Expert Systems with Applications 39 (2012) 1156–1164

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A context-aware data mining process model based framework for supporting evaluation of data mining results

Kweku-Muata Osei-Bryson [*]

Department of Information Systems & The Information Systems Research Institute, Virginia Commonwealth University, Richmond, VA 23284, USA

ARTICLE INFO

Keywords:
Context
Data mining process
KDDM
Evaluation
Decision analysis
Multi-criteria decision analysis
Post-processing

ABSTRACT

The knowledge discovery via data mining process (KDDM) is a multiple phase that aims to at a minimum semi-automatically extract new knowledge from existing datasets. For many data mining tasks, the evaluation phase is a challenging one for various reasons. Given this challenge several studies have presented techniques that could be used for the semi-automated evaluation of data mining results. When taken together, these studies suggest the possibility of a common multi-criteria evaluation framework. The use of such a multi-criteria evaluation framework, however, requires that relevant objectives, measures and preference function be identified. This implies that the context of the DM problem is particularly important for the evaluation phase of the KDDM process. Our framework utilizes and integrates a pair of established tightly coupled techniques (i.e. Value Focused Thinking (VFT) and the Goal–Question–Metric (GQM) methods) as well as established techniques from multi-criteria decision analysis in order to explicate and utilize context information in order to facilitate semi-automated evaluation.

1. Introduction

The knowledge discovery via data mining process (KDDM) is a multiple phase process (see Fig. 1) that aims to, at a minimum, semi-automatically extract new knowledge from existing data sets. This process that has been described in various ways (e.g. Cios, Teresinska, Konieczna, Potocka, & Sharma, 2000; Shearer, 2000) but essentially consists of the following steps: Business (or Application Domain) Understanding (which includes definition of business and data mining goals), Data Understanding, Data Preparation, Data Mining (or Modeling), Evaluation (e.g. evaluation of results based on Data Mining goals), and Deployment (Kurgan & Musilek, 2006). CRISP-DM (cross-industry standard procedure for data mining), the most popular of the KDDM process model was developed by multi-industry collective of practitioners after the practitioner community became aware of the need for formal data mining process models that prescribe the journey from data to discovering knowledge. The original model was further extended by researchers (e.g. Cios et al., 2000; Sharma & Osei-Bryson, 2010).

For many data mining tasks, the evaluation phase is a challenging one for various reasons. For example, with regard to decision tree (DT) induction although the performance measures may be clear (e.g. accuracy, simplicity, lift), challenges include the need to evaluate a large number of DTs. Gersten, Wirth, & Arundt (2000) noted that regards to setting parameter values, there is "*no practicable approach to select ... the most promising combinations early in the process*" *and as such* "*it is necessary to experiment with different combinations*" but "*it is very hard to compare that many models and pick the optimal one reliably*". Given this challenge Osei-Bryson (2004) proposed an approach for comparing and selecting the 'optimal' decision tree (DT) model given preference and value functions specified by the domain expert(s). Choi, Ahn, and Kim (2005) and Chen (2007) presented approaches for prioritizing association rules. Osei-Bryson (2005, 2010) also presented approaches for selecting the most appropriate segmentation. Overall these papers describe techniques that could be used for the semi-automated evaluation of data mining results. When taken together, these papers suggests the possibility of a common context-aware multi-criteria framework for evaluating the results of data mining that accommodates multiple performance measures, supports adequate data mining experimentation and the non-burdensome semi-automated evaluation of results from the application of data mining techniques. The use of such a multi-criteria evaluation framework, however, requires that relevant objectives, measures and preference function be identified.

This implies that the context of the DM problem is particularly important for the evaluation phase of the KDDM process. The Stakeholders, Business Objectives, Data Mining Objectives and associated performance measures, and the preference function are the major important elements of the context of the particularly DM problem, with the stakeholders' perspectives being a major factor for determining the other elements. Given the identification and definition of the objectives, associated measures and preference

[*] Tel.: +1 804 827 3632; fax: +1 804 828 3199.
E-mail addresses: Kweku.Muata@isy.vcu.edu, KMOsei@VCU.Edu

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.07.117

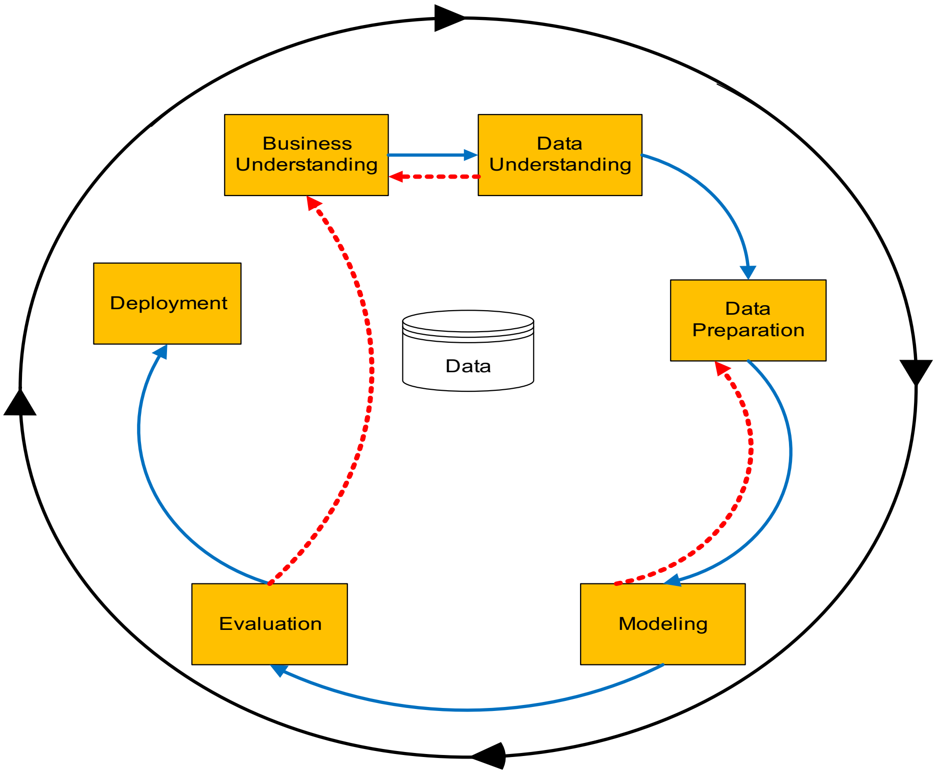

Fig. 1. CRISP process model (CRISP-DM, 2003).

function, then a multi-criteria approach could be used to automatically determine the ranking of the data mining results during the evaluation phase. Several studies including Osei-Bryson (2004, 2005, 2007, 2008, 2010), Choi et al. (2005) and Chen (2007) have offered this type of context-aware multi-criteria approach for post-processing. However, apart from Osei-Bryson (2010), the solution methods of those studies were not explicitly situated within the context of KDDM process models and none (including Osei-Bryson, 2010) described how the implications of a given problem context could be explicated in a manner that would facilitate the evaluation of DM output.

As noted by Kurgan and Musilek (2006), 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*". In this paper we present a KDDM process model based common context-aware multi-criteria framework for evaluating data mining results that includes the explication of business and data mining objectives and performance measures. Our research problem can be considered to involve context-aware support for the selection of *a limited set of the 'best'* models (Zopounidis & Doumpos, 2002) in order to reduce the cognitive burden on the domain experts in the evaluation phase of the KDDM process.

2. Description of proposed framework

2.1. Description of process

In this section we will both describe our extended KDDM process for doing context-aware evaluation. Our description covers the Business Understanding (BU), Modeling/Data Mining (DM), and Evaluation (EV) phases, and assumes that activities that are equivalent to other phases (e.g. Data Understanding, and Data Preparation steps) are done in the usual manner (see Table 1).

Sharma and Osei-Bryson (2010) in an earlier work explicated the major links and outputs between the phases of the KDDM process model (e.g. Fig. 2). Assuming that the Stakeholders, Business Objectives, Data Mining Objectives and associated performance measures, and the preference function are the major elements of the context of the context of a given DM problem, then this context would be explicated in the Business Understanding (BU), and utilized in later phases of the KDDM process.

Given the activities listed in Table 1, outputs of the BU phase would include the set of performance measures (i.e. substeps BU: b, d, e), preference function (i.e. substep BU: j), value functions (i.e. substep BU: k). Given the preference function (e.g. including weights if relevant) and the set of data mining performance measures it is then possible to use multi-criteria decision making (MCDM) techniques such as the AHP to do automatic ranking in the evaluation phase of the multiple data mining models that would have been generated in the DM phase.

2.2. Generating data mining goals from business goals

Several approaches could be used to identify the appropriate set of Data Mining Objectives (DMO) that correspond to given a set of Business Objectives (BO). One such approach is the Value-Focused Thinking (VFT) which was proposed by Keeney (1992, 1996) and which provides explicit guidance on the formulation of objectives, an indispensable task in any decision making situation. VFT has been applied across a wide variety of domains (Kajanus, Kangas, & Kurtilla, 2004), and systems engineering (Boylan, Tollefson, Kwinn, & Guckert, 2006).

VFT assumes three different types of objectives: fundamental objectives (FO), mean objectives (MO), and strategic objectives. Fundamental objectives concern the ends that decision makers value in a particular decision context whereas means objectives are the methods to achieve the ends. Strategic objectives provide common guidance for more detailed fundamental objectives.