# Data Mining and Knowledge Discovery in Databases

D

**Ana Azevedo**
*Polytechnic Institute of Porto, 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. 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.

## 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

DOI: 10.4018/978-1-5225-2255-3.ch166

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

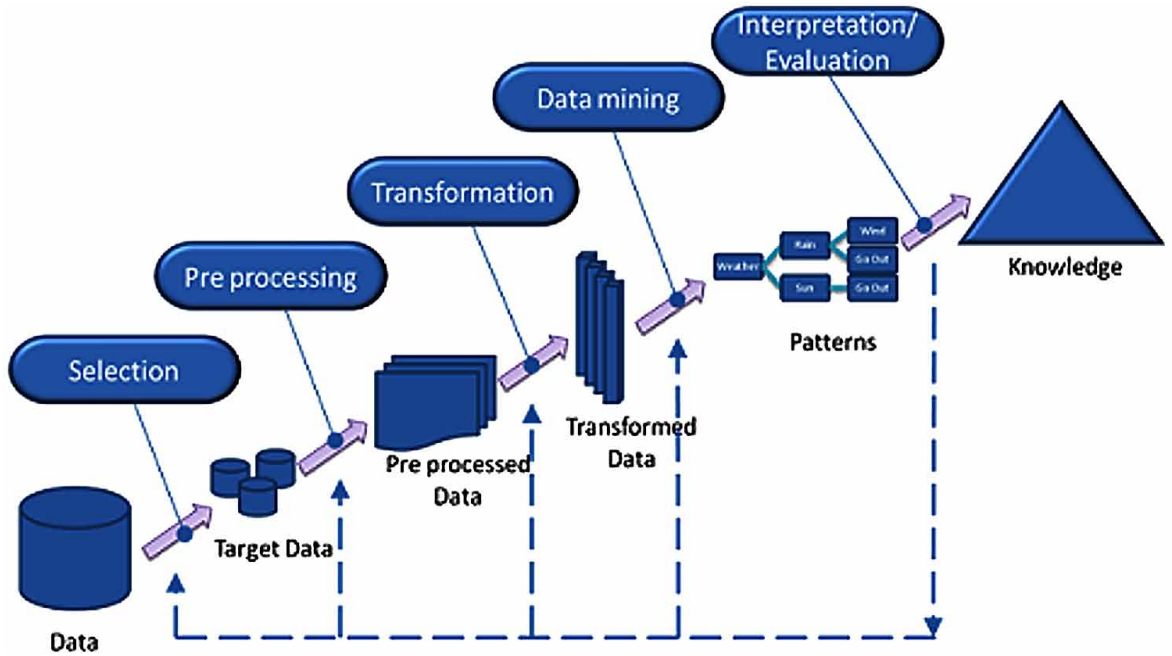
of variables or data samples, on which discovery is to be performed;

- **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);
- **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, J., & 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 U. M., 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).

*Figure 1. The KDD process*



10 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/data-mining-and-knowledge-discovery-in-databases/183906

## Related Content

A Semiosis Model of the Natures and Relationships among Categories of Information in IS
Tuan M. Nguyenand Huy V. Vo (2013). *International Journal of Information Technologies and Systems Approach (pp. 35-52).*
www.irma-international.org/article/a-semiosis-model-of-the-natures-and-relationships-among-categories-of-information-in-is/78906

An Efficient Random Valued Impulse Noise Suppression Technique Using Artificial Neural Network and Non-Local Mean Filter
Bibekananda Jena, Punyaban Pateland G.R. Sinha (2018). *International Journal of Rough Sets and Data Analysis (pp. 148-163).*
www.irma-international.org/article/an-efficient-random-valued-impulse-noise-suppression-technique-using-artificial-neural-network-and-non-local-mean-filter/197385

GIS-Based Quantitative Landslide Risk Assessment Approach for Property and Life at Bartin Hepler
Arzu Erener, Gülcan Sarpand ebnem Düzgün (2021). *Encyclopedia of Information Science and Technology, Fifth Edition (pp. 1623-1636).*
www.irma-international.org/chapter/gis-based-quantitative-landslide-risk-assessment-approach-for-property-and-life-at-bartin-hepler/260292

Creativity of End Users in Theory and in Practice
Malgorzata Pankowska (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 4080-4089).*
www.irma-international.org/chapter/creativity-of-end-users-in-theory-and-in-practice/112851

An Approach to Distinguish Between the Severity of Bullying in Messages in Social Media
Geetika Sarnaand M.P.S. Bhatia (2016). *International Journal of Rough Sets and Data Analysis (pp. 1-20).*
www.irma-international.org/article/an-approach-to-distinguish-between-the-severity-of-bullying-in-messages-in-social-media/163100