Employing Neural Networks in Data Mining

Mohamed Salah Hamdi
UAE University, UAE

INTRODUCTION

Data-mining technology delivers two key benefits: (i) a descriptive function, enabling enterprises, regardless of industry or size, in the context of defined business objectives, to automatically explore, visualize, and understand their data and to identify patterns, relationships, and dependencies that impact business outcomes (i.e., revenue growth, profit improvement, cost containment, and risk management); (ii) a predictive function, enabling relationships uncovered and identified through the data-mining process to be expressed as business rules or predictive models. These outputs can be communicated in traditional reporting formats (i.e., presentations, briefs, electronic information sharing) to guide business planning and strategy. Also, these outputs, expressed as programming code, can be deployed or hard wired into business-operating systems to generate predictions of future outcomes, based on newly generated data, with higher accuracy and certainty.

However, there also are barriers to effective large-scale data mining. Barriers related to the technical aspects of data mining concern issues such as large datasets, highly complex data, high algorithmic complexity, heavy data management demands, and qualification of results (Musick, Fidelis & Slezak, 1997). Barriers related to attitudes, policies, and resources concern potential dangers such as privacy concerns (Kobsa, 2002).

The number of research projects and publications reporting experiences with data mining has been growing steadily. Researchers in many different fields, including database systems, knowledge-base systems, artificial intelligence, machine learning, knowledge acquisition, statistics, spatial databases, data visualization, and Internet computing, have shown great interest in data mining. In this contribution, the focus is on the relationship between artificial neural networks and data mining. We review some of the related literature and report on our own experience in this context.

BACKGROUND

Artificial Neural Networks

Neural networks are patterned after the biological ganglia and synapses of the nervous system. The essential element of the neural network is the neuron. A typical neuron j receives a set of input signals from the other connected neurons, each of which is multiplied by a synaptic weight of w_{ij} (weight for connection between neurons i and j). The resulting activation weights are then summed to produce the activation level for the neuron j. Learning is carried out by adjusting the weights in a neural network. Neurons that contribute to the correct answer have their weights strengthened, while other neurons have their weights reduced. Several architectures and error correction algorithms have been developed for neural networks (Haykin, 1999).

In general, neural networks can help where an algorithmic solution cannot be formulated, where lots of examples of the required behavior are available, and where picking out the structure from existing data is needed. Neural networks work by feeding in some input variables and producing some output variables. Therefore, they can be used where some known information is available and some unknown information should be inferred.

Data Mining

The data-mining revolution started in the mid-1990s. It was characterized by the incorporation of existing and already well-established tools and algorithms such as machine learning. In 1995, the International Conference on Knowledge Discovery and Data Mining became the most important annual event for data mining. The framework of data mining also was outlined in many books, such as *Advances in Knowledge Discovery and Data Mining* (Fayyad et al., 1996). Data-mining conferences like ACM SIGKDD, SPIE, PKDD, and SIAM, and journals like *Data Mining and Knowledge Discovery Journal* (1997), *Journal of Knowledge and Information Systems* (1999), and *IEEE Transactions on Knowledge and Data Engineering* (1989) have become an integral part of the data-mining field.

The trends in data mining over the last few years include OLAP (Online Analytical Processing), data warehousing, association rules, high performance data-mining systems, visualization techniques, and applications of data mining. Recently, new trends have emerged that have great potential to benefit the data-mining field, like XML (eXtensible Markup Language) and XML-related technologies, database products that incorporate data-mining tools, and new developments in the design and

implementation of the data-mining process. Another important data-mining issue is concerned with the relationship between theoretical data-mining research and data-mining applications. Data mining is an exponentially growing field with a strong emphasis on applications.

A further issue of great importance is the research in data-mining algorithms and the discussion of issues of scale (Hearst, 1997). The commonly used tools may not scale up to huge volumes of data. Scalable data-mining tools are characterized by the linear increase of their runtime with the increase of the number of data points within a fixed amount of available memory. An overview of scalable data-mining tools is given in Ganti, Gehrke, and Ramakrishnan (1999). In addition to scalability, robust techniques to model noisy data sets containing an unknown number of overlapping categories are of great importance (Krishnapuram et al., 2001).

MAIN THRUST

Exploiting Neural Networks in Data Mining

How is data mining able to tell you important things that you didn't know or tell you what is going to happen next? The technique that is used to perform these feats is called *modeling*. Modeling is simply the act of building a model based on data from situations where the answer is known, and then applying the model to other situations where the answers aren't known. Modeling techniques have been around for centuries, but it is only recently that data storage and communication capabilities required to collect and to store huge amounts of data and the computational power to automate modeling techniques to work directly on the data have been available. Modeling techniques used for data mining include decision trees, rule induction, genetic algorithms, nearest neighbor, artificial neural networks, and many other techniques (Chen, Han & Yu, 1996; Cios, Pedrycz & Swiniarski, 1998; Hand, Mannila & Smyth, 2000).

Exploiting artificial neural networks as a modeling technique for data mining is considered to be an important direction of research. Neural networks can be applied to a number of data-mining problems, including classification, regression, and clustering, and there are quite a few interesting developments and tools that are being developed in this field. Lu, Setiono, and Liu (1996) applied neural networks to mine symbolic classification rules from large databases. They report that neural networks were able to deliver a lower error rate and are more robust against noise than decision trees. Ainslie and Drèze (1996) show how effective data mining can be achieved by combining the power of neural networks with the rigor of more traditional statistical tools. They argue that this alliance can generate important synergies. Craven and Shavlik (1997) describe neural network learning algorithms for data mining that are able to produce comprehensible models and that do not require excessive training times. They argue that neural network methods deserve a place in the toolboxes of data-mining specialists. Mitra, Pal, and Mitra (2002) provide a survey of the available literature on data mining using soft computing methodologies, including neural networks. They came to the conclusion that neural networks are suitable in data-rich environments and are typically used for extracting embedded knowledge in the form of rules, quantitative evaluation of these rules, clustering, self-organization, classification, and regression. Vesely (2003) argues that from the methods of data mining based on neural networks, the Kohonen's self-organizing maps are the most promising, because, by using a self-organizing map, one can more easily visualize high-dimensional data. Self-organizing maps also outperform other conventional methods such as the popular Principal Component Analysis (PCA) method for screening analysis of high-dimensional data. Although highly successful in typical cases, PCA suffers from the drawback of being a linear method. Furthermore, real-world data manifolds, besides being nonlinear, often are corrupted by noise and embed into high-dimensional spaces. Self-organizing maps are more robust against noise and are often used to provide representations that can be analyzed successfully using conventional methods like PCA.

In spite of their excellent performance in concept discovery, neural networks do suffer from some shortcomings. They are sensitive to the net topology, initial weights, and the selection of attributes. If the number of layers is not selected suitably, the learning efficiency will be affected. Too many irrelevant nodes can cause unnecessary computational expense and overfit (i.e., the network creates meaningless concepts); randomly selected initial weights sometimes can trap the nets in so-called pitfalls; that is, neural nets stabilize around local minima instead of the global minimum. Background knowledge remains unused in neural nets. The knowledge discovered by nets is not transparent to users. This is perhaps the main failing of neural networks, as they are unintelligible black boxes.

Our own work is focused on mining educational data to assist e-learning in a variety of ways. In the following section, we report on the experience with MASACAD (Multi-Agent System for ACademic ADvising), a data-mining, multi-agent system that advises students using neural networks.