Data Mining Tool Selection

Christophe Giraud-Carrier
Brigham Young University, USA

INTRODUCTION

It is sometimes argued that all one needs to engage in Data Mining (DM) is data and a willingness to "give it a try." Although this view is attractive from the perspective of enthusiastic DM consultants who wish to expand the use of the technology, it can only serve the purposes of one-shot proofs of concept or preliminary studies. It is not representative of the complex reality of deploying DM within existing business processes. In such contexts, one needs two additional ingredients: a process model or methodology, and supporting tools.

Several Data Mining process models have been developed (Fayyad et al, 1996; Brachman & Anand, 1996; Mannila, 1997; Chapman et al, 2000), and although each sheds a slightly different light on the process, their basic tenets and overall structure are essentially the same (Gaul & Saeuberlich, 1999). A recent survey suggests that virtually all practitioners follow some kind of process model when applying DM and that the most widely used methodology is CRISP-DM (KDnuggets Poll, 2002). Here, we focus on the second ingredient, namely, supporting tools.

The past few years have seen a proliferation of DM software packages. Whilst this makes DM technology more readily available to non-expert end-users, it also creates a critical decision point in the overall business decision-making process. When considering the application of Data Mining, business users now face the challenge of selecting, from the available plethora of DM software packages, a tool adequate to their needs and expectations. In order to be informed, such a selection requires a standard basis from which to compare and contrast alternatives along relevant, business-focused dimensions, as well as the location of candidate tools within the space outlined by these dimensions. To meet this business requirement, a standard schema for the characterization of Data Mining software tools needs to be designed.

BACKGROUND

The following is a brief overview, in chronological order, of some of the most relevant work on DM tool characterization and evaluation.

Information Discovery, Inc. published, in 1997, a taxonomy of data mining techniques with a short list of products for each category (Parsaye, 1997). The focus was restricted to implemented DM algorithms.

Elder Research, Inc. produced, in 1998, two lists of commercial desktop DM products (one containing 17 products and the other only 14), defined along a few, yet very detailed, dimensions (Elder & Abbott, 1998; King & Elder, 1998). Another 1998 study contains an overview of 16 products, evaluated against pre-processing, data mining and post-processing features, as well as additional features such as price, platform, release date, etc. (Gaul & Saeuberlich, 1999). The originality of this study is its very interesting application of multidimensional scaling and cluster analysis to position 12 of the 16 evaluated tools in a four-segment space.

In 1999, the Data & Analysis Center for Software (DACS) released one of its state-of-the-art reports, consisting of a thorough survey of data mining techniques, with emphasis on applications to software engineering, which includes a list of 55 products with both summary information along a number of technical as well as process-dependent features and detailed descriptions of each product (Mendonca & Sunderhaft, 1999). Exclusive Ore, Inc. released another study in 2000, including a list of 21 products, defined mostly by the algorithms they implement together with a few additional technical dimensions (Exclusive Ore, 2000).

In 2004, an insightful article discussing high-level considerations in the choice of a DM suite—highlighting the fact that no single suite is best overall—together with a comparison of five of the then most widely used commercial DM software packages, was published in a well-read trade magazine (Nisbett, 2004). About the same time, an extensive and somewhat formal DM tool characterization was proposed, along with a dynamic database of the most popular commercial and freeware DM tools on the market (Giraud-Carrier & Povel, 2003).[1] In a most recent survey of the field, the primary factors considered in selecting a tool were highlighted, along with a report on tool usage and challenges faced (Rexer et al, 2007).

Finally, it is worth mentioning a number of lists of DM tools that, although not including any characterization or evaluation, provide an useful starting point for tool evaluation and selection exercises by centralizing (and generally maintaining in time) basic information for each tool and links to the vendor's homepage for further details (KDNet, 2007; KDnuggets, 2007; Togaware, 2007).

MAIN FOCUS

The target audience of this chapter is business decision-makers. The proposed characterization, and accompanying database, emphasize the complete Data Mining process and are intended to provide the basis for informed, business-driven tools comparison and selection. Much of the content of this chapter is an extension of the work in (Giraud-Carrier & Povel, 2003), with additional insight from (Worthen, 2005; Smalltree, 2007; SPSS, 2007).

The set of characteristics for the description of DM tools can be organized naturally in a hierarchical fashion, with the top level as depicted in Figure 1. Each branch is expanded and discussed in the following sections.

Business Context and Goals

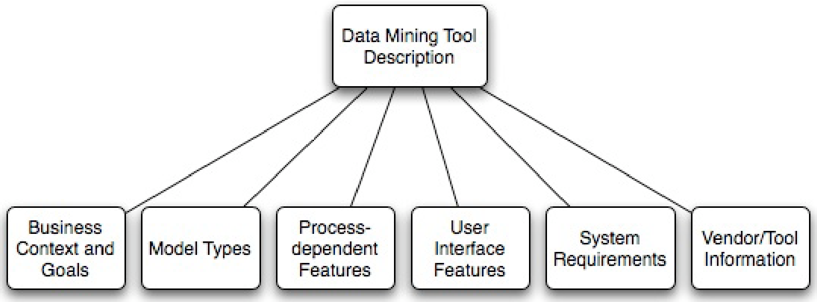
Data mining is primarily a *business-driven process* aimed at the discovery and consistent use of profitable knowledge from corporate data. Naturally, the first questions to ask, when considering the acquisition of supporting DM tools, have to do with the business context and goals, as illustrated in Figure 2.

Since different business contexts and objectives call for potentially different DM approaches, it is critical to understand what types of questions or business problems one intends to solve with data mining, who will be responsible for executing the process and presenting the results, where the data resides, and how the results of the analysis will be disseminated and deployed to the business. Answers to these high-level questions provide necessary constraints for a more thorough analysis of DM tools.

Model Types

The generative aspect of data mining consists of the building of a model from data. There are many available (machine learning) algorithms, each inducing one of a variety of types of models, including predictive models (e.g., classification, regression), descriptive models (e.g., clustering, segmentation), dependency

Figure 1. Top level of the DM tool description hierarchy


Data Mining Tool Description
Business Context and Goals
Model Types
Process-dependent Features
User Interface Features
System Requirements
Vendor/Tool Information