Expert Systems with Applications 36 (2009) 4114–4124

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Framework for formal implementation of the business understanding phase of data mining projects

Sumana Sharma *, Kweku-Muata Osei-Bryson

Virginia Commonwealth University, United States

ARTICLE INFO

Keywords:
Data mining
Framework
Business understanding
CRISP-DM

ABSTRACT

Various data mining methodologies have been proposed in the literature to provide guidance towards the process of implementing data mining projects. The methodologies describe a data mining project as comprised of a sequence of phases and highlight the particular tasks and their corresponding activities to be performed during each of the phases. It seems that the large number of tasks and activities, often presented in a checklist manner, are cumbersome to implement and may explain why all the recommended tasks are not always formally implemented. Additionally, there is often little guidance provided towards how to implement a particular task. These issues seem to be especially dominant in case of the business understanding phase which is the foundational phase of any data mining project. In this paper, we present an organizationally grounded framework to formally implement the business understanding phase of data mining projects. The framework serves to highlight the dependencies between the various tasks of this phase and proposes how and when each task can be implemented. An illustrative example of a credit scoring application from the financial sector is used to exemplify the tasks discussed in the proposed framework.

1. Introduction

Data mining (DM) has been defined as the non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Frawley, Piatetsky-Shapiro, & Matheus, 1992). Research in data mining has addressed a broad range of applications as diverse as sales and customer relationship management (Berry & Linoff, 1997, 2000; Hung, Yen, & Wang, 2006), financial forecasting (Chun & Park, 2006), fraud detection (Fawcett & Provost, 1997), gene mapping (Kantardzic & Zurada, 2005) and mining of health care data (Alonso, 2002; Phillips-Wren, Sharkey, &Dy,Dy,) to name a few. The interest in the field of DM has surged mainly due to the rapid growth in size of data generated and collected by companies (Han & Kamber, 2006). A recent KD nuggets poll (June 2007), http://www.kdnuggets.com/polls/ based on the largest data size data-mined found that nearly 22% of the respondents reported mining databases of 1 terabyte or more which is double the 11.5% of respondents who mined terabyte size databases in 2006. However there has also been the increasing recognition that mere access to data is not sufficient to learn about interesting patterns found in the data or to uncover novel relationships. There is value in having a formal process that details how the DM project can be implemented (Berry & Linoff, 2000). Data mining (DM) methodologies (Anand & Buchner, 1998; Berry & Linoff, 1997; Cabena, 1998; Cios & Kurgan, 2005; CRISP-DM, 2003; Fayyad, Piatetsky-Shapiro, & Smyth, 1996b) address this issue by providing explicit guidance regarding implementation of data mining projects. They describe the data mining project as consisting of various phases and suggest how each of the phases can be carried out. The methodologies differ somewhat in their prescribed phases and the sequence of these phases, so also the particular tasks needed to implement the various phases However, most methodologies recommend starting a data mining project with developing an understanding of the business domain. This phase generally encompasses determining of business and data mining objectives of the project, the associated success criteria and an assessment of the resources required to execute the project. Certain methodologies such as CRISP-DM (CRISP-DM, 2003), acronym for CRoss Industry Standard Process for Data Mining, also recommend developing a plan for the remaining phases of the project in addition to the above objectives. Fig. 1 shows the process model for the CRISP-DM methodology.

It appears from our review of published data mining case studies that the business understanding (BU) phase of data mining (DM) projects is often implemented in an ad hoc manner. Hardly any published data mining case studies actually provide a detailed description of how this phase was formally implemented. We believe that the reason behind such an unstructured approach is the general lack of support towards how this phase can be

* Corresponding author. Tel.: +1 804 519 8085.
E-mail address: sharmas5@vcu.edu (S. Sharma).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.03.021

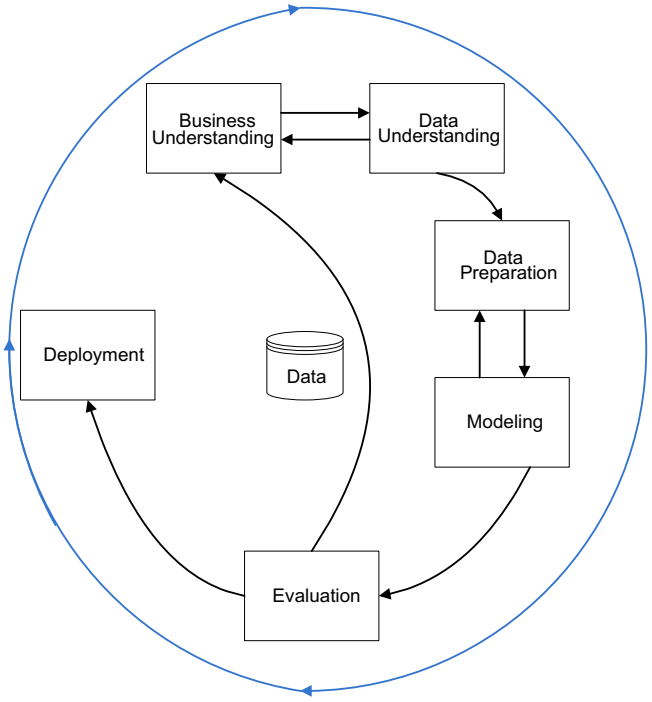

Fig. 1. Phases of CRISP-DM reference model (source: CRISP-DM v.10 guide).

