Chapter 6
DATA MINING PROCESSES AND
COLLABORATION PRINCIPLES

Alfpio Jorge, Steve Moyle, Hendrik Blockeel, and Angi Voß

Abstract: Data mining is a process involving the application of human skill as well as technology, and as such it can be supported by clearly defined processes and procedures. This chapter presents the CRISP-DM process, one well developed standard data mining process, which contains clearly defined phases with clearly defined steps and deliverables. The nature of some of the CRISP-DM phases is such that it is possible to perform them in an e-collaboration setting. The principles for extending the CRISP-DM process to support collaborative data mining are described in the RAMSYS approach to data mining. The tools, systems, and evaluation procedures that are required for the RAMSYS approach to reach its potential are described.

1. INTRODUCTION

The core of data mining is the extraction of useful information (models) from data (Hand, et al., 2001). However, data mining is not simply about technology. It requires the application of significant human skill. Data mining (see Chapter 1) brings together many different disciplines with different focuses including: engineering (database systems and implementations), computer science (efficient and scalable implementations), analysis techniques (statistics and machine learning), human-computer interaction (visualization techniques). Facets of all these disciplines (and more) must be combined – typically by employing a team of individuals – together for successful data mining project outcomes. Some of the most important areas not to be neglected are the human factors including project management and control.

Getting from business opportunities to actionable results is a long and non-trivial process (Berry and Linoff, 1997) involving aspects of business and technology (Pyle, 1999). A well defined process is of importance to achieving successful data mining results in a data mining project, particularly as the number of participants involved in carrying out the data mining tasks grows (Pyle, 1999). Many authors
have suggested broadly defined process models to perform data mining (Adriaans and Zantinge, 1996, Fayyad, et al., 1996). The emerging standard data mining process model is the CRoss Industry Standard Process for Data Mining (CRISP-DM) (Chapman, et al., 2000). CRISP-DM subdivides a data mining project into the six, interrelated phases of: (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation, and (6) deployment. Like the alternative data mining processes there are numerous feedback loops connecting the phases in CRISP-DM.

As data mining is multidisciplinary it requires the expertise from numerous individuals. The business understanding phase requires communication skills to work closely with the data mining client (the organization interested in the data mining results), while the modeling phase requires the use of statistics or machine learning. Acquiring the appropriate blend of people with sufficient skills can be a daunting process – particularly if they are to be assembled in the same location. Fortunately, the nature of the CRISP-DM process allows some phases to be undertaken largely independently of others. This makes it possible to perform parts of a data mining process in an e-collaborative setting (see Chapter 5). To ensure that e-collaborative data mining is successful, well defined collaboration principles and support tools are required.

Naturally, the undertaking of a collaborative data mining project increases the complexity of the process (this is clearly described in Chapter 5), but there are potential benefits by the combination of expertise. To achieve such benefits it is vital that all collaborating data miners can share their results (either complete or intermediate), and that other data miners utilize those results to best effect in their own work. For this sharing it is also required that the data miners ‘speak the same language’. For example, in the data preparation phase, any data transformations should be made available. In the modeling phase, the models should also be available in a standardized format. All such data mining project information needs to be securely but easily shared by using an appropriate e-collaboration system.

The evaluation phase is important in the data mining process for it is in this phase that the key results – the models – are compared with respect to the initial project objectives. When working in a collaborative setting it is important that all produced models are evaluated fairly and consistently. One way of achieving these goals is to centralize the model evaluation as much as practicable.

The remainder of this chapter is organized as follows. Section 2 describes in some detail the CRISP-DM process. This is then followed by Section 3 on principles for extending CRISP-DM for collaborative data mining. Section 4 presents tools and services to support such a process. These are (1) an e-collaboration platform – ZENO-for-RAMSYS; and (2) processes for evaluating models in a fair and centralized manner. Other important tools for collaborative data mining such as (3) the SumatraTT data preprocessing tool; (4) tools for sharing and understanding data mining models in a standard format – VizWiz are described in Chapters 18 and 10, respectively. Some case studies are briefly introduced in Section 5 before the final conclusions are presented.