Induction of Selective Bayesian Networks from Data

Moninder Singh*

Department of Computer and Information Science
University of Pennsylvania, Philadelphia, PA 19104-6389
msingh@gradient.cis.upenn.edu

Bayesian networks (Pearl 1988), which provide a compact graphical way to express complex probabilistic relationships among several random variables, are rapidly becoming the tool of choice for dealing with uncertainty in knowledge based systems. Amongst the many advantages offered by Bayesian networks over other representations such as decision trees and neural networks are the ease of comprehensibility to humans, effectiveness as complex decision making models and elicitability of informative prior distributions.

However, approaches based on Bayesian networks have often been dismissed as unfit for many real-world applications because they are difficult to construct and probabilistic inference is intractable for most problems of realistic size. Given the increasing availability of large amounts of data in most domains, learning of Bayesian networks from data can circumvent the first problem. This research deals primarily with the second problem. We address this issue by learning **selective** Bayesian networks — a variant of the Bayesian network that uses only a subset of the given attributes to model a domain. Our aim is to learn networks that are smaller, and hence computationally simpler to evaluate, but display accuracy comparable to that of networks induced using all attributes.

We have developed two methods for inducing selective Bayesian networks from data. The first method, K2-AS (Singh & Provan 1995), selects a subset of attributes that maximizes predictive accuracy prior to the network learning phase.The idea behind this approach is that attributes which have little or no influence on the accuracy of learned networks can be discarded without significantly affecting their performance. The second method we have developed, Info-AS (Singh & Provan 1996), uses information-theoretic metrics to efficiently select a subset of attributes from which to learn the classifier. The aim is to discard those attributes which can give us little or no information about the class variable, given the other attributes in the network. We have showed that relative to networks learned using all attributes, networks learned by both K2-AS and Info-AS are significantly smaller and computationally simpler to evaluate but display comparable predictive accuracy. Moreover, they display faster learning rates, hence requiring smaller datasets to achieve their asymptotic accuracy. We have also shown that both methods significantly outperform the *naive* Bayesian classifier, one of the most widely-studied Bayesian methods within the machine learning community.

These results have several important ramifications. First, they give us a way of applying Bayesian networks to problems where it was not possible to do so previously, due to computational intractability. Second, they show that decreasing the size of the networks does not significantly reduce the classification accuracy which may be very important in some applications (e.g. medicine). Third, in real world applications, features may have an associated cost (e.g. a feature representing an expensive test). The learning algorithms proposed can be modified to prefer removal of such high-cost tests.

Since databases from most real life domains, especially medicine, are replete with missing data, we are also working on extending our learning algorithms to deal with such data.Previous work in this area basically deals with learning the conditional probability tables assuming that the Bayesian network structure is known. We are trying to extend this work to learn both network structure as well as the probability tables from data that has missing values/variables and incorporate it with the feature selection approaches presented in this paper. Moreover, we would like to test our methods of learning selective Bayesian networks on two real-world databases in the domain of acute abdominal pain. This domain is relatively hard, yielding about 65% accuracy with other learning methods. Thus, it offers a good test bed for our ideas.

References

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann.

Singh, M., and Provan, G. M. 1995. A comparison of induction algorithms for selective and non-selective Bayesian classifiers. In *Proc. 12th Intl. Conference on Machine Learning*, 497–505.

Singh, M., and Provan, G. M. 1996. Efficient learning of selective Bayesian network classifiers. To appear in *Proc. 13th Intl. Conference on Machine Learning.*

*This work is funded by an IBM Cooperative fellowship.