Ad Hoc Attribute-Value Prediction

Gabor Melli
Simon Fraser University
British Columbia, Canada, V5A 1S6
melli@cs.sfu.ca

The evolving ease and efficiency in accessing large amounts of data presents an opportunity to execute prediction tasks based on this data (Hunt, Marin, & Stone 1964). Research in learning-from-example has addressed this opportunity with algorithms that induce either decision structures (ID3) or classification rules (AQ15). Lazy learning research on the other hand, delay the model construction to strictly satisfy a prediction task (Aha, Kibler, & Albert 1991). To support a prediction query against a data set, current techniques require a large amount of preprocessing to either construct a complete domain model, or to determine attribute relevance. Our work in this area is to develop an algorithm that will automatically return a probabilistic classification rule for a prediction query with equal accuracy to current techniques but with no preprocessing requirements. The proposed algorithm, DBPredictor, combines the delayed model construction approach of lazy learning along with the information theoretic measure and top-down heuristic search of learning-from-example algorithms. The algorithm induces only the information required to satisfy the prediction query and avoids the attribute relevance tests required by the nearest-neighbour measures of lazy learning.

Given a data set in some domain, an attribute-value prediction query requests the prediction of an attribute's value for some partially described event drawn from this domain. Applicable classification rules for an attribute-value prediction query are shown to be based on the exponential number of combinations of the attribute-values specified in the query. DBPredictor performs an informed top-down search that incrementally specializes one attribute-value at a time to locate a maximally valued classification rule. In a sense the algorithm iteratively selects the next most relevant attribute-value for this query. If interrupted, the algorithm reports the best encountered classification rule to date. Given a query with n instantiated values the search is contained to $n + (n-1) + ... + 1 = n(n+1)/2$ evaluations. We use the information-theoretic J-Measure (Smyth & Goodman 1992) to evaluate the quality of a probabilistic classification rule.

A corresponding program based on DBPredictor has been developed to satisfy ad hoc attribute-value prediction queries against SQL-based relational database management systems. The performance and accuracy of DBPredictor is empirically tested against both real-world and synthetic data. Furthermore the results are contrasted to the ID3, AQ15, and ITRULE algorithms. DBPredictor commonly requires two orders of magnitude less processing than the other algorithms for a single query. Finally, as expected, DBPredictor's accuracy is equivalent to that of the other algorithms.

To exemplify the increased access to large amounts of information and the opportunities provided by this technique, an interactive version of the program has be placed on the World Wide Web. Users first choose the real-world database they want to query against, next they choose the attribute whose value will be predicted and finally, with the use of pull down menus, describe the event's instantiated attributes. Because of the inconclusive nature of most prediction queries based on real-world databases, the generated report provides a ranked distribution of the values to expect, rather than only the most likely value.

Several interesting future research directions are possible for DBPredictor. First, the search technique could be extended to support requests for more accurate classification rules. Finally the caching of discovered rules could be used to expedite future queries and eventually to better understand the underlying model of a large data set.

The DBPredictor algorithm's flexibility and efficiency can support ad hoc attribute-value prediction queries against large and accessible data sets of the near future.

References

Aha, D. W.; Kibler, D.; and Albert, M. K. 1991. Instance-based learning algorithms. *Machine Learning* 6:37–66.

Hunt, E. B.; Marin, J.; and Stone, P. J. 1964. *Experiments in Induction.* New York: Academic Press.

Smyth, P., and Goodman, R. M. 1992. An information theorctic approach to rule induction. *IEEE Transactions on Knowledge and Data Engineering* 4(4):301–316.