On Explanation-Oriented Data Mining

Yiyu Yao
University of Regina, Canada

Yan Zhao
University of Regina, Canada

INTRODUCTION

The objective of data mining is to discover new and useful knowledge, in order to gain a better understanding of nature. This in fact is the goal of scientists when carrying out scientific research, independent in their various disciplines. This goal-oriented view enables us to re-examine data mining in a wider context of scientific research. The consequence after the immediate comparison between scientific research and data mining is that, an explanation discovery and evaluation task is added to the existing data mining framework. In this chapter, we elaborate the basic concerns and methods of explanation discovery and evaluation. Explanation-oriented association mining is employed as a concrete example to show the whole framework.

BACKGROUND

Scientific research and data mining have much in common in terms of their goals, tasks, processes and methodologies. As a recently emerged multi-disciplinary study, data mining and knowledge discovery can benefit from the long established studies of scientific research and investigation (Martella *et al*., 1999). By viewing data mining in a wider context of scientific research, we can obtain insights into the necessities and benefits of explanation discovery. The model of explanation-oriented data mining is a recent result from such an investigation (Yao *et al*., 2003). The basic idea of explanation-oriented data mining has drawn attentions from many researchers (Lin & Chalupsky, 2004; Yao, 2003) ever since the introduction of it.

Common Goals of Scientific Research and Data Mining

Scientific research is affected by the perceptions and the purposes of science. Martella *et al*. (1999) summarized the main purposes of science, namely, to describe and predict, to improve or manipulate the world around us, and to explain our world. The results of the scientific research process provide a description of an event or a phenomenon. The knowledge obtained from research helps us to make predictions about what will happen in the future. Research findings are useful for us to make an improvement in the subject matter. Research findings can be used to determine the best or the most effective interventions that will bring 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 *et al*. (1996) identified two high-level goals of data mining as prediction and description. Prediction involves the use of some variables to predict the values of some other variables, and description focuses on patterns that describe the data. Ling *et al*. (2002) studied the issue of manipulation and action based on the discovered knowledge. Yao *et al*. (2003) introduced the notion of explanation-oriented data mining, which focuses on constructing models for the explanation of data mining results.

Common Processes of Scientific Research and Data Mining

The process of scientific research includes idea generation, problem definition, procedure design and planning, observation and experimentation, data analysis, result interpretation, and communication. It is possible to combine several phases into one, or to divide one phase into more detailed steps. The division between phases is not a clear cut. The research process does not follow a rigid sequencing of the phases. Iteration of different phrases may be necessary (Graziano & Raulin, 2000, Martella *et al*., 1999).

Many researchers have proposed and studied models of data mining processes (Fayyad *et al*.,

1996; Mannila, 1997; Yao *et al*., 2003; Zhong *et al*., 2001). The model that adds the explanation facility to the commonly used models is recently proposed by Yao *et al*. The process thus is composed by: date preprocessing, data transformation, pattern discovery and evaluation, explanation discovery and evaluation, and pattern representation. Like the research process, the process of data mining also is an iterative process and there is no clear cut between different phases. In fact, Zhong, *et al*. (2001) argued that it should be a dynamically organized process. The whole framework is illustrated in Figure 1.

There is a parallel correspondence between the processes of scientific research and data mining. The 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 of, and connections between, scientific research and data mining may 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 OF THE CHAPTER

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. In this section, the ideas and processes of explanation profile construction, explanation discovery and explanation evaluation are demonstrated by explanation-oriented association mining.

Basic Issues

- Explanation-oriented data mining explains and interprets the knowledge discovered from data. Knowledge can be discovered by unsupervised learning methods. Unsupervised learning studies how systems can learn to represent, summarize, and organize the data in a way that reflects the internal structure (namely, a pattern) of the

Figure 1. A framework of explanation-oriented data mining