implemented. Charest et al. (2006) believe this to be a broader issue and state that "DM methodologies provide very little detailed advice to the novice miner on how to actually carry out a given step". In our view, this issue is more dominant in case of the BU phase.

It appears that the opposite situation is found in the case of modeling phase, about which relatively more information is generally available. We argue that formally implementing the business understanding phase is just as important as implementing the modeling phase or any other phase of the data mining project. Perhaps, the business understanding phase is even somewhat more important than other phases given that a number of decisions about other phases, such as the modeling as well as other phases (such as data preparation, data understanding, evaluation, etc.) are made, or ideally *should* be made, during the BU phase. Fig. 2 shows how the BU phase is pervasive to all other phases of the DM project.

Not making appropriate decisions during the BU phase seems to lead to two problems. First, it creates inefficiencies as these decisions have to be dealt with in later phases taking away the time and resources that were allocated to accomplish the tasks associated with that phase. The second problem is even more detrimental as not making certain decisions during the BU phase can lead to the DM project taking a completely different direction than what was intended. The second problem originates from the numerous dependencies that exist between the various phases and tasks of a data mining project. These dependencies need to be clearly identified and effectively managed in order to formally implement the BU phase.

Accordingly, the objective of this paper is to present an organizationally grounded framework to implement the various tasks of the BU phase, to identify dependencies existing between the tasks of this phase and to explain the various facets of each task such as its desired output, motivation behind the task, role of organizational actors involved in the task, when it should be performed and its predecessor tasks, and how it can be performed. We use an illustrative example of a typical data mining application from the financial services sector to elucidate our approach. By carefully streamlining the various tasks of the BU phase, the proposed framework allows for formal implementation of the various tasks of this phase. This is likely to result in improving the efficiency and reliability with which such projects can be implemented.

2. Framework for implementation of BU phase

Creating a framework to formally implement the BU phase first requires selection of a DM methodology to serve as an anchor. We reviewed various DM methodologies proposed in the literature (Anand & Buchner, 1998; Berry & Linoff, 1997; Cabena, 1998; Cios & Kurgan, 2005; CRISP-DM, 2003; Fayyad et al., 1996b) to study their suitability of serving as an anchor. Since we intend the framework to be used across all methodologies, we wanted to select a detailed methodology that also covered the various aspects of the BU phase described in other competing DM methodologies. Methodologies that were narrow in focus and did not provide detailed guidance regarding the various aspects were therefore eliminated.

2.1. Selection of CRISP-DM methodology to guide development of framework

Based on our review of the BU phase of various methodologies, we finally selected CRISP-DM (CRISP-DM, 2003) methodology to guide our approach. We selected CRISP-DM due to various reasons. First, this methodology is popularly used in real world organizations. In a 2004 KD Nuggets poll, nearly 42% of the responders reported using CRISP-DM methodology in their data mining projects http://kdnuggets.com/polls/2004/data_mining_methodology.htm. Second, it is more detailed than any other DM methodology and provides extended guidance towards the various tasks to be executed during the BU phase of the project. Third, it covers all the aspects of this phase as described in its competing methodologies. The next section describes the BU phase of the CRISP-DM methodology.

2.2. Business understanding phase of CRISP-DM methodology

Per CRISP-DM methodology, the BU phase of DM projects consists of four different tasks: determination of business objectives, assessment of situation, determination of data mining goals, and production of a project plan. The first task, 'determining of business objectives' aims at developing a thorough understanding from a business perspective, of what the customer really wants to accomplish. Such a task has to take into consideration the fact that the client may have competing objectives and constraints that must be balanced. The second task, 'assessment of situation' is a fact finding exercise about the resources, constraints, assumptions and other factors related to the project. The third task, 'determining of data mining goals' is related to translating the business goal(s) into a data mining goal(s). The fourth and final task, 'production of a project plan' describes the intended plan for achieving the data mining goals(s) and business goal(s). CRISP-DM user guide also describes the desired outputs expected to be generated after accomplishing each of these tasks (see Table 1).

2.3. Framework for implementing the business understanding phase of DM projects

Our framework for implementing the BU phase is inspired from the analogy between the CRISP-DM methodology and the Input–Output Model. The proposed framework describes the dependencies