RSD: Relational Subgroup Discovery through First-Order Feature Construction

Nada Lavrač¹, Filip Železný², and Peter A. Flach³

¹ Institute Jožef Stefan, Ljubljana, Slovenia
nada.lavrac@ijs.si
² Czech Technical University, Prague, Czech Republic
zelezny@fel.cvut.cz
³ University of Bristol, Bristol, UK
peter.flach@bristol.ac.uk

Abstract. Relational rule learning is typically used in solving classification and prediction tasks. However, relational rule learning can be adapted also to subgroup discovery. This paper proposes a propositionalization approach to relational subgroup discovery, achieved through appropriately adapting rule learning and first-order feature construction. The proposed approach, applicable to subgroup discovery in individual-centered domains, was successfully applied to two standard ILP problems (East-West trains and KRK) and a real-life telecommunications application.

1 Introduction

Developments in descriptive induction have recently gained much attention. These involve mining of association rules (e.g., the APRIORI association rule learning algorithm [1]), subgroup discovery (e.g., the MIDOS subgroup discovery algorithm [22]), symbolic clustering and other approaches to non-classificatory induction.

The methodology presented in this paper can be applied to relational subgroup discovery. As in the MIDOS approach, a subgroup discovery task can be defined as follows: given a population of individuals and a property of those individuals we are interested in, find population subgroups that are statistically ‘most interesting’, e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest. This paper aims at solving a slightly modified subgroup discovery task that can be stated as follows. Again, the input is a population of individuals and a property of those individuals we are interested in, and the output are population subgroups that are statistically ‘most interesting’: are as large as possible, have the most unusual statistical (distributional) characteristics with respect to the property of interest and are sufficiently distinct for detecting most of the target population.

Notice an important aspect of the above two definitions. In both, there is a predefined property of interest, meaning that both aim at characterizing popu-
lation subgroups of a given target class. This property indicates that rule learning may be an appropriate approach for solving the task. However, we argue that standard propositional rule learning \cite{6} and relational rule learning algorithms \cite{19} are unsuitable for subgroup discovery. The main drawback is the use of the covering algorithm for rule set construction. Only the first few rules induced by a covering algorithm may be of interest as subgroup descriptions with sufficient coverage, thus representing a ‘chunk’ of knowledge characterizing a sufficiently large population of covered examples. Subsequent rules are induced from smaller and strongly biased example subsets, i.e., subsets including only positive examples not covered by previously induced rules. This bias prevents a covering algorithm to induce descriptions uncovering significant subgroup properties of the entire population. A remedy to this problem is the use of a weighted covering algorithm, as demonstrated in this paper, where subsequently induced rules with high coverage allow for discovering interesting subgroup properties of the entire population.

This paper investigates how to adapt classification rule learning approaches to subgroup discovery, by exploiting the information about class membership in training examples. This paper shows how this can be achieved by appropriately modifying the covering algorithm (weighted covering algorithm) and the search heuristics (weighted relative accuracy heuristic). The main advantage of the proposed approach is that each rule with high weighted relative accuracy represents a ‘chunk’ of knowledge about the problem, due to the appropriate tradeoff between accuracy and coverage, achieved through the use of the weighted relative accuracy heuristic.

The paper is organized as follows. In Section 2 the background for this work is explained: propositionalization through first-order feature construction, irrelevant feature elimination, the standard covering algorithm used in rule induction, the standard heuristics as well as the weighted relative accuracy heuristic, probabilistic classification and rule evaluation in the ROC space. Section 3 presents the proposed relational subgroup discovery algorithm. Section 4 presents the experimental evaluation in two standard ILP problems (East-West trains and KRK) and a real-life telecommunications application. Section 5 concludes by summarizing the results and presenting plans for further work.

2 Background

This section presents the backgrounds: propositionalization through first-order feature construction, irrelevant feature elimination, the standard covering algorithm used in rule induction, the standard heuristics as well as the weighted relative accuracy heuristic, probabilistic classification and rule evaluation in the ROC space.

2.1 Propositionalization through First-Order Feature Construction

The background knowledge used to construct hypotheses is a distinctive feature of relational rule learning (and inductive logic programming, in general). It is