New Ensemble Learning Approaches for Microarray Data Classification

Ching-Wei Wang
University of Lincoln, UK

INTRODUCTION

A reliable and precise classification of tumors is essential for successful treatment of cancer. Microarray technologies allow the rapid and comprehensive assessment of the transcriptional activity of a cell, leading to a more comprehensive understanding of the molecular variations among tumors and, hence, to a finer informative classification. However, the major challenge in using this technology is the analysis of its massive data output, which requires powerful computational means for interpretation. Hence, the ability to interpret the information in gene expression data becomes a critical issue in Genetics.

One of the most active areas of research in supervised machine learning has been to study methods for constructing good ensembles of classifiers. The main discovery is that the ensemble classifier often performs much better than single classifiers that make them up. Recent researches (Dettling, 2004, Tan & Gilbert, 2003) have confirmed the utility of ensemble machine learning algorithms for gene expression analysis. The motivation of this work is to investigate a suitable machine learning algorithm for classification and prediction on gene expression data.

The research starts with analyzing the behavior and weaknesses of three popular ensemble machine learning methods—Bagging, Boosting, and Arcing—followed by presentation of a new ensemble machine learning algorithm. The proposed method is evaluated with the existing ensemble machine learning algorithms over 12 gene expression datasets (Alon et al., 1999; Armstrong et al., 2002; Ash et al., 2000; Catherine et al., 2003; Dinesh et al., 2002; Gavin et al., 2002; Golub et al., 1999; Scott et al., 2002; van 't Veer et al., 2002; Yeoh et al., 2002; Zembutsu et al., 2002). The experimental results show that the proposed algorithm greatly outperforms existing methods, achieving high accuracy in classification.

The outline of this chapter is as follows: Ensemble machine learning approach and three popular ensembles (i.e., Bagging, Boosting, and Arcing) are introduced first in the Background section; second, the analyses on existing ensembles, details of the proposed algorithm, and experimental results are presented in Method section, followed by discussions on the future trends and conclusion.

BACKGROUND

Ensemble methods are learning algorithms that construct a set of base classifiers and then classify new data points by taking a vote of their predictions. The spirit in ensemble machine learning is to combine a number of rough "rules-of-thumb" into a more accurate aggregate class prediction rule. The learning procedure for ensemble algorithms can be divided into the following two parts. The first stage is constructing base classifiers/base models. The main tasks of this division are (1) data processing: prepare the input training data for building base classifiers by perturbing the original training data; and (2) base classifier constructions: build base classifiers on the perturbed data with a learning algorithm as the base learner. In this project, the C4.5 decision tree algorithm (Quinlan, 1996) is employed as the base learner. The second stage is voting, which combines the base models built in the previous stage into the final ensemble model. There are various kinds of voting systems. Two main voting systems are generally utilized; namely, weighted voting and unweighted voting. In the weighted voting system, each base classifier holds different voting power. On the other hand, in the unweighted system, an individual base classifier has equal weight, and the winner is the one with the most number of votes. Figure 1 illustrates the structure of the ensemble machine learning approach.

There are three types of generally adopted ensembles (i.e., Bagging, Boosting, and Arcing). Bagging algorithm, which is introduced by Breiman (1996), constructs base classifiers with inputs generated by the bootstrapping technique. The construction process of every base classifier is independent of each other. It perturbs the training set repeatedly to generate multiple predictors and combines these base classifiers by simple voting (classification) or averaging (regression) in order to obtain an aggregated predictor. The multiple input data for building base classifiers is formed by bootstrapping replicates of the original learning data.

Boosting was introduced by Schapire (1990) as a method to enhance the performance of a weak learning algorithm. Freund and Schapire (1996) proposed an algorithm called AdaBoost. There are lots of varieties of Boosting algorithms, and AdaBoostM1 is chosen as the Boosting method used in this project. Boosting adaptively reweights the training set in a way based on an error rate of the previous base classifier. The Boosting algorithm improves its behavior in reflection to the latest faults it makes. Moreover, if the error rate of a base classifier is greater than 0.5 or equal to 0, the sequential construction of base classifiers stops.

The framework of Arcing introduced by Breiman (1998) is similar to the one employed in Boosting. They both proceed in sequential steps. The major difference between Arcing and Boosting is that Arcing improves its behavior based on the accumulation of its faults in history. It examines all previous base classifiers' faults for construction of a new base classifier, while Boosting only checks the previous one base classifier. Apart from this, Arcing adopts unweighted voting system, whereas Boosting uses weighted voting. In addition, unlike Boosting, no checking procedure exists through the constructions of base classifiers.

METHOD

In this section, the behavior and weaknesses of the existing ensembles are first analyzed, followed by presentation of the design and detailed algorithm of the proposed approach.

Behavior and Weaknesses of Existing Ensembles

1. **Boosting.** The accuracy of Boosting models will remain the same after specific numbers of base models are established, due to the checking mechanism following each construction of base classifiers. The specific criterion in Boosting stops further construction of base classifiers while its error rate is equal to 0 or greater than 0.5 (see Figure 2). Therefore, if the sequential construction halts after building six base classifiers, the same result will be obtained on evaluating over Boosting

Figure 1. Ensemble machine learning algorithm