A Representation Reducing Approach to Scientific Discovery

Joseph Phillips
University of Michigan
1101 Beal Ave., Ann Arbor, MI 48109-2110, USA
josephp@eecs.umich.edu

The proliferation of sensors and the ease of large dataset maintenance have given many scientists more data than can be analyzed by traditional means. Computers have long been used to help scientists calculate. Science, however, is more than calculating. Science at least also involves hypothesis generation and testing, planning, and integrating prior knowledge with new ideas. Artificial intelligence (AI) and database (DB) technologies have grown to the point where they may be able to help scientists in more meaningful ways. We investigate a principled approach to semi-automated knowledge discovery in databases (KDD) for integrated scientific discovery and the rationale for this approach.

Nordhausen and Langley developed perhaps the first system for integrated automated scientific discovery (Nordhausen & Langley 1990). IDS (for Integrated Discovery System) takes an initial **is-a** hierarchy and a sequential list of "histories" (qualitative states) and produces a fuller, richer hierarchy of classified history abstractions with associated state and transition laws. The system can learn taxonomic, qualitative and numeric relations and can make predictions.

Unfortunately, IDS' hierarchy imposes a scientifically irrelevant bias. Operations are limited to local tree manipulations. This constrains search by limiting operator applicability. However, it also limits the resulting knowledge to be obtainable by local comparison.

Some sort of bias is needed to control search. We suggest replacing this external bias with one that many believe is already employed in scientific discovery: Occam's Razor. Occam's razor has the advantage of having a history in both machine learning (Briscoe & Caelli 1996) and scientific discovery (Mach 1895). Our approach is to try to minimize the combined size of a stylized Prolog program that can predict the data with some degree of accuracy, and, a correcting string composed of the errors (*i.e* observed value minus predicted). The method is motivated by Minimum Description Length encoding (Rissanen 1983). Prolog programs are chosen from a restricted class that have some of the computational machinery of modern science. Like IDS, these programs can represent and use taxonomies, equations and symbolic knowledge. Correcting strings have a field for each observed value. The length of each field increases monotonically with the magnitude of the error. Unless we want to compress the table without loss, it suffices to estimate a correcting string's length rather than actually identify it.

Just because we change the bias to the "tried-and-true" Occam's razor does not make the problem simpler, however. One problem is that we do not know *a priori* which operators are best. This may be cast as a multi-armed bandit problem where a discovery operator instance's probability of success depends on its type. Operator instance types group instances by the operator, the kind of data that the operator is applied to, the results of prior operator instances, *etc*. Our research investigates a method of intelligently estimating the operator instance's probability of success from estimates generated by its and similar types. Success estimates generated at different levels of type specificity are combined so that more specific information is relied upon more heavily as it becomes more statistically significant. The main responsibility for estimating success probability smoothly transitions from very general to very specific estimators for as many levels of specificity as desired.

I present criteria for KDD specific to scientific discovery that allow me to characterize how well my approach does when compared with IDS and other KDD methods. Seismology is used as a test domain.

References

Briscoe, G., Caelli, T., 1996. *A Compendium of Machine Learning, Volume 1: Symbolic Machine Learning*. Norwood, New Jersey: Ablex.

Mach, E. 1895. The Economical Nature of Physics in *Popular scientific lectures*. tr. by T. J. McCormack, Chicago: The Open Court Publishing Company, 1895.

Nordhausen, B., Langley, P., 1990. An Integrated Approach to Empirical Discovery. In J. Shrager and P. Langley (Eds.) *Computational Models of Scientific Discovery and Theory Formation*, San Mateo, CA: Morgan Kaufmann, 1990.

Rissanen, J., 1983. A Universal Prior For Integers And Estimation By Minimum Description Length, *The Annals of Statistics*, Vol 11, No. 2, p. 416-431.