Knowledge Structure and Data Mining Techniques

Rick L. Wilson
Oklahoma State University, USA

Peter A. Rosen
University of Evansville, USA

Mohammad Saad Al-Ahmadi
Oklahoma State University, USA

INTRODUCTION

Considerable research has been done in the recent past that compares the performance of different data mining techniques on various data sets (e.g., Lim, Low, & Shih, 2000). The goal of these studies is to try to determine which data mining technique performs best under what circumstances. Results are often conflicting—for instance, some articles find that neural networks (NN) outperform both traditional statistical techniques and inductive learning techniques, but then the opposite is found with other datasets (Sen & Gibbs, 1994; Sung, Chang, & Lee, 1999: Spangler, May, & Vargas, 1999). Most of these studies use publicly available datasets in their analysis, and because they are not artificially created, it is difficult to control for possible data characteristics in the analysis. Another drawback of these datasets is that they are usually very small.

With conflicting empirical results in the knowledge discovery/data mining literature, there have been numerous calls for a more systematic study of different techniques using synthetic, well-understood data. The rationale for synthetic data is that various factors can be manipulated while others are controlled, which may lead to a better understanding of why technique X outperforms technique Y in some, but not all, circumstances (Scott & Wilkins, 1999).

This call for research dates back to Quinlan's seminal work in inductive learning algorithms. In his 1994 study that analyzed the difference between neural networks and inductive decision trees, Quinlan conjectures the existence of what he called S-problems and P-problems. In his definition, S-problems are those that are unsuited for NN's, while P-problems are those unsuited for decision tree induction. More recently, the review work on neural networks by Tickle, Maine, Bologna, Andrews, and Diederich (2000) propose that determining whether a classification task belongs to the P-problem or S-problem set is a very important research question.

Recently, other researchers have proposed that the composition of the underlying knowledge in a dataset, or knowledge structure (KS), may be pertinent in understanding why knowledge discovery techniques perform well on one dataset and poorly on others. This term has been used by Hand, Mannila, and Smyth (2001), and Padmanabhan and Tuzhilin (2003) to refer to this phenomenon, while Scott and Wilkins (1999) used a similar term, structural regularities, to describe the same concept.

The goal of this article is to explore in more detail how the existence of a database's underlying *knowledge structure* might help explain past inconsistent results in the knowledge discovery literature. Management scholars will recognize the term knowledge structure, as Walsh (1995) refers to it as a "mental template…imposed on an information environment to give it form and meaning." Therefore, for the knowledge discovery context, we propose that knowledge structure is analogous to the form and meaning of the knowledge to be discovered in a database. Though we will not explore the concept too deeply, one also can define knowledge structure through the use of a parameter set P as proposed by Hand et al. (2001). The parameter set would be attribute-value pairs that detail the existence of a specific knowledge structure for a given knowledge concept/database pair.

This knowledge structure concept is an abstract concept, which may make it hard to visualize. Typically, when a knowledge worker is using a technique to extract knowledge from a database, they will not have any idea about the underlying knowledge structure of the concept of interest. But, researchers have hypothesized that knowledge discovery in a database is optimized when the formalism of the tool matches this underlying structure of the knowledge (Hand et al., 2001). Based on this,

we conjecture that if a knowledge worker did know the knowledge structure parameter values prior to exploring the data, he or she could find the optimal tool for the knowledge discovery process.

From a historical perspective, past knowledge discovery and data mining research results could be explained by whether a particular knowledge discovery tool was or was not a good "match" with the underlying knowledge structure. The idea of matching the tool to the structure is somewhat analogous to the concept of task-technology fit, studied in the MIS literature during the mid 1990s (Goodhue, 1995).

Recent research in other related areas has found that contradictory or difficult to explain results could be related to the concept of knowledge structure (Wilson & Rosen, 2003). In this study, the well-known IRIS and BUPA Liver datasets were used to examine the efficacy of knowledge discovery tools in protected (by data perturbation) confidential databases. The IRIS dataset is known to possess linearly separable classes, while the BUPA Liver dataset cases has been historically difficult to correctly classify for all knowledge discovery tools. An outcome of this research was the proposal that knowledge discovery tool effectiveness in a protected (perturbed) database could be impacted by both the database's underlying knowledge structure and the *noise* present in the database. The concept of *noise* is simply the degree to which the different classes can be separated or differentiated by the optimal tool, or, alternatively, a surrogate measure of how difficult cases are to classify (e.g., Li & Wang, 2004).

Through a simple example, the article will attempt to provide some evidence that the underlying knowledge structure present in a database could have significant impacts on the performance of knowledge discovery tools. Building on past postulation, the example also will explore whether the so-called "match" between the knowledge structure and the knowledge discovery tools' own formalism is important to the classification accuracy of the knowledge discovery task.

BACKGROUND

To investigate the possible impact of what has previously been defined as knowledge structure, a hypothetical database/classification task will be formulated. Thus, the investigation of knowledge structure in this article will be limited to a classification domain. The concepts of knowledge structure can be extended to all kinds of knowledge discovery tasks: prediction/regression, clustering, and so forth. We choose classification as our focus because it is a well-studied area and is easily illustrated in this experiment.

To this end, a 50,000 record fictitious bank database, previously used in another work (see Muralidhar, Parsa, & Sarathy, 1999), will serve as the database for the study. The data, in its original form, has five attributes (Home Equity, Stock/Bonds, Liabilities, Savings/Checking, and CD's) with known means, standard deviations, and so forth.

To simulate the existence of an important knowledge concept, a sixth binary categorical (class) variable was systematically added to the database, representing some important knowledge to a data analyst (perhaps differentiating between profitable customers and not-so-profitable customers). How the class variable was systematically created is addressed and is related to the knowledge structure parameters.

We chose a very simplistic definition of knowledge structure types in our continuing example. Two different knowledge structures were employed, decision tree and linear. The *decision tree* (DT) structure means that the researchers created a decision tree using all five variables, and then the data was applied to the tree to determine class membership (either '0' or '1') in the sixth variable, for each individual case. The specific tree used was chosen such that all variables were found in the tree and that there were an equal number of the two distinct classes created (25,000 cases each of 'class 0' and 'class 1'). The tree itself was obviously somewhat arbitrary, but does represent a scenario where the underlying structure of the knowledge concept was in a decision tree form.

The second structure used was a linear format (LINEAR). A strictly linear relationship was created using all five variables, to determine class membership (either '0' or '1') in the sixth variable for each individual case. Again, the resulting values for the sixth variable included 25,000 cases for each 'class 0' and 'class 1.' This again represents the situation where the underlying knowledge concept of interest is in a linear form.

While these two structures may be overly simplistic, their choice allows us to explore the possible impact of the concept of knowledge structure with minimal moderating factors. Ultimately, the notation of Padmanabhan and Tuzhilin (2003) may be a more formal and more accurate approach to describe this phenomenon, and we will return to this later in the article.

For each of the two exemplar knowledge structures (DT and LINEAR), we created another database that involved adding "noise" to the class variable. The purpose of adding noise is to have the synthesized datasets replicate knowledge discovery situations where a perfect discrimination between classes is not possible, as is true in the case of the previously mentioned BUPA Liver dataset. One could argue that real-world databases are more likely than not to have a high degree of noise.