Expert Systems with Applications 39 (2012) 11335–11348

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Evaluation of an integrated Knowledge Discovery and Data Mining process model

Sumana Sharma *, Kweku-Muata Osei-Bryson, George M. Kasper

Virginia Commonwealth University, School of Business, 301 W Main St., Richmond, VA 23220, United States

ARTICLE INFO

Keywords:
Evaluation
Knowledge Discovery and Data Mining (KDDM) process models
CRISP-DM
IKDDM
Analytical testing

ABSTRACT

Data Mining projects are implemented by following the knowledge discovery process. This process is highly complex and iterative in nature and comprises of several phases, starting off with business understanding, and followed by data understanding, data preparation, modeling, evaluation and deployment or implementation. Each phase comprises of several tasks. Knowledge Discovery and Data Mining (KDDM) process models are meant to provide prescriptive guidance towards the execution of the end-to-end knowledge discovery process, i.e. such models prescribe how exactly each one of the tasks in a Data Mining project can be implemented. Given this role, the quality of the process model used, affects the effectiveness and efficiency with which the knowledge discovery process can be implemented and therefore the outcome of the overall Data Mining project. This paper presents the results of the rigorous evaluation of the Integrated Knowledge Discovery and Data Mining (IKDDM) process model and compares it to the CRISP-DM process model. Results of statistical tests confirm that the IKDDM leads to more effective and efficient implementation of the knowledge discovery process.

1. Introduction

Today, data driven decision making is considered as the cornerstone of modern organizational strategy. It involves the mining of large volumes of data, in the quest for discovering nuggets of knowledge. In recent years Data Mining practitioners and researchers (e.g. CRISP-DM; Cios, Teresinska, Konieczna, Potocka, & Sharma, 2000; Kurgan & Musilek, 2006; Shearer, 2000) have recognized the need for formal Data Mining process models that prescribes the journey from converting data into knowledge. Kurgan and Musilek (2006), noted with regards to Data Mining "*Before any attempt can be made to perform the extraction of this useful knowledge, an overall approach that describes how to extract knowledge needs to be established*". The Knowledge Discovery and Data Mining (KDDM) process is a multiphase process that includes: business understanding (also sometimes referred to as domain understanding), data preparation, modeling, evaluation and deployment or implementation phases (see Fig. 1). The KDDM process is highly iterative and complex, as each phase involves multiple tasks, and there are numerous intra-phase and inter-phase dependencies that exist between the various tasks of the process.

Several KDDM process models have been proposed by researchers and practitioners. Examples include, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996), Cabena, Hadjinian, Stadler, and Verhees (1998), Cios et al. (2000), CRISP-DM (2003), and Berry and Linoff (1997). In a poll conducted by KDNuggets, 42% of the respondents chose CRISP-DM is the main methodology used by them for Data Mining (KDNuggets, 2007).

Sharma and Osei-Bryson (2010) identified some significant limitations in existing KDDM process models and presented an integrated KDDM (IKDDM) process model to address these limitations. Since a KDDM process model is a design artifact, it should be subjected to formal evaluation as such an evaluation provides essential feedback which can then be used to refine the given artifact. It should be noted that to-date there has be no published research studies on formal evaluation of any of the KDDM process models. In this paper, we follow the methodology of Hevner, March, Park, and Ram (2004) to present the results of the formal evaluation of the static qualities of the IKDDM process model. We also compare the performance of the IKDDM process model with that of the CRISP-DM process model.

The rest of the paper is organized as follows: Section 2 provides an overview of the KDDM process and includes a discussion on several serious limitations with previously proposed KDDM process models; Section 3 describes the measurement instrument used for comparing the quality of the IKDDM process model versus the CRISP-DM process model. Section 4 presents our evaluation methodology and the statistical results of the analytical testing and Section 5 presents a discussion of significant findings.

2. Overview of the KDDM process

Knowledge Discovery and Data Mining or KDDM process models serve the purpose of a roadmap or guide, that provide

* Corresponding author. Tel.: +1 804 519 8085.
E-mail addresses: sumanasharma09@gmail.com (S. Sharma), kmuata@isy.vcu.edu (K.-M. Osei-Bryson), gmkasper@vcu.edu (G.M. Kasper).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.044

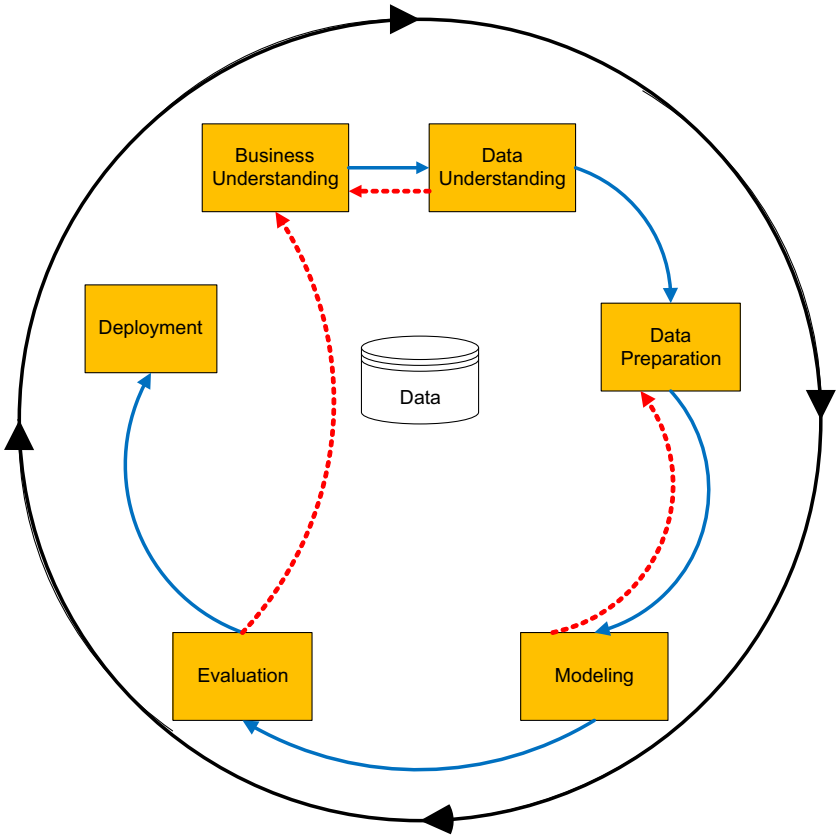

Fig. 1. Typical view of KDDM process models.

prescriptive guidance towards how each task in the end-to-end process can be implemented. They can be regarded as a reference guide or manual that describes 'what' tasks should be executed in the context of a Data Mining project and 'how' they should be executed.

2.1. Overview on phases of the KDDM process

Table 1 provides a brief description of the different phases of the KDDM process, as presented by various process models.

2.2. Limitations of previous KDDM models

Given their prescriptive role and the fact that one essentially relies on a chosen KDDM process model to execute a Data Mining project, it is apparent that the quality of model used to implement the knowledge discovery process has a strong effect on the effectiveness and efficiency with which the relevant Data Mining project can be executed, as well as on the outcome of the project. Our previous detailed review (Sharma & Osei-Bryson, 2010) of the previously proposed KDDM process models revealed that they suffer from several significant limitations that are discussed below.

2.2.1. Checklist oriented description and lack of tool support

While all KDDM process models acknowledge the complexity of the KDDM process, they still describe the complicated KDDM process in terms of a list of steps or tasks. While the steps outlined may be valid (such as formulate a Data Mining objective, chose an appropriate modeling algorithm, evaluate the modeling results, etc.), they are at best, a broad guideline, and do not provide assistance towards "how to" execute the tasks. The issue of lack of tool support by KDDM models has been previously identified in the literature (Charest, Delisle, Cervantes, & Shen, 2006). It is important to note that this problem is especially compounded in the case of a process model such as CRISP-DM which outlines a total of 288 activities to be executed in the context of a Data Mining project. Without support in form of 'how' to execute this long list of activities, it is likely that many of these tasks will be completely overlooked or not adequately implemented when the Data Mining project is implemented.

Given that a KDDM process requires a user to make numerous decisions (Fayyad et al. 1996), it is only necessary that the process models be complemented by support in form of appropriate tools and techniques for carrying out the various tasks. Charest et al. (2006) note that existing process models 'only provide general directives, however what a non-specialist really needs are explanations, heuristics and recommendations on how to effectively carry out the particular steps of the methodology'. Lack of decision support towards tasks can result in certain tasks not being executed during the knowledge discovery process. Given the numerous task–task dependencies, each task helps drive other tasks (its output may be used as input by one or more tasks). Therefore not executing a task can quickly translate into either not implementing or not effectively implementing the succeeding tasks in the model.

2.2.2. Fragmented design

The limitation outlined above leads to another issue in form of the fragmented design of the existing KDDM process models. In other words, the process models do not capture or highlight the important dependencies that exist in a typical KDDM process. By dependencies we mean the interrelationships between the various steps, or between the various phases and tasks (of the same and different phases) of a KDDM project. The dependency which is most obvious from Fig. 1 is the phase-phase dependency resulting