Making the Most of What You've Got: using Models and Data to Improve Learning Rate and Prediction Accuracy

Julio Ortega
Computer Science Dept., Vanderbilt University
P.O. Box 1679, Station B
Nashville, TN 37235
julio@vuse.vanderbilt.edu

Abstract

Prediction and classification in areas such as engineering, medicine, and applied expert systems often relies on two sources of knowledge: actual data and a model of the domain. Recent efforts in machine learning (Ourston 1991) (Towell, Shavlik, & Noordewier 1990) have developed techniques that take advantage of both sources, but the methods are often tied to particular types of models and induction techniques. We propose two general techniques that allow induction methods, C4.5(Quinlan 1993) in our case, to take advantage of an available model(Ortega 1994).

Our first technique exploits the implicit information in the model, which is used as a feature generator for induction. In particular, "model" simulation on input data computes many intermediate and output values which can serve to extend the features that describe the data. For example, in models expressed as propositional theories, we generate extended features from the proofs of the intermediate concepts in the theory. As the number of extended features proliferates quickly, we use feature selection techniques borrowed from the statistical recognition literature (Kittler 1985) to reduce the number of features considered during induction. The original data is re-expressed in terms of the selected features, and induction (i.e. C4.5) is run over this set of data.

Our second technique is motivated by the observation that the reliability of both the available model and the available data may vary widely from one domain (or situation) to another. Our approach consists of evaluating the effectiveness of the model as a predictor using the available data. The data set available for training is divided in two categories: data on which the model is accurate, and data on which the model is inaccurate. This divided data set is used to a build a "Model Reliability" predictor (using our default inductive method, i.e. C4.5) that provides a mechanism for deciding in which situations the model should be chosen for the prediction of future instances. Another predictor, the "Data" predictor is built using induction on the available data. This predictor is used on future instances where the "Model Reliability" predictor indicates that the model is unreliable.

We have conducted experiments using some databases (recognizing DNA promoter sequences, soybean and audiology diseases) which are often used in the Machine Learning community as a benchmark. The results show that the combined use of our techniques compare favorably to existing approaches, both in terms of efficiency and accuracy. However, unlike other techniques for combining inductive and deductive learning, the techniques we are developing are quite general and can be adapted to apply in non-propositional domains. We are also implementing our techniques in a domain where the model is of a mathematical nature (prediction of diabetes glucose levels), and a domain where the model is of a qualitative nature (Reaction Control System of the Space Shuttle).

References

Kittler, T. 1985. Feature selection and extraction. In Young, T. Y., and Fu, K. S., eds., *Handbook of Pattern Recognition and Image Processing.* Orlando, FL: Academic Press. 59–83.

Ortega, J. 1994. Making the most of what you've got: using models and data to improve learning rate and prediction accuracy. Technical Report TR-94-01, Computer Science Dept., Vanderbilt University.

Ourston, D. 1991. *Using Explanation-Based and Empirical Methods in Theory Revision.* Ph.D. Dissertation, University of Texas, Austin, TX.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

Towell, G. G.; Shavlik, J. W.; and Noordewier, M. O. 1990. Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, 861–866.

Acknowledgments

This research is supported by NASA Ames grant NAG 2-834 to D.H. Fisher.