Title	Using Prior Knowledge in Rule Induction
Author(s)	Nguyen, Dung Duc
Citation	
Issue Date	2003-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/439
Rights	
Description	Supervisor:Ho Tu Bao, 知識科学研究科, 修士

Using Prior Knowledge in Rule Induction

Nguyen Dung Duc

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2003

Keywords: data mining, rule induction, prior knowledge, classification, rule-based classifier.

Learning a set of rules from data is a problem that has attracted considerable interest because a rule provides a concise statement of potential useful information that is easily understood by end users. Rule induction algorithms aim at finding a set of rules that satisfy some predefined criteria. Sequential covering algorithms like AQ, CN2, search for a set of rules that covers all training samples. Simultaneous covering algorithms like ID3, C4.5, extract rules from a decision tree built by dividing samples until some conditions are satisfied. In all of these covering algorithms, learning process will stop when all training samples are covered, or they satisfy some predefined conditions. There are also attempts to integrating domain knowledge into induction process to obtain the benefits from both purely inductive and analytical learning, the learning methods that use prior knowledge to derive general hypothesis deductively. One of the approaches is to use prior knowledge, in terms of domain theory, to augment search operators. The FOCL algorithm uses the domain theory to increase the number of candidate specializations considered at each step of the search for a single Horn clause. Candidate hypotheses are then evaluated based on their performance over the training data. In this way, FOCL combines the greedy, general-to-specific inductive search strategy with the rule-chaining, analytical reasoning of analytical methods.

In this work we propose a new approach to use prior knowledge in rule induction. The concept "prior knowledge" is extended to a broader sense; it includes domain knowledge, prior rule set that has existed as the result of previous learning processes, and user's constraints. The objective of learning is also changed from finding a new hypothesis in the hypotheses space, to finding a better hypothesis based on the existing one, and data.

The approach consists of two main steps. In the first step prior knowledge, or prior rule set, is used to generate a set of simple rules called *rule seeds*. The role of these *rule seeds* is to direct the search for new rules that are different from those in the prior rule set. The second step is to specialize these simple, low confidence *rule seeds* to achieve more accurate rules. The newly learned rules are then used to improve the prior rule set to get a better set of rules. Because no example is removed from training data, the fragmentation is avoided; the newly learned rules are globally significant.

To illustrate the effectiveness, we have conducted two experiments. In the first experiment we apply the proposed approach to improve the classification accuracy of rule-based classifiers produced by LUPC, a sequential covering rule induction algorithm. Experiment result on 26 datasets downloaded from the UCI machine learning repository shows that the proposed approach is effective: it can improve LUPC's classifiers on 7 out of 26 domains. In the second experiment, we apply the proposed approach to improve rule-based classifiers produced by the See5 system. Experiment on 9 datasets shows that the proposed approach effectively improves 3 out of 9 classifiers: classification errors are reduced significantly while the size of classifiers is not changed.