Chapter 103
Optimization of a Hybrid Methodology (CRISP-DM)

José Nava
Centro de Investigación en Ciencias Aplicadas para la Industria, México

Paula Hernández
ITCM, México

ABSTRACT

Data mining is a complex process that involves the interaction of the application of human knowledge and skills and technology. This must be supported by clearly defined processes and procedures. This Chapter describes CRISP-DM (Cross-Industry Standard Process for Data Mining), a fully documented, freely available, robust, and non proprietary data mining model. The chapter analyzes the contents of the official Version 1.0 Document, and it is a guide through all the implementation process. The main purpose of data mining is the extraction of hidden and useful knowledge from large volumes of raw data. Data mining brings together different disciplines like software engineering, computer science, business intelligence, human-computer interaction, and analysis techniques. Phases of these disciplines must be combined for data mining project outcomes. CRISP-DM methodology defines its processes hierarchically at four levels of abstraction allowing a project to be structured modularly, being more maintainable, scalable and the most important, to reduce complexity. CRISP-DM describes the life cycle of a data mining project consisting of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

1. BRIEF HISTORY

Developed by industry experts and leaders, with input from more than 200 data mining users, data mining tools and service providers from all over the world. In late 1996, CRISP-DM was conceived by three companies, experts in the young and immature data mining market. DaimlerChrysler (then Daimler-Benz) ahead of most industrial and commercial organizations applying data mining techniques in its business. SPSS (then ISL), providing data mining services since 1990 and launched the first commercial data mining workbench – Clementine – in 1994. NCR, with teams of data mining consultants and specialists.

DOI: 10.4018/978-1-4666-2455-9.ch103

In the 1990's, data mining market was showing an exponential demand in several countries, which was both exciting and terrifying. All data mining users were developing their approaches on demand, and as they went along. Every data mining developer was learning by trial and error. Was that the best approach? Were they doing right? And the most important, how could they demonstrate to the world's prospective customers that data mining was mature enough to be adopted as a key part of their business processes? Then they thought that a standard process model, non-proprietary and freely available, would address those issues for them and for all practitioners.

A year later, they formed a consortium, and come out with the acronym (Cross-Industry Standard Process for Data Mining). They obtained funding from the European Commission and begun to work in their initial ideas. As this methodology was intended to be industry, tool, and application neutral, they knew they had to get input from as wide range as possible of data mining practitioners and others (such data warehouse vendors, and consultancies) with vested interest in this area.

They launched the CRISP-DM Special Interest Group (The SIG, as it became known), by broadcasting an invitation to interested parties to join in Amsterdam for a workshop to share ideas, invite people to present their, and openly discuss how to work together in CRISP-DM project.

This event significantly surpassed expectations: Twice as many people turned up as they had initially expected, there was an overwhelming consensus that the data mining industry needed urgent standardization, and there was a tremendous common ground in how people viewed the general process of data mining.

Once the workshop was ended, they felt confident they could deliver, along with the SIG's input and work, a standard process model to service the data mining community.

They worked hard the next two and a half years developing CRISP-DM, they ran trials in large-scale data mining projects in world-class companies like Mercedes-Benz, and OHRA.

They worked on the integration of CRISP-DM commercial data mining tools. The SIG was invaluable, with a growing to over 200 members and holding workshops in London, New York and Brussels.

As of mid-1999, they produced a draft of the process model. Actual version, CRISP-DM 1.0 is not radically different. In later years, CRISP-DM was strongly tested by DaimlerChrysler in a wide range of applications, having a big success. SPSS' and NCR's Professional Services groups have adopted CRISP-DM and used it successfully on numerous customer engagements covering many industries and business problems. Since this version, interested institutions have been releasing Data Access, Model Generation and Model Representation Standards for data mining, which we will present later in this chapter.

2. CRISP-DM METHODOLOGY

2.1. Description of the Methodology

The most important and core in CRISP-DM methodology is that defines its processes hierarchically, consisting of a set of tasks at four levels of abstraction, from general to specific, explained (Figure1):

- Using levels of abstraction, at the top level we have the phases, the data mining process is organized into a number of phases, where each phase consists of several second-level generic tasks.
- At the second level we have the generic tasks, which must be general enough to cover all possible data mining situations and must be as complete and stable as possible. Complete means covering all possible applications and the whole process