Extracting Knowledge from Neural Networks

Christie M. Fuller
Oklahoma State University, USA

Rick L. Wilson
Oklahoma State University, USA

INTRODUCTION

Neural networks (NN) as classifier systems have shown great promise in many problem domains in empirical studies over the past two decades. Using case classification accuracy as the criteria, neural networks have typically outperformed traditional parametric techniques (e.g., discriminant analysis, logistic regression) as well as other non-parametric approaches (e.g., various inductive learning systems such as ID3, C4.5, CART, etc.).

In spite of this strong evidence of superior performance, the use of neural networks in organizations has been hampered by the lack of an “easy” way of explaining what the neural network has learned about the domain being studied. It is well known that knowledge in a neural network is “mysteriously” encapsulated in its connection weights. It is well accepted that decision-makers prefer techniques that can provide good explanations about the knowledge found in a domain even if they are less effective in terms of classification accuracy.

Over the past decade, neural network researchers have begun an active research stream that focuses on developing techniques for extracting usable knowledge from a trained neural network. The literature has become quite vast and, unfortunately, still lacks any form of consensus on the best way to help neural networks be more useful to knowledge discovery practitioners.

This article will then provide a brief review of recent work in one specific area of the neural network/knowledge discovery research stream. This review considers knowledge extraction techniques that create IF-THEN rules from trained feed-forward neural networks used as classifiers.

We chose this narrow view for a couple of important reasons. First, as mentioned, the research in this area is extraordinarily broad and a critical review cannot be done without focusing on a smaller subset within the literature. Second, classification problems are a familiar problem in business. Third, creating basic IF-THEN rules from a trained neural network is viewed as the most useful area in the entire research stream for the knowledge management and data mining practitioner.

With this narrow focus, some aspects of knowledge extraction from neural networks are obviously not mentioned here. With the focus on deterministic IF-THEN rules, outputs that include “fuzziness” (fuzzy logic) are omitted. In addition, research that involves different neural network architectures (e.g., recurrent networks) and/or different knowledge discovery problem areas (e.g., regression/prediction rather than classification) are also excluded from the review.

BACKGROUND

The discussion of the different neural network knowledge extraction techniques are organized around the fundamental premise or process used for rule extraction. Previous researchers (including Tickle, Maire, Bologna, Andrews, & Diederich, 2000) have used the following terms to help segment the different approaches: decompositional, pedagogical, and eclectic.

Decompositional techniques for rule extraction are approaches that perform rule extraction at the individual neuron (or neural component) level. Pedagogical approaches, on the other hand, extract knowledge by treating the entire NN as a “black box,” creating rules by correlating inputs to the neural network to the resultant outputs (without considering anything about the structure or weights of the NN). It is reasonable to think of these two terms as extreme points in a continuous spectrum of approaches. Eclectic approaches are techniques that borrow some aspects from each of the two extremes.

Figure 1 helps visualize how these algorithms work. Figure 1 shows a 6-input, 3 hidden neuron, 2 output neural network. Assuming no bias inputs and a fully connected neural network, there would be 24 connection weights (not shown) which represent the knowledge stored in the neural network (after, of course, the NN has been trained on a set of data). The decompositional approaches will examine (at least) the connection weights that lead to each hidden neuron and will “discover rules” such as IF X2 < 7, THEN CONCLUDE Class A. Pedagogical approaches would present systematic random inputs to the neural
network, observe the output of the neural networks, and “learn” rules like above through studying the relationship between input and output variations.

The review of pertinent neural network rule extraction algorithms also will include three different measures of technique usefulness (accuracy, fidelity, and comprehensibility) when such measures have been studied. These three different measures of technique usefulness are important in assessing the quality of the different methodologies. Accuracy measures the ability of the derived rule set to classify cases from the problem domain. This is typically reported as percentage correctly classified. Fidelity measures how well classification of cases using rules extracted from the trained NN mimic the classification of just the original NN. Comprehensibility measures the size of the rule set extracted from the neural network, looking at both number of rules and the number of antecedents in the rule. This is a measure of rule complexity.

Additionally, any empirical comparisons to other “competing” rule extraction techniques will also be shared, as well as any empirical comparisons to other comparable well-known knowledge discovery techniques such as inductive learning systems (which might include such well-studied techniques as ID3, C4.5, CART, and See5, among others).

The comparisons to other knowledge discovery techniques that result in decision trees, rules, and so forth are particularly relevant to practice. The techniques discussed in this article use a two-step approach to knowledge elicitation: first, one trains a neural network on a set of cases, and then rules are found through the specific technique. Inductive learning approaches accomplished this in one step directly from the cases. The ongoing research into the two-step neural network process seeks to explore the contention by Tickle et al. (2000) and others who claim that rule extraction from trained neural networks will (someday) be a better approach than inductive learning systems. This claim stems from the potential for realizing additional correct classifications (which is well documented) possible with neural networks. The quest for this still continues, and results (as shown) show promise. The next section outlines recent progress in this area.

**NN RULE EXTRACTION REVIEW**

**Decompositional Approaches**

Su, Hsu, and Tsai (2002) present a decompositional methodology that is similar to many in this category. After a feed-forward NN is trained, important inputs are identified and unnecessary connections are pruned from the network in an iterative fashion. Input data is then transformed into binary form, and then the knowledge (in rules) is extracted by building truth tables for each hidden neuron. The rules are simplified using Karnaugh maps and removing repetitive statements. The classification accuracy of this algorithm showed mixed results when compared to the original neural network model and the inductive learning approach See5. However, their algorithm had increased comprehensibility compared to See5, requiring far fewer attributes to classify the data.

The GLARE approach (GeneraLized Analytic Rule Extraction) is similar, extracting rules from standard feed-forward backpropagation networks requiring binary inputs (Gupta, Park, & Lam, 1999). First, weights between input nodes and hidden nodes are ranked, determining the largest connection weights. Next, another index calculates the importance of the connections between the hidden and output layers using the ranking. This step leads to the extraction of symbolic rules. Using classification accuracy as a benchmark with strictly categorical datasets, GLARE performs as well or better than the inductive learning system C4.5 and the original backpropagation NN. When used with datasets possessing continuous inputs that were converted to Boolean inputs, GLARE’s performance suffered significantly.

Vaughn, Cavill, Taylor, Foy, and Fogg (2000) report on a rule extraction method developed for multi-layer feed-forward neural networks with binary inputs. Their method ranks the activation contributions to each of the hidden neurons, and then repeats the process to determine the significant input neurons. Rules are extracted by combining the significant inputs related to a particular class, in ranked order.