Multistrategy Learning: A Case Study

Pedro Domingos*

Department of Information and Computer Science
University of California, Irvine
Irvine, California 92717, U.S.A.
pedrod@ics.uci.edu
http://www.ics.uci.edu/~pedrod

Two of the most popular approaches to induction are instance-based learning (IBL) and rule generation. Their strengths and weaknesses are largely complementary. IBL methods are able to identify small details in the instance space, but have trouble with attributes that are relevant in some parts of the space but not others. Conversely, rule induction methods may overlook small exception regions, but are able to select different attributes in different parts of the instance space. The two methods have been unified in the RISE algorithm (Domingos 1995). RISE views instances as maximally specific rules, forms more general rules by gradually clustering instances of the same class, and classifies a test example by letting the nearest rule win. This approach potentially combines the advantages of rule induction and IBL, and has indeed been observed to be more accurate than each on a large number of benchmark datasets. However, it is important to determine if this performance is indeed due to the hypothesized advantages, and to define the situations in which RISE's bias will and will not be preferable to those of the individual approaches. This abstract reports experiments to this end in artificial domains.

Compared to rule induction algorithms, RISE should have an advantage when the concept to learn is best described by fairly specific rules, and vice-versa. If concepts are defined by Boolean DNF formulas, their degree of specificity can be measured by L, the average length of the disjuncts. Twenty artificial datasets were randomly generated for each value of L from 1 to 32. C4.5RULES was chosen for comparison. The average accuracies obtained are shown in Figure 1. They indicate that RISE's bias is indeed more appropriate when concepts are fairly to very specific, with the advantage increasing with specificity.[1] More general concepts were learned to a similar degree by both systems. We were unable to determine conditions where C4.5RULES's bias would be preferable to RISE's; corrupting the data with 10% and 20% class noise resulted in similar degradation for the two systems. Another surprising observation is that C4.5RULES's accuracy has an upward trend for length ≥ 14. This is due to the fact that, as the concepts become very specific, it becomes easier to induce short rules for their negation, and C4.5RULES finds these.

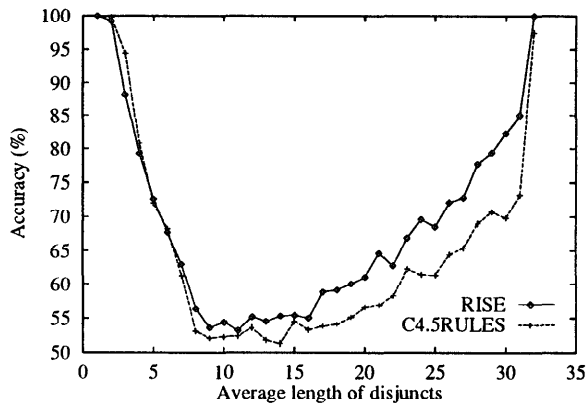

Figure 1: Accuracy as a function of concept specificity.

Compared to IBL algorithms, RISE should have an advantage when some features are relevant in some parts of the instance space, but not in others. A natural measure of the extent to which this happens is D, the average number of different features for all pairs of disjuncts in the target concept description. Twenty mixed Boolean and numeric domains were generated at random for a succession of values of D, and the accuracy of RISE's feature selection method was compared with that of two methods commonly used in IBL, forward and backward selection. RISE's approach performed best throughout, with the advantage increasing with D. However, in separate studies it was observed to be at a disadvantage when each feature is either globally relevant or globally irrelevant, and the dataset is small and noisy.

Lesion studies were also conducted, showing that each of RISE's components is essential to its performance.

References

Domingos, P. 1995. Rule Induction and Instance-Based Learning: A Unified Approach. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1226–1232. Montréal: Morgan Kaufmann.

*Partly supported by a PRAXIS XXI scholarship.

[1] All accuracy differences for length ≥ 12 are significant at the 5% level.