Data Mining and Knowledge Discovery in Databases

D

Ana Azevedo
Polytechnic Institute of Porto, ISCAP, Portugal & Algoritmi R&D Center, University of Minho, Portugal

INTRODUCTION

The term knowledge discovery in databases or KDD, for short, was coined in 1989 to refer to the broad process of finding knowledge in data, and to emphasize the "high-level" application of particular Data Mining (DM) methods (Fayyad, Piatetski-Shapiro, & Smyth, 1996). Fayyad considers DM as one of the phases of the KDD process. The DM phase concerns, mainly, the means by which the patterns are extracted and enumerated from data. The literature is sometimes a source of some confusion because the two terms are indistinctively used, making it difficult to determine exactly each of the concepts (Benoît, 2002). Nowadays, the two terms are, usually, indistinctly used.

Efforts are being developed in order to create standards and rules in the field of DM with great relevance being given to the subject of inductive databases (De Raedt, 2003) (Imielinski & Mannila, 1996). Within the context of inductive databases a great relevance is given to the so called DM languages.

This article presents a comprehensive introduction and summary of the main basic concepts and bibliography in the area of DM, nowadays. Thus, the main contribution of this article is that it can be considered as a good starting point for newcomers in the area.

The remaining of this article is organized as follows. Firstly, DM and the KDD process are introduced. Following, the main DM tasks, methods/algorithms, and models/patterns are organized and succinctly explained. SEMMA and CRISP-DM are next introduced and compared with KDD. A brief explanation of standards for DM is then presented. The article concludes with possible future research directions and conclusion.

BACKGROUND

In recent years, we have witnessed the growth and consolidation of the DM area. Since the first Workshop, IJCAI-89 Workshop on Knowledge Discovery in Databases, which took place at Detroit in 1989 and that led, in 1995, to the nowadays main annual conference in the area, ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, the number of publications and conferences dedicated to the area presents a significant growth. These conferences as well as several seminal article, helped in the consolidation of the area. Since then, the evolution has been overwhelming, and DM can be considered as a consolidated research area.

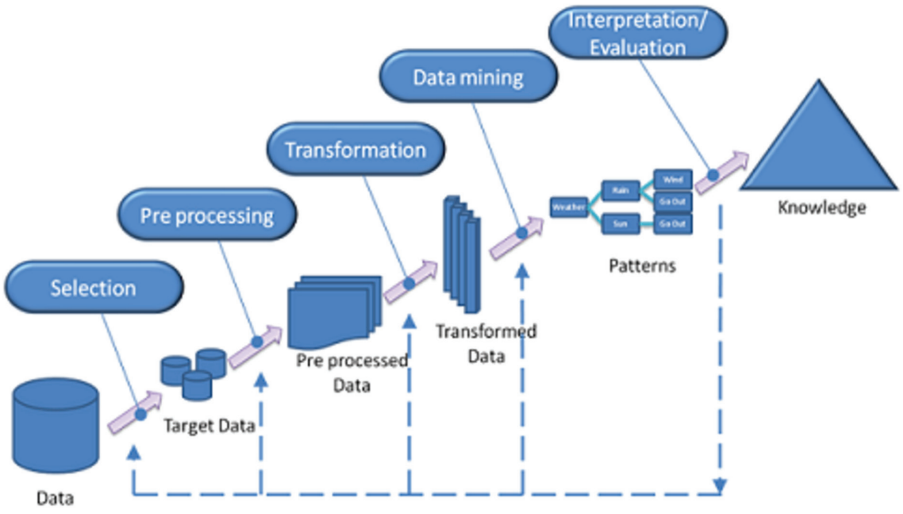
DATA MINING AND THE KNOWLEDGE DISCOVERY IN DATABASES PROCESS

"The KDD process, as presented in (Fayyad, Piatetski-Shapiro, & Smyth, 1996), is the process of using DM methods to extract what is considered knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. There are five stages considered, namely, selection, preprocessing, transformation, data mining, and interpretation/evaluation as presented in Figure 1:

- **Selection**: This stage consists on creating a target data set, or on focusing in a subset of variables or data samples, on which discovery is to be performed;

DOI: 10.4018/978-1-4666-5888-2.ch164

Figure 1. The KDD process



- **Preprocessing**: This stage consists on the target data cleaning and preprocessing in order to obtain consistent data;
- **Transformation**: This stage consists on the transformation of the data using dimensionality reduction or transformation methods;
- **Data Mining**: This stage consists on the searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction);
- **Interpretation/Evaluation**: This stage consists on the interpretation and evaluation of the mined patterns." (Azevedo & Santos, 2008, p. 183)

The KDD process is preceded by the development of an understanding of the application domain, the relevant prior knowledge, and the goals of the end-user. It must be continued by knowledge consolidation, incorporating this knowledge into the system. The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user (Brachman & Anand, 1996)

As of the foundations of KDD and DM, several applications were developed in many diversified fields. The growth of the attention paid to the area emerged from the rising of big databases in an increasing and differentiated number of organizations. Nevertheless, there is the risk of wasting all the value and wealthy of information contained in these databases, unless the adequate techniques are used to extract useful knowledge (Chen, Han, & Yu, 1996; Fayyad, 1997; Simoudis, 1996). The application of DM techniques with success can be found in a wide and diversified range of applications, for instance, bioinformatics, ecology and sustainability, finance, industry, marketing, scientific research, telecommunications, and other applications (Azevedo & Santos, 2011) .

DATA MINING TASKS, METHODS/ALGORITHMS, AND MODELS/PATTERNS

Prediction and description were identified as the two "high-level" primary goals of DM (Fayyad, Piatetski-Shapiro, & Smyth, 1996). "Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Description focuses on finding human-interpretable patterns on finding the data" (Fayyad, Piatetski-Shapiro, & Smyth, 1996, p. 12).

To achieve these goals some DM tasks were used and its description can be found in the literature. Some of the most common DM tasks are classification, prediction, clustering, association, and summarization: