Mining Predictive Patterns in Sequences of Events

Gary M. Weiss

Department of Computer Science
Rutgers University
New Brunswick, NJ 08903
gweiss@cs.rutgers.edu

Extended Abstract

Learning to predict rare events from sequences of events with categorical features is an important, real-world, problem. Unfortunately, most machine learning methods that learn classification "rules" are not suited to solving this type of problem because they assume an unordered set of examples and cannot identify patterns *between* "examples" (i.e., events). Statistical time-series prediction methods are also not suitable, since they assume numerical features. Genetic algorithms, however, which have often been used to find patterns in data, are well suited to finding predictive temporal and sequential patterns in the event sequence data.

In order to solve the event prediction problem, we developed Timeweaver, a genetic-based machine learning system that, given a pre-specified "target" event, learns to identify patterns in the data that successfully predict the future occurrence of that event. Timeweaver has been applied to the task of predicting telecommunication failures from time-stamped alarm messages and has outperformed several simple prediction methods. The event prediction task and Timeweaver are described in a KDD paper (Weiss & Hirsh 1998), as well as in a GECCO paper that provides a more detailed description of the GA (Weiss 1999). Due to the availability of these papers, only a short description of Timeweaver is provided in these workshop notes.

We employ a Michigan-style GA to evolve a set of prediction rules. The main issue we faced when applying an evolutionary algorithm to the event prediction problem was finding a balance between exploration of the search space and efficiency of the search. In particular, we needed to find an appropriate fitness function and diversity maintenance strategy. Our fitness function factored in both the *recall* of each rule (the percentage of target events predicted) and the *precision* of each rule (the percentage of correct predictions). The way in which these two measures were combined was found to have a dramatic impact on the search. The strategy we adopted involves varying the importance of these two measures, so that a population is evolved that contains some highly precise rules and some less precise rules that cover more target events. A niching strategy called sharing was used in order to ensure that a diverse set of rules were developed, so that collectively the prediction rules predict the majority of the target events.

While the issues just described—of balancing exploration and exploitation and of maintaining diversity—are issues that are always important when using evolutionary methods, we believe they have even greater importance when developing evolutionary methods for data mining. The reason for this belief is that in data mining the goal is not to find just one solution, but many "solutions". We believe that it is still an open question whether evolutionary methods are widely applicable to data mining problems and believe that progress is needed in managing and coordinating evolutionary search before such problems can be effectively handled. Because of the difficult issues just mentioned, it is not surprising that many data mining methods (especially those that find association rules) utilize relatively simple, complete, algorithms.

The event prediction problem we have described is not a particularly good example of a data-mining problem, since one must specify the specific target event to be predicted. The problem would have much more of the "flavor" of a data-mining problem if this requirement were relaxed, so that the problem is to find *all* predictive patterns. With some extensions, Timeweaver should, in principle, be able to solve this more general problem, but for the reasons mentioned earlier, we are unsure whether such a system would be computationally tractable.

It is worth noting that efficient data mining algorithms exist for a problem that appears, on the surface, to be similar to the event prediction problem. Manilla, Toivonen & Verkamo (1995) have developed data mining methods for finding common patterns in sequential data. However, although common patterns can be used to predict some future, the problem of finding common patterns is actually much easier than learning to predict future events.

References

Manilla, H., Toivonen, H. & Verkamo, A. 1995. Discovering Frequent Episodes in Sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 210-215. AAAI Press.

Weiss, G. M. 1999. Timeweaver: a Genetic Algorithm for Identifying Predictive Patterns in Sequences of Events. In *Proceedings of the Genetic and Evolutionary Computation Conference*. San Francisco, CA: Morgan Kaufmann.

Weiss, G. M., and Hirsh, H. 1998. Learning to Predict Rare Events in Event Sequences. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 359-363. Menlo Park, CA: AAAI Press.