Available online at www.sciencedirect.com

ScienceDirect

Physics Procedia 62 (2015) 18 – 22

Physics Procedia

3rd International Conference Frontiers in Diagnostic Technologies, ICFDT3 2013

From Patterns in Data to Knowledge Discovery: What Data Mining Can Do

Francesco Gullo*

Yahoo Labs
Barcelona, Spain
gullo@yahoo-inc.com

Abstract

Data mining is defined as the computational process of analyzing large amounts of data in order to extract patterns and useful information. In the last few decades, data mining has been widely recognized as a powerful yet versatile data-analysis tool in a variety of fields: information technology in primis, but also clinical medicine, sociology, physics.

In this technical note we provide a high-level overview of the most prominent tasks and methods that form the basis of data mining. The note also focuses on some of the most recent yet promising interdisciplinary aspects of data mining.

Peer-review under responsibility of the ENEA Fusion Technical Unit

Keywords: data mining, knowledge discovery, graph mining

1. Knowledge Discovery in Databases and Data Mining

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying novel, valid, potentially useful, and ultimately understandable patterns in data Fayyad et al. (1996a). The term "pattern" refers to a subset of the data expressed in some language or a model exploited for representing such a subset. KDD aims at discovering patterns that *(i)* do not result in straightforwardly computing predefined quantities (i.e., non-trivial), *(ii)* can apply to new data with some degree of certainty (i.e., valid), *(iii)* have been unknown so far (i.e., novel), *(iv)* provide some benefit to the user or to further tasks (i.e., potentially useful), and *(v)* lead to useful insights, immediately or after some post-processing (i.e., understandable).

The KDD process is an iterative and interactive sequence of the following main steps (Figure 1):

- *selection*, whose main goal is to create a target data set from the original data, i.e., selecting a subset of variables or data samples, on which discovery has to be performed;
- *preprocessing*, which aims to "clean" data by performing various operations, such as noise modeling and removal, defining proper strategies for handling missing data fields, accounting for time-sequence information;

* Corresponding author. Tel.: +34-93-183-8891 ; fax: +34-93-183-8901.
E-mail address: gullo@yahoo-inc.com

Peer-review under responsibility of the ENEA Fusion Technical Unit
doi:10.1016/j.phpro.2015.02.005

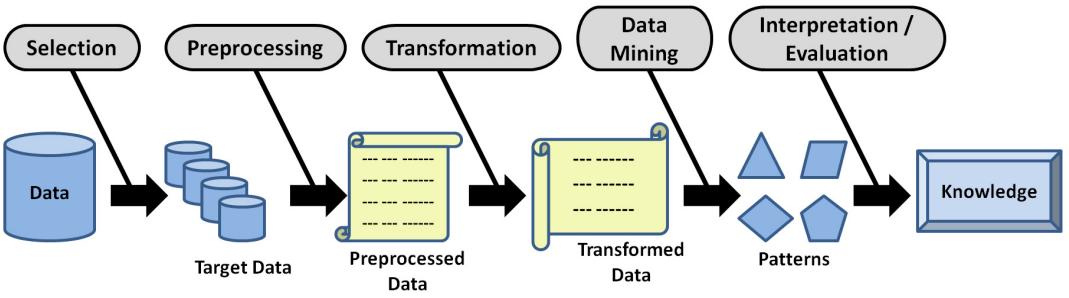

Fig. 1. The Knowledge Discovery in Databases (KDD) process

- *transformation*, which is in charge of reducing and projecting the data, in order to derive a representation suitable for the specific task to be performed; it is typically accomplished by involving transformation techniques or methods that are able to find invariant representations of the data;
- *data mining*, which deals with extracting interesting patterns by choosing *(i)* a specific data-mining *method* or *task* (e.g, summarization, classification, clustering, regression, and so on), *(ii)* proper *algorithm(s)* for performing the task at hand, and *(iii)* an appropriate representation of the output results;
- *interpretation/evaluation*, which is exploited by the user to interpret and extract knowledge from the mined patterns, by visualizing the patterns; this interpretation is typically carried out by visualizing the patterns, the models, or the data given such models and, in case, iteratively looking back at the previous steps of the process.

Data mining represents the "core" step of the KDD process, so much so that the "data mining" and "KDD" terms are often treated as synonyms Han and Kamber (2001). Several definitions of what data mining is have been used, e.g., "*automated yet non-trivial extraction of implicit, previously unknown, and potentially useful information from data*", "*automated exploration and analysis of large quantities of data in order to discover meaningful patterns*", "*computational process of automatically extracting useful knowledge from large amounts of data*". All definitions are all roughly equivalent to each other. They all agree on the main aspects of data mining, which are: *(i)* huge quantity of data that *(ii)* should be analyzed so as to *(iii)* extract what is called "knowledge", or "useful information", or "patterns", i.e., *(iv)* something that can be processed and profitably exploited by human beings.

The importance of data mining nowadays is mainly motivated by the lots of data that is collected and stored by a variety of today's prominent applications. This data includes Web data, e-commerce data, purchases, bank transactions, and so on. Also, the number of applications dealing with data that needs to be processed at enormous speeds (GB/seconds or even more) is rapidly increasing; examples include remote sensors on satellite, telescopes scanning the skies, microarray generating gene-expression data, scientific simulations. Due to the peculiarity of the underlying data, it is apparent that data analysis in such challenging contexts cannot be performed with traditional data-analysis techniques, either manual or automated. Data mining aims at filling this gap, with its intrinsic interdisciplinary nature that poses it at the intersection of a number of more classical fields, such as artificial intelligence, statistics, database systems, machine learning.

2. Data-Mining Tasks

Data mining comprises a number of tasks that can be used, even in combination, based on the requirements of the specific application context. Data-mining tasks are usually classified into *predictive* and *descriptive* Fayyad et al. (1996b). Predictive tasks refer to building a model useful for predicting future behavior or values for certain features. Among others, these include *classification* and *prediction*, i.e., deriving some models (or functions) that describe data classes or concepts by a set of data objects whose class label is known (i.e., the *training set*), so as to predict the class of objects whose class label is unknown; *deviation detection*, i.e., dealing with *deviations* in data, which are defined as differences between measured values and corresponding references such as