are scarce. Second, such data sets usually contain a large number of genes, for example, several thousands. That is, the data are high dimensional. Therefore, this is a special pattern recognition problem with relatively small number of patterns and very high dimensionality. To provide such a problem with a good solution, appropriate algorithms should be designed.

In fact, a number of different approaches such as k-nearest neighbor (Olshen and Jain, 2002), support vector machines (Guyon *et al.,*2002), artificial neural networks (Khan *et al.*, 2001) and some statistical methods have been applied to this problem since 1995. Among these approaches, some obtained very good results. For example, Khan *et al.* (2001) classified small round blue cell tumors (SRBCTs) with 100% accuracy by using 96 genes. Tibshirani *et al.* (2002) successfully classified SRBCTs with 100% accuracy by using only 43 genes. They also classified three different subtypes of lymphoma with 100% accuracy by using 48 genes. (Tibshirani, Hastie, Narashiman, & Chu, 2003)

However, there are still a lot of things can be done to improve present algorithms. In this work, we use and compare two gene selection schemes, i.e., principal components analysis (PCA) (Simon, 1999) and a t-test-based method (Tusher, Tibshirani, & Chu, 2001). After that, we introduce an RBF neural network (Fu & Wang, 2003) as the classification algorithm.

MAIN THRUST

After a comparative study of gene selection methods, a detailed description of the RBF neural network and some experimental results are presented in this section.

Microarray Data Sets

We analyze three well-known gene expression data sets, i.e., the SRBCT data set (Khan *et al.,* 2001), the lymphoma data set (Alizadeh *et al.,* 2000), and the leukemia data set (Golub *et al.,* 1999).