Explanation-Oriented Data Mining

Yiyu Yao
University of Regina, Canada

Yan Zhao
University of Regina, Canada

INTRODUCTION

Data mining concerns theories, methodologies, and, in particular, computer systems for knowledge extraction or mining from large amounts of data (Han & Kamber, 2000). The extensive studies on data mining have led to many theories, methodologies, efficient algorithms and tools for the discovery of different kinds of knowledge from different types of data. In spite of their differences, they share the same goal, namely, to discover new and useful knowledge, in order to gain a better understanding of nature.

The objective of data mining is, in fact, the goal of scientists when carrying out scientific research, independent of their various disciplines. Data mining, by combining research methods and computer technology, should be considered as a research support system. This goal-oriented view enables us to re-examine data mining in the wider context of scientific research. Such a re-examination leads to new insights into data mining and knowledge discovery.

The result, after an immediate comparison between scientific research and data mining, is that an explanation construction and evaluation task is added to the existing data mining framework. In this chapter, we elaborate upon the basic concerns and methods of explanation construction and evaluation. Explanation-oriented association mining is employed as a concrete example to demonstrate the entire framework.

BACKGROUND

Scientific research and data mining have much in common in terms of their goals, tasks, processes and methodologies. As a recently emerged area of multi-disciplinary study, data mining and knowledge discovery research can benefit from the long established studies of scientific research and investigation (Martella, Nelson, & Marchand-Martella, 1999). By viewing data mining in a wider context of scientific research, we can obtain insights into the necessities and benefits of explanation construction. The model of explanation-oriented data mining is a recent result from such an investigation (Yao, 2003; Yao, Zhao, & Maguire, 2003).

Common Goals of Scientific Research and Data Mining

Scientific research is affected by the perceptions and the purposes of science. Martella et al. summarized the main purposes of science, namely, to describe and predict, to improve or manipulate the world around us, and to explain our world (Martella, et al., 1999). The results of the scientific research process provide a description of an event or a phenomenon. The knowledge obtained from this research helps us to make predictions about what will happen in the future. Research findings are a useful tool for making an improvement in the subject matter. Research findings also can be used to determine the best or the most effective ways of bringing about desirable changes. Finally, scientists develop models and theories to explain why a phenomenon occurs.

Goals similar to those of scientific research have been discussed by many researchers in data mining. For example, Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy identified two high-level goals of data mining as prediction and description (Fayyad, et al., 1996). Prediction involves the use of some variables to predict the values of some other variables, while description focuses on patterns that describe the data. Ling, Chen, Yang and Cheng studied the issue of manipulation and action based on discovered knowledge (Ling *et al.,* 2002). Yao, Zhao, et al. introduced the notion of explanation-oriented data mining, which focuses on constructing models for the explanation of data mining results (2003).

Common Processes of Scientific Research and Data Mining

Research is a highly complex and subtle human activity, which is difficult to formally define. It seems impossible to give any formal instruction on how to do research. On the other hand, some lessons and general principles can be learnt from the experience of scien-

Table 1. The model of scientific research processes

▪ Idea-generation phase: to identify a topic of interest. ▪ Problem-definition phase: to precisely and clearly define and formulate vague and general ideas generated in the previous phase. ▪ Procedure-design/planning phase: to make a workable research plan by considering all issues involved. ▪ Observation/experimentation phase: to observe real world phenomenon, collect data, and carry out experiments. ▪ Data-analysis phase: to make sense out of the data collected. ▪ Results-interpretation phase: to build rational models and theories that explain the results from the data-analysis phase. ▪ Communication phase: to present the research results to the research community.

Table 2. The model of data mining processes

▪ Data pre-processing phase: to select and clean working data. ▪ Data transformation phase: to change the working data into the required form. ▪ Pattern discovery and evaluation phase: to apply algorithms to identify knowledge embedded in data, and to evaluate the discovered knowledge. ▪ Explanation construction and evaluation phase: to construct plausible explanations for discovered knowledge, and to evaluate different explanations. ▪ Pattern presentation: to present the extracted knowledge and explanations.

tists. There are some basic principles and techniques that are commonly used in most types of scientific investigations. We adopt the model of the research process from Garziano and Raulin (2000), and combine it with other models (Martella, et al., 1999). The basic phases and their objectives are summarized in Table 1. It is possible to combine several phases into one, or to divide one phase into more detailed steps. The division between phases is not clear-cut. The research process does not follow a rigid sequencing of the phases. Iteration of different phrases may be necessary (Graziano & Raulin, 2000).

Many researchers have proposed and studied models of data mining processes (Fayyad, et al. 1996; Mannila, 1997; Yao, Zhao, et al., 2003; Zhong, Liu, & Ohsuga, 2001). A model that adds the explanation facility to the commonly used models has been recently proposed by Yao, Zhao, et al.; it is remarkably similar to the model of scientific research. The basic phases and their objectives are summarized in Table 2. Like the research process, the data mining process is also an iterative process and there is no clear-cut difference among the different phases. In fact, Zhong, et al. argue that it should be a dynamically organized process (Zhong, et al., 2001). The entire framework is illustrated in Figure 1.

There is a parallel correspondence between the processes of scientific research and data mining. Their main difference lies in the subjects that perform the tasks. Research is carried out by scientists, and data mining is done by computer systems. In particular, data mining may be viewed as a study of domain-independent research methods with emphasis on data analysis. The higher and more abstract level of comparisons and connections between scientific research and data mining can be further studied in levels that are more concrete. There are bi-directional benefits. The experiences and results from the studies of research methods can be applied to data mining problems; the data mining algorithms can be used to support scientific research.

MAIN THRUST

Explanations of data mining address several important questions. What needs to be explained? How to explain the discovered knowledge? Moreover, is an explanation correct and complete? By answering these questions, one can better understand explanation-oriented data mining. The ideas and processes of explanation construction and explanation evaluation are demonstrated by explanation-oriented association mining.

Figure 1. A framework of explanation-oriented data mining