Data Mining and the KDD Process

D

Ana Funes
Universidad Nacional de San Luis, Argentina

Aristides Dasso
Universidad Nacional de San Luis, Argentina

INTRODUCTION

The constant search for diseases causes, the improvement of automatic diagnoses methods, financial data analytics and market tendencies, among others, are only some of the innumerable applications where analysis and discovery of new patterns have fuelled the research and development of new methods, all related to machine intelligence, knowledge extraction from what is now being called 'big data', Knowledge Discovery in Databases or KDD, and Data Mining.

The development of these fields has benefited from the existence of large volumes of data proceeding from the most diverse sources and domains, e.g. entrepreneurial historical data bases, medical data bases, biological data bases, astronomical data, etc. KDD process and methods of Data Mining allows for the discovery of knowledge in data that is hidden to humans, particularly when data volumes are large or even extremely large, presenting the knowledge extracted under different ways: rules, equations, decision trees, etc., and helping to answer questions such as *what are the groups from a population of individuals with common characteristics?*, *is this client reliable?*, *is this e-mail spam?,* etc.

Answers to these questions, as well as to many others, are different from the traditional answers obtained from queries in On Line Transactions Processing (OLTP), where the information is not hidden neither is discovered, but it is presented summarized in an agreed format or report. They also differ from information proceeding from Online Analytical Processing (OLAP), which can be presented in different perspectives or aggregated in different ways and not just summarized as in OLTP, and that can even escalate to the use of big data, where OLTP fails. However, both of these methods are not capable of discover new knowledge neither producing new patterns and rules as is the case with the KDD process.

In this chapter, an overview of the KDD process and all its stages is given, including Data Selection, Cleaning, etc., with especial attention to the phase of Data Mining, its tasks and methods, as well as its relation to other areas such as Machine Learning, Inductive Logic Programming (ILP), Statistics, etc. A discussion of a possible classification of Data Mining methods is also given as well as an overview of future challenges in the field.

BACKGROUND

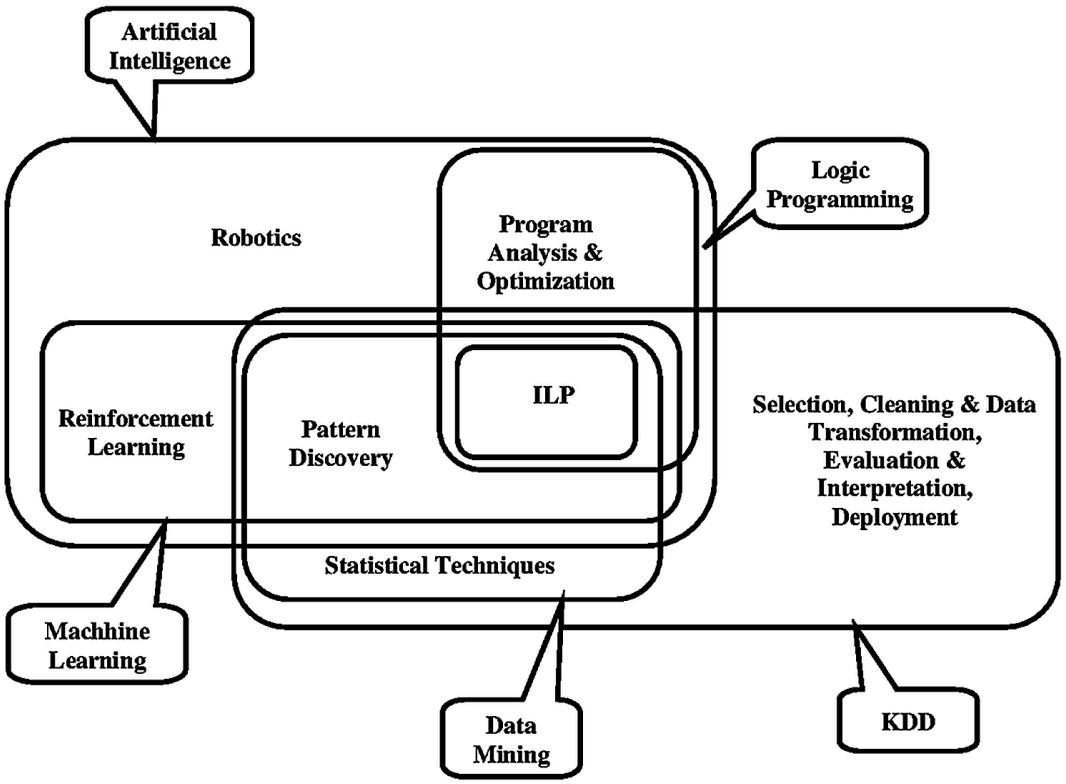
There exists some confusion in the use of the terms of *Knowledge Discovery in Databases* or KDD and *Data Mining*. Frequently these terms are interchanged, using Data Mining as synonym of KDD. Although they are strongly related, it is important to clarify the differences between them.

Several definitions of Data Mining can be found in the literature. Witten and Frank (2000) refers to Data Mining as the process of extraction of previously-unknown, useful and understandable knowledge from big volumes of data, which can be in different formats and come from different sources. In a much more short way, Hernández-Orallo, Ferri and Ramírez-Quintana (2004) define Data Mining as the process of converting data

DOI: 10.4018/978-1-5225-2255-3.ch167

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

Figure 1. Data mining and its relationships with other fields



into knowledge. Sometimes Data Mining is also referred by many other names including *knowledge extraction*, *information discovery*, *information harvesting*, *data archeology*, and *data pattern processing* (Fayyad et al, 1996a).

The notion of Data Mining is not new. Since the 60s, other terms as *Data Fishing* or *Data Dredging* have been used by statisticians to refer to the idea of finding correlations in data without a previous hypothesis as underlying causality. However, it is not until the late 80s that Data Mining became a discipline of Computer Science and scientific community adopted the term. In fact, as Witten and Frank (2005) point out, *the first book on data mining appeared in 1991 (Piatetsky-Shapiro and Frawley, 1991)–a collection of papers presented at a workshop on knowledge discovery in databases in the late 1980s.*

Data Mining is a branch of Artificial Intelligence, closely related to Machine Learning, where Machine Learning provides the technical basis for Data Mining (Witten and Frank, 2005). Data Mining deals with *inductive learning* in a practical and not theoretical way (as Machine Learning does), making use of tools provided by Machine Learning. It applies Machine Learning techniques as well as other statistical and algebraic techniques to find structural patterns hidden in data, with the main objective of describing data or making predictions from them.

Artificial Intelligence comprises not only Machine Learning but other disciplines such as *Robotics*, *Logic Programming* and *Inductive Logic Programming* (ILP), which is a field of Machine Learning and Data Mining. Figure 1 shows a schematic view of the relationships between Data Mining, Machine Learning, Artificial Intelligence and other associated disciplines.

As it can be seen in Figure 1, Data Mining adopts techniques and methods not only from Machine Learning –area with which is closely related– but also from Statistics.

13 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-the-kdd-process/183907

Related Content

Performance Measurement of Technology Ventures by Science and Technology Institutions
Artie W. Ng, Benny C. F. Cheungand Peggy M. L. Ng (2018). *Encyclopedia of Information Science and Technology, Fourth Edition (pp. 4774-4784).*
www.irma-international.org/chapter/performance-measurement-of-technology-ventures-by-science-and-technology-institutions/184182

Reversible Data Hiding Scheme for ECG Signal
Naghma Tabassumand Muhammed Izharuddin (2018). *International Journal of Rough Sets and Data Analysis (pp. 42-54).*
www.irma-international.org/article/reversible-data-hiding-scheme-for-ecg-signal/206876

Improving Health Care Management Through the Use of Dynamic Simulation Modeling and Health Information Systems
Daniel Goldsmithand Michael Siegel (2012). *International Journal of Information Technologies and Systems Approach (pp. 19-36).*
www.irma-international.org/article/improving-health-care-management-through/62026

Assessment in Academic Libraries
Gregory A. Smith (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 4823-4832).*
www.irma-international.org/chapter/assessment-in-academic-libraries/112928

Virtual Reality Exposure Therapy for Anxiety and Specific Phobias
Thomas D. Parsons (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 6475-6483).*
www.irma-international.org/chapter/virtual-reality-exposure-therapy-for-anxiety-and-specific-phobias/113105