Hierarchical Multi-classification with Predictive Clustering Trees in Functional Genomics

Jan Struyf[1], Sašo Džeroski[2], Hendrik Blockeel[1], and Amanda Clare[3]

[1] Katholieke Universiteit Leuven, Dept. of Computer Science,
Celestijnenlaan 200A, B-3001 Leuven, Belgium
{Jan.Struyf, Hendrik.Blockeel}@cs.kuleuven.be
[2] Jozef Stefan Institute, Dept. of Knowledge Technologies,
Jamova 39, 1000 Ljubljana, Slovenia
Saso.Dzeroski@ijs.si
[3] The University of Wales, Aberystwyth, Dept. of Computer Science,
Penglais, Aberystwyth, Ceredigion, SY23 3DB, Wales, UK
afc@aber.ac.uk

Abstract. This paper investigates how predictive clustering trees can be used to predict gene function in the genome of the yeast *Saccharomyces cerevisiae*. We consider the MIPS FunCat classification scheme, in which each gene is annotated with one or more classes selected from a given functional class hierarchy. This setting presents two important challenges to machine learning: (1) each instance is labeled with a set of classes instead of just one class, and (2) the classes are structured in a hierarchy; ideally the learning algorithm should also take this hierarchical information into account. Predictive clustering trees generalize decision trees and can be applied to a wide range of prediction tasks by plugging in a suitable distance metric. We define an appropriate distance metric for hierarchical multi-classification and present experiments evaluating this approach on a number of data sets that are available for yeast.

1 Introduction

Saccharomyces cerevisiae (baker's or brewer's yeast) is one of biology's classic model organisms, and has been the subject of intensive study for years. Its genes have annotations provided by the Munich Information Center for Protein Sequences (MIPS) under their FunCat scheme for classifying the functions of the products of genes. FunCat is a hierarchical system of functional classes. A small part of this hierarchy is shown in Fig. 1. Many yeast genes are annotated with more than one functional class.

This classification setting presents two main challenges to machine learning: (1) each instance (gene) is labeled with a set of classes instead of just one class, and (2) the classes are structured in a hierarchy; ideally the learning algorithm should also take this hierarchical information into account.

A simple approach is to ignore the hierarchy and to learn separate models for each class (indicating whether a single instance belongs to the class or not).

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 272–283, 2005.

(a)
```
1 METABOLISM
1/1 amino acid metabolism
1/2 nitrogen and sulfur metabolism
...
2 ENERGY
2/1 glycolysis and gluconeogenesis
...
```

(b)

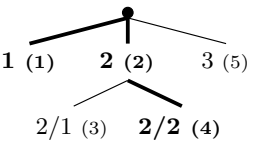

Fig. 1. (a) A part of the hierarchical FunCat classification scheme. (b) A toy hierarchy that will be used as example throughout the text. (Note that the class labels indicate the position in the hierarchy: 2/1 is a subclass of the class 2.)

In this work, we consider instead the task of learning one model for all classes. This has the advantage that the total size of the predictive theory is typically smaller, and that dependencies between different classes w.r.t. membership can be taken into account and may even be explicitated. Advantages of learning a single model for multiple related prediction tasks have been reported several times in the literature (see e.g., [6] for decision trees, [8,1] for neural networks, [15] for text classification).

Taking into account the hierarchical structure of the classes is also important while learning. The hierarchy concisely conveys relevant information about the similarity and differences between classes and also expresses the constraint that an instance belonging to a class also belongs to the parent class. The combination of multi-classification and hierarchical classification is known as hierarchical multi-classification [3].

Blockeel et al. [3] show how predictive clustering trees (PCTs) can be applied to hierarchical multi-classification. PCTs form a very general framework for prediction that can be instantiated to a particular prediction task by defining a distance metric and prototype. The distance metric used in [3] is a generic distance between sets of classes that is subsequently instantiated for hierarchical classification by plugging in the weighted shortest path distance between individual classes. The set distance has however two major disadvantages: (1) the distance between two given sets is difficult to interpret (it involves the computation of a kernel), and (2) it is not guaranteed to be positive because the kernel matrix is not positive definite. The impact of the latter is difficult to evaluate.

In this work we take an approach similar to that of [3]. We also use PCTs, but we introduce a new distance metric that is specific to hierarchical multi-classification and that does not have the disadvantages of the distance metric used in [3].

Recently, an extension to C4.5 [11] has been introduced by Clare [9] that is also capable of hierarchical multi-classification. This is accomplished by adapting the definition of entropy to take into account both the multi-class aspect and the hierarchical relationship between the classes. In the experimental evaluation included in this paper we compare our approach to the results presented in [9].

This paper is organized as follows. We first define hierarchical multi-classification more formally (Section 2). Then follows a brief description of PCTs (Section 3). Section 4 shows how PCTs can be instantiated for hierarchical multi-classification. This approach is validated experimentally in Section 5. Section 6 discusses further work, and Section 7 states the main conclusions.