Chapter 37
Data Mining and Knowledge Discovery in Databases

Ana Azevedo
Polytechnic Institute of Porto, Portugal

ABSTRACT

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. The DM phase concerns, mainly, the means by which the patterns are extracted and enumerated from data. 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. Within the context of inductive databases, a great relevance is given to the so-called DM languages. This chapter explores DM in KDD.

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. 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 chapter presents a comprehensive introduction and summary of the main basic topics and bibliography in the area of DM, nowadays. Thus, the main contribution of this chapter 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 methodologies 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.

DOI: 10.4018/978-1-5225-7598-6.ch037

BACKGROUND

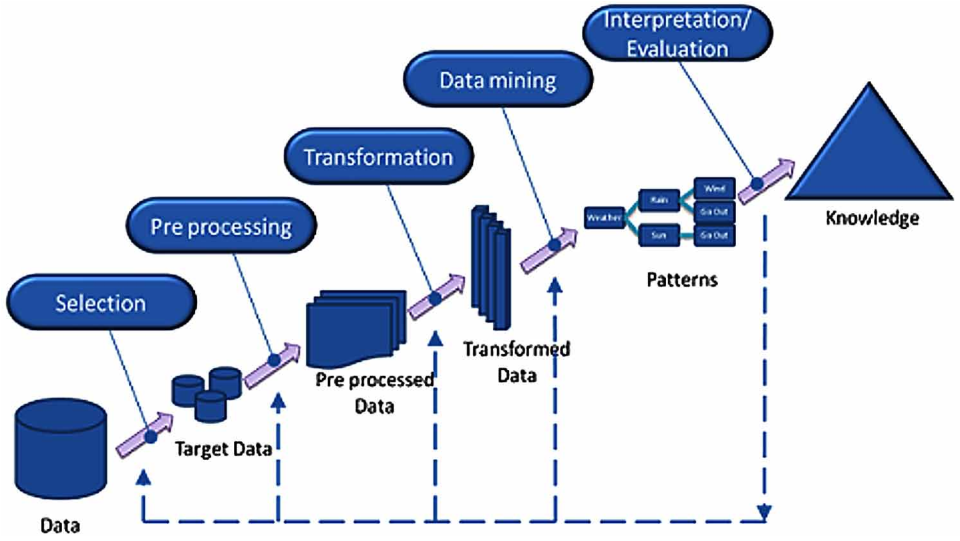
In recent years, we have witnessed the growth and consolidation of the DM area. Since the first Workshop, IJCAI-89 Workshop on Knowledge Discovery in Databases, which took place at Detroit in 1989 and that led, in 1995, to the nowadays main annual conference in the area, ACM SIGKDD International Conference 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 papers, helped in the consolidation of the area. Since then, the evolution has been overwhelming, and DM can be considered as a consolidated research area (Azevedo, 2015).

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;
- **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);

Figure 1. The KDD process


Interpretation/
Evaluation
Data mining
Transformation
Pre processing
Selection
Data
Target Data
Pre processed
Data
Transformed
Data
Weather
Rain
Sun
Wind
Go Out
Go Out
Patterns
Knowledge