Modeling the KDD Process

Vasudha Bhatnagar
University of Delhi, India

S. K. Gupta
IIT, Delhi, India

INTRODUCTION

Knowledge Discovery in Databases (KDD) is classically defined as the "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large databases" (Fayyad, Piatetsky-Shapiro & Smyth, 1996a). The recently developed KDD technology is based on a well-defined, multi-step "**KDD process**" for discovering knowledge from large data repositories. The basic problem addressed by the KDD process is one of mapping low-level data (operational in nature and too voluminous) to a more abstract form (descriptive approximation or model of the process that generated the data) or a useful form (for example, a predictive model) (Fayyad, Piatetsky-Shapiro & Smyth, 1996b). The KDD process evolves with pro-active intervention of the domain experts, data mining analyst and the end-users. It is a 'continuous' process in the sense that the results of the process may fuel new motivations for further discoveries (Chapman et al., 2000). Modeling and planning of the KDD process has been recognized as a new research field (John, 2000).

In this chapter we provide an introduction to the *process of knowledge discovery in databases (KDD process),* and present some models (conceptual as well as practical) to carry out the KDD endeavor.

BACKGROUND

The process of Knowledge Discovery in Databases consists of multiple steps, and is inherently iterative in nature. It requires human interaction for its applicability, which makes the process subjective. Various parameters require to be adjusted appropriately before the outcome of the process can be applied for decision making.

The process starts with the task of understanding the domain in the context of the goal of the endeavor, and ends with the task of interpretation and evaluation of the discovered patterns. Human centric nature of the process has been emphasized since the early days of inception of the KDD technology (Brachman & Anand, 1996) and vindicated by the veterans (Ankerst, M. 2002). The core of KDD process employs "data mining" algorithms, which aim at searching for interesting models and patterns in the vast search space, and are responsible for actually discovering nuggets of knowledge (Fayyad, Piatetsky-Shapiro & Smyth, 1996a). Often, the key to a successful KDD effort lies not so much in the act of data mining alone but in the pre and post mining steps of the process. Understanding the whole KDD process therefore becomes imperative for the success of the knowledge discovery exercise. Further, a formal model of the KDD process also helps in differentiating and comparing two or more approaches for KDD endeavor.

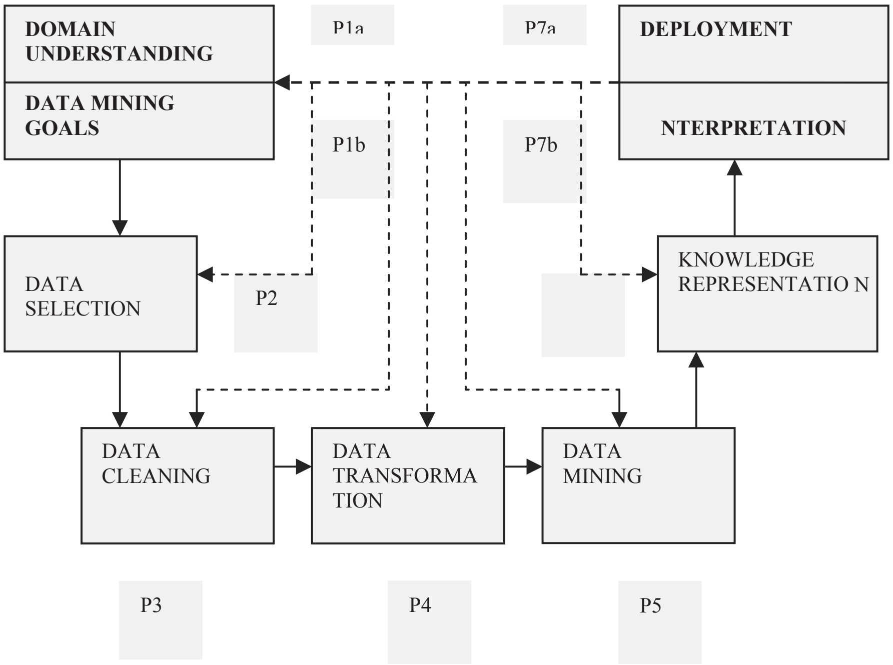
GENERIC STEPS OF THE KDD PROCESS

Figure 1 shows a simple model of the KDD process exhibiting the logical sequencing of the various process steps. The model allows the data miner to effortlessly map the logical process steps (P1 to P7) to the corresponding physical computing processes.

The data flows in a straight forward manner from each process step to the subsequent step as shown by solid lines. The dash lines show the control flow and indicate optional iteration of process steps after the discovered knowledge has been evaluated. We describe below the generic steps of a KDD process.

Domain Understanding and Defining Data Mining Requirements (Step P1(a, b)) are the tasks that

Figure 1. Steps of the KDD process



comprise the first step of the KDD process. Identifying and defining the project goals, assessing the feasibility, planning and preparation of subsequent steps, are some of the sub-tasks that need to be performed during this step. Understanding of the basic characteristics of available data and knowledge of the domain, play an important role in crystallizing the project objectives. Experience and expertise of the domain experts, decision-makers and the end users help in translating the project objectives into knowledge discovery goals.

Data Selection (Step P2) involves laying down the criteria for including the data to be analyzed or excluding the unwanted data. Since all the data may not be relevant for the knowledge discovery goals, the data mining analyst selectively identifies the relevant data, depending on the business questions that need to be answered. For instance, in order to explore the buying patterns of the customers of a particular geographical area, customer addresses can be the selection criterion.

Data Cleaning (Step P3) is performed to ensure domain and semantic validity of data with respect to the goals set in Step P1b. Faulty data capturing methods, transmission errors or legal issues are some causes of "noisy data" in a database apart from the practical data gathering pitfalls. Strategies to handle missing and noisy data have to be decided by the end user and/or domain expert, and must be implemented faithfully before applying the mining algorithm.

Data cleaning assumes added significance for the successful outcome of the KDD process, since most of the mining algorithms do not address the problem of dirty or missing data.

Data Transformation (Step P4) is often required because of syntactic reasons. For example, in a distributed database if the salary attribute is stored in different currencies then it has to be transformed to a pre-decided common currency before undertaking mining. Some data mining algorithms require data to be in a specific format. Transforming numerical attributes to categorical attributes is commonly required before applying *clustering* or *classification* algorithms. For *multilevel association rule* mining, data transformations are required to facilitate mining at different levels of abstraction.

Data Mining (Step P5) results into discovery of hidden patterns by applying a suitable mining algorithm over the "*pre-processed*" (selected, cleaned and transformed) data. The choice of the mining algorithm is influenced by the specification of the discovery goals, type of knowledge to be discovered and nature of data. The mining analyst may have to re-engineer an algo-