Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data

Abstract. In many application domains, there is a need for learning algorithms that can effectively exploit attribute value taxonomies (AVT)—hierarchical groupings of attribute values—to learn compact, comprehensible and accurate classifiers from data—including data that are partially specified. This paper describes AVT-NBL, a natural generalization of the naïve Bayes learner (NBL), for learning classifiers from AVT and data. Our experimental results show that AVT-NBL is able to generate classifiers that are substantially more compact and more accurate than those produced by NBL on a broad range of data sets with different percentages of partially specified values. We also show that AVT-NBL is more efficient in its use of training data: AVT-NBL produces classifiers that outperform those produced by NBL using substantially fewer training examples.

Keywords. Attribute value taxonomies · AVT-based naïve Bayes learner · Partially specified data

1 Introduction

Synthesis of accurate and compact pattern classifiers from data is one of the major applications of data mining. In a typical inductive learning scenario, instances
to be classified are represented as ordered tuples of attribute values. However, attribute values can be grouped together to reflect assumed or actual similarities among the values in a domain of interest or in the context of a specific application. Such a hierarchical grouping of attribute values yields an attribute value taxonomy (AVT). Such AVTs are quite common in biological sciences. For example, the Gene Ontology Consortium is developing hierarchical taxonomies for describing many aspects of macromolecular sequence, structure and function [5]. Undercoffer et al. have developed a hierarchical taxonomy that captures the features that are observable or measurable by the target of an attack or by a system of sensors acting on behalf of the target [41]. Several ontologies being developed as part of the Semantic Web-related efforts [7] also capture hierarchical groupings of attribute values. Kohavi and Provost have noted the need to be able to incorporate background knowledge in the form of hierarchies over data attributes in e-commerce applications of data mining [25, 26]. Against this background, algorithms for learning from AVT and data are of significant practical interest for several reasons:

(a) An important goal of machine learning is to discover comprehensible, yet accurate and robust, classifiers [34]. The availability of AVT presents the opportunity to learn classification rules that are expressed in terms of abstract attribute values leading to simpler, accurate and easier-to-comprehend rules that are expressed using familiar hierarchically related concepts [25, 44].

(b) Exploiting AVT in learning classifiers can potentially perform regularization to minimize overfitting when learning from relatively small data sets. A common approach used by statisticians when estimating from small samples involves shrinkage [29] to estimate the relevant statistics with adequate confidence. Learning algorithms that exploit AVT can potentially perform shrinkage automatically, thereby yielding robust classifiers and minimizing overfitting.

(c) The presence of explicitly defined AVTs allows specification of data at different levels of precision, giving rise to partially specified instances [45]. The attribute value of a particular attribute can be specified at different levels of precision in different instances. For example, the medical diagnostic test results given by different institutions are presented at different levels of precision. Partially specified data are unavoidable in knowledge acquisition scenarios that call for integration of information from semantically heterogeneous information sources [10]. Semantic differences between information sources arise as a direct consequence of differences in ontological commitments [7]. Hence, algorithms for learning classifiers from AVT and partially specified data are of great interest.

Against this background, this paper introduces AVT-NBL, an AVT-based generalization of the standard algorithm for learning naïve Bayes classifiers from partially specified data. The rest of the paper is organized as follows: Sect. 2 formalizes the notions on learning classifiers with AVT taxonomies; Sect. 3 presents the AVT-NBL algorithm; Sect. 4 discusses briefly on alternative approaches; Sect. 5 describes our experimental results and Sect. 6 concludes with summary and discussion.