Evaluating the Role of Background Knowledge in Enhancing Knowledge Discovery in Databases

Venkateswarlu Kolluri
Department of Information Science
University of Pittsburgh
Pittsburgh, PA 15260
venkat@sis.pitt.edu

In the field of Knowledge Discovery in Databases (KDD), background knowledge is usually available in the form of taxonomies (*is-a* hierarchies) over the features and the corresponding feature-value hierarchies. Using the RL inductive rule learning system (Clearwater & Provost 1991) as a test bed, we are trying to evaluate the effectiveness of such background knowledge in enhancing the KDD process. The improvements in the learned concept definitions are evaluated based on the learnt rule set's predictive power and simplicity. [1]

In order to estimate the improvement in the learning process, with the addition of background knowledge, rule sets generated (using both artificial and real world data sets) before and after the inclusion of feature/value hierarchies are compared. In the presence of multiple hierarchies, we are trying to look at various ways in which we can introduce domain constraints that will prevent the formulation of overlapping rules (i.e. rules that use the same feature redundantly, across multiple hierarchies)

In most domains, it is common to find very large taxonomies over features. Borrowing ideas from conceptual clustering literature (Fisher, 1987) we are trying to use information theoretic measures to identify the appropriate level of abstraction, for a given learning task.

Each node in the hierarchy, for a given feature, can be considered as a spliting function that splits the given set of training instances into respective subsets (depending on the number of children for this node in the hierarchy). Given a set of subsets, the *categorical utility* value (Fisher, 1987) measures the inter-class dissimilarity and intra-class similarity. For an attribute value pair $A_i = V_{ij}$, and class, C_k, intra-class similarity is measured in terms of a conditional probability $P(A_i = V_{ij}|C_k)$; Similarly, the inter-class similarity is measured in terms of $P(C_k|A_i = V_{ij})$; Using this approach, the quality of the partition can be measured by (Fisher 1987): $\sum_{k=1}^{n} \sum_i \sum_j P(A_i = V_{ij})P(C_k|A_i = V_{ij})P(A_i = V_{ij}|C_k)$, which gives the tradeoff of between *predictability* and *predictiveness*, that is summed for all classes (k), attributes (i), and values (j), over (n) instances. For a given hierarchy, the level of abstraction that maximizes this measure, can be considered to be the appropriate level for the learning task in hand. By using only those generalized feature values the induction system might come up with simplified and intuitive concept definitions.

Although most feature/value hierarchies are fixed in advance, usually by the domain experts, some learning tasks require few minor adjustments to suit the learning task (Han and Fu 1994). Given a feature/value hierarchy, a learning task, and a relevent data set, nodes in the hierarchy can be either *split* into further child nodes, or *merged* into superior nodes, to best suit the learning task. The *categorical utility* value measure can be used to decide whether to *split*, *merge* or leave the hierarchy as it is. To make the process computationally inexpensive, only the nodes at the level, found to be appropriate for the learning task can be considered for refinement. Such selective refinement of existing background knowledge gives us an opportunity to explore the advantages of using *constructive induction* in KDD.

Acknowledgements

The author is grateful to Dr. Bruce Buchanan and Dr. John Aronis for their help and guidance. This work is supported by National Science Foundation grant IRI-9412549.

References

[1] Clearwater, S. and Provost, F.J. RL4: A Tool for Knowledge-Based Induction. In Proc. of the Second Intl. IEEE Conf. on Tools for Artificial Intelligence, pp:24-30, IEEE C.S. Press.

[2] Fisher, D. Improving inference through conceptual clustering. In Proc. of the Natl. Conf. on Artificial Intelligence. (AAAI'97), Seattle, Washington, July 1987.

[3] Han, J. and Fu, Y. Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. In Proc. of the Workshop on Knowledge Discovery in Databases (KDD'94), WA, July 1994, pp:157-168.