

Additional Paper for “Biological Data Mining”

Paper reference

A Clare and R D King. Machine Learning of Functional class from Phenotype Data. *Bioinformatics*, 18(1):160-166, 2002.

Abstract

Clare and King applied supervised machine learning to *Saccharomyces cerevisiae* phenotype to detect the functional class of unknown Open Reading Frames (ORFs). They use data from three databases: TRIPLES¹, EUROFAN², and MIPS³ to train their model, which is based on the C4.5 algorithm. Using their model they were able to correctly classify unknown yeast ORFs as well determine relationships that could help in biologist’s experimental design.

Clare and King use a modified version of the C4.5 algorithm as their learning system, which generates decision trees or sets of rules. They extended the algorithm to handle the problem that ORFs may have more than one function, and therefore, more than one label. Generally, decision trees do not provide meaningful rules under the condition that child classes within a class hierarchy are sparsely populated. Clare and King decided to use a bootstrap approach that employed 500 random samples from the training a training set of 1461 classified ORFs taken from the TRIPLES, EUROFAN, and MIPS databases. C4.5 was run on each of the 500 samples, generating 500 rule-sets. Those rules that appeared across most of the rule-sets were considered the most reliable. Clare and King found that useful rules were produced at levels 1 and 2 in the hierarchy, but not at levels 3 and 4. Consequently, their approach only found simple rules to be reliable. For example, their approach found that:

if the ORF deletant is sensitive to calcofluor white
and the ORF deletant is sensitive to zymolyase
then its class is “biogenesis of cell wall (cell envelope)”

This particular rule had a mean accuracy of 90.9%. However, the combination of the two statements provided much better discrimination than just the first, whose mean accuracy in predicting cell wall biogenesis was only 43.8%.

The rules learned by their approach were able to predict the function of 83 ORFs of unknown function with an estimated accuracy of 80%. Furthermore, since the rules learned by the decision tree usually had a straightforward relationship to biology, they provide useful experimental design information for biologists such as “which media provide the most discrimination between functional classes”. For example, the above example rule shows that if you want to determine whether or not a gene is involved in cell wall biogenesis, these two sensitivity assays are sufficient to provide a high level of certainty; however, the calcofluor white assay alone is insufficient. Clare and King point out that their method could be used for other organisms with a large amount of genomic data.

Discussion

Like both papers presented in class (Wieser et al and Pappa et al) Clare and King used the C4.5 algorithm to analyze a database of biological data. Like the Pappa et al. paper Clare and King use the decision tree algorithm to classify unclassified instances of the object being classified. Whereas, Pappa et al were looking for a highly specific type of protein, Clare and King were looking at a much larger class of labels – a problem perhaps less suited to decision trees. However, Clare and King were able to find meaningful sets of experiments for biologists, which is quite impressive even though they were very simple. Wieser et al were filtering erroneous protein annotations using this method, which suggests that the same thing could be done for gene annotations in yeast databases that are determined by automatic programs like Clare and King’s, especially since there is so much yeast genomic data.

¹ Data generated by randomly inserting transposons into the yeast genome.

² Library of yeast ORF deletion mutants.

³ Catalogue of yeast phenotype data.