Available online at www.sciencedirect.com

ScienceDirect

Data & Knowledge Engineering 61 (2007) 540–553

DATA & KNOWLEDGE ENGINEERING

www.elsevier.com/locate/datak

Utilizing hierarchical feature domain values for prediction ☆

Yiqiu Han, Wai Lam *

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong

Received 24 June 2006; accepted 27 June 2006
Available online 25 July 2006

Abstract

We propose a Bayesian learning framework which can exploit hierarchical structures of discrete feature domain values to improve the prediction performance on sparse training data. One characteristic of our framework is that it provides a principled way based on mean–variance analysis to transform an original feature domain value to a coarser granularity by exploiting the underlying hierarchical structure. Through this transformation, a tradeoff between precision and robustness is achieved to improve the parameter estimation for prediction. We have conducted comparative experiments using three real-world data sets. The results demonstrate that utilizing domain value hierarchies gains benefits for prediction.

Keywords: Hierarchical domain value; Machine learning; Classification and prediction

1. Introduction

Hierarchical structures are commonly found in data sets of various applications. In this paper, we deal with supervised learning problems with one extra property: the features used to describe examples have a hierarchical structure. We refer *hierarchical feature* to a feature with categorical values organized in a hierarchical structure. For example, the features handled by "drill down" or "scroll up" operations in Online Analytical Processing (OLAP) are basically hierarchical features. Such a feature hierarchy reflects existing knowledge about feature values and reveals their inter-relationships in different levels. We can exploit this sort of underlying knowledge for classification and numeric prediction.

Consider a "protein class" feature in the biomedical domain as shown in Fig. 1, it may take on the values "alpha subunits", "beta subunits", "gamma subunits", "RGS", "RAS", and so on. In this specific application, each of the values mentioned is a specialized concept of "trimeric GTP-proteins", which is in turn a specialized concept of "GTP-binding proteins". There is a known tree structure imposed on feature values, and it

☆ The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Nos.: CUHK 4179/03E and CUHK 4193/04E), the Direct Grant of the faculty of Engineering, CUHK (Project code: 2050363), and CUHK Strategic Grant (No.: 4410001). This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies.

* Corresponding author. Tel.: +852 2609 8306; fax: +852 2603 5505.
E-mail addresses: yqhan@se.cuhk.edu.hk (Y. Han), wlam@se.cuhk.edu.hk (W. Lam).

0169-023X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2006.06.018

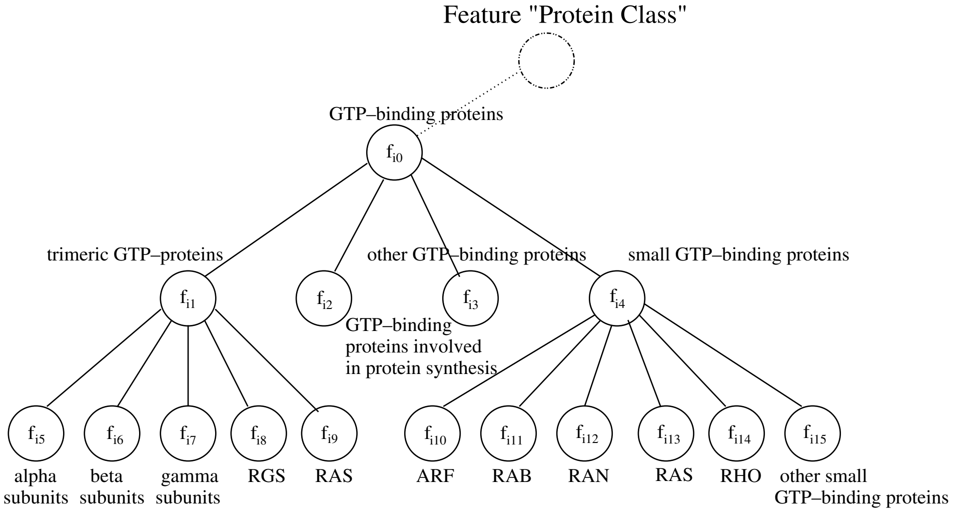

Fig. 1. One branch of the domain value hierarchy of the “protein class” feature.

can be exploited in cases where, for example, data is sparse, by replacing values such as “RGS” with a label higher in the tree in order to make better use of existing data. This would have the effect of allowing us to turn those examples with the feature labeled “RGS”, “RAS”, or “beta subunit” into examples where the feature has the common value “trimeric GTP-proteins” or “GTP-binding proteins”.

In many data sets, only the most atomic feature values, which correspond to leaf nodes in a hierarchy, are used. In practice, however, it may be difficult and quite expensive to precisely collect the most atomic feature values for every data record. The information of a feature value can be incomplete, for instance, only known to be among a set of values. This can be a challenge to many classical learning models.

Another challenge is the data sparseness problem which arises as the number of valid domain values becomes large for a hierarchical feature. Common maximum likelihood estimation for model parameters might be unreliable since a large domain of feature values reduces the effective amount of samples used to estimate those parameters. This situation becomes worse if the data set used for training is already sparse. The key to solve this problem may lie in the hierarchy itself. As stated above, the hierarchy in fact reflects existing knowledge about feature values, revealing their inter-relationships in different levels. When the training data is too sparse to estimate parameters associated with some feature values, one straightforward method is to utilize the global data distribution for estimation. However, this approach actually discards all information contained by those feature values. The hierarchical structure among feature values provides a means to find a favorable tradeoff between two primary learning goals, i.e., to utilize the information of feature values as much as possible, and to obtain reliable estimates.

The hierarchy of class labels is commonly utilized to divide the classification task into a sequence of sub-class classification tasks where each sub-class is associated with a certain node in the class label hierarchy. It has already been shown that hierarchical structure of feature domain values is helpful in OLAP. Utilizing hierarchical features in prediction problems has also been explored. Forman [5] engineered additional hierarchy prevalence features for handling the hierarchical features in the data set of KDD Cup 2002 Task 2. Svatek [16] has discussed in his work the impact of feature-value hierarchies on learning rules. In this method, abstract terms are introduced into the premises of rules and one-shot classification can be shifted to hierarchical refinement of the conclusion with the hierarchies of goal classes. An objective function of hierarchical clustering is developed for constructing man-made feature value hierarchies. desJardins [2] explored the utilization of feature value hierarchies in Bayesian Network learning model and obtained better scoring networks that also have better generalization performance.

Zhang and Honavar [19] addressed the problem of learning from user-supplied attribute value taxonomies (AVT) using the decision tree algorithm. Their algorithm works in a top-down fashion starting at the root of each AVT and builds a decision tree that uses the sufficiently informative abstraction level of attribute values