Impact of Purity Measures on Knowledge Extraction in Decision Trees

Mitja Lenič, Petra Povalej, Peter Kokol

Laboratory for system design, Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

Symbolic knowledge representation is crucial for successful knowledge extraction and consequently for successful data mining. Therefore decision trees and association rules are most commonly used symbolic knowledge representations. Often some sorts of purity measures are used to identify relevant knowledge in data. Selection of appropriate purity measure can have important impact on quality of extracted knowledge. In this paper a novel approach for combining purity measures and thereby altering background knowledge of extraction method is presented. An extensive case study on 42 UCI databases using heuristic decision tree induction as knowledge extraction method is also presented.

Introduction

An important step in the successful data mining process is to select appropriate attributes, which may have significant impact on observed phenomena, before collecting the data. This step introduces an important part of knowledge about the problem in database and therefore represents background knowledge of the database. However, the knowledge extraction method used in the data mining process also includes some predefined background knowledge about the method algorithm, which in the entry uses the background knowledge of the database (its attributes, their types and definition domain) and the data collected.

Background knowledge of knowledge extraction method is usually hard coded into the induction algorithm and depends on its target knowledge representation.

Generally knowledge extraction methods can be divided into search space methods that use some sort of non-deterministic algorithm and heuristic methods that use heuristic function to accomplish goal. Heuristic function is therefore an important part of background knowledge of knowledge extraction method. For decision tree and rule induction methods this heuristics are named (im)purity measures. Since induction algorithm for decision trees and rules is commonly fixed, the only tunable parameters that can adjust background knowledge are the selection of purity measure and discretization method. In this paper the main focus is on impact of purity measures on induction method although also different discretization methods are applied. We introduce new hybrid purity measures that change background knowledge of induction method. Hybrid purity measures are composed out of most commonly used single purity measures. In order to demonstrate the effectiveness of newly introduced hybrid purity measures, a comparison to its single components is performed on 42 UCI databases. We also study the impact of boosting to hybrid and commonly used single purity measures as an alternative (second opinion) knowledge. Additionally the effect of pruning is considered. The paper is focused on the use of purity measures on induction of decision trees, however our findings are not limited only to decision trees and can be applied also in other knowledge extraction methods like rule extraction method, discretization method, etc.

Heuristic induction of decision tree

Decision tree is a hierarchical knowledge representation consisting of tests in inner nodes and their consequences in leafs. Tests define specific attributes and values, representing the inherent knowledge hidden in a database. Since tests and thereafter the extracted knowledge depends mainly on a selected purity measure, the relation purity measure – knowledge is very important and we study and analyze it very profoundly in the present paper.