ISMB 2003 Abstract

Multi-class Protein Fold Classification using an Integrative Machine Learning Approach

Aik Choon TAN[1], David GILBERT[1], Yves DEVILLE[2]
[1]Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, 17 Lilybank Gardens, G12 8QQ Glasgow, United Kingdom
{actan, drg}@brc.dcs.gla.ac.uk
[2]Department of Computing Science and Engineering, Université catholique de Louvain, Place Sainte Barbe, 2 B-1348 Louvain-la-Neuve, Belgium
deville@info.ucl.ac.be

One of the current research trends in machine learning applied to bioinformatics is to combine several sophisticated learning algorithms in order to increase a classifier's predictive accuracy (credibility) and its explanatory power (comprehensibility). When trying to learn from large and diverse data sets (e.g. biological databases) it is important to produce hypotheses that encapsulate all the information from different sources. The classifiers that are used to characterise and/or classify the data must be accurate and easily understandable by the human expert. Most methods in bioinformatics only concentrate on the classifier's credibility and less often emphasise its comprehensibility.

For some multi-class classification problems (e.g. C1, C2, C3, …, Cn), the set of positive examples (C1) is very small compared to the set of negative examples (C2, C3,…,Cn); this is the common scenario in the functional annotation problem where there exist a lot of classes but the number of the examples (proteins) in each class is relatively low. This imbalanced proportion of examples in each class contributes to the poor performance of standard machine learning techniques (e.g. decision trees). Existing machine learning approaches tend to produce a strong discrimination classifier (high accuracy) with very low sensitivity (also called completeness) when learning on these types of problem.

The aim of this research is to construct a novel approach to integrate rules/patterns induced from multi-class and unbalanced data sets; and to demonstrate its usefulness in biological data. Specifically this method has been designed to increase the sensitivity of the classifiers. However, one consequence of this approach is the decrease in classifier's specificity.

We have applied this method to multi-class protein fold classification. The data set contains 700 examples for 27 SCOP folds. We showed that this approach is useful when the ratio of positive/negative examples is very low, and when the initial classifiers yield little sensitivity. In this case, the loss of specificity is small compared to the increase of sensitivity, yielding more useful classifiers. We are now working on improving the specificity of the integrated classifiers.

50 words
We devised a novel approach to integrate rules induced from multi-class and unbalanced data sets; and to demonstrate its usefulness to multi-class protein fold classification which contains 700 examples for 27 SCOP folds. We showed that this approach increases the sensitivity of the classifiers and yielding more useful classifiers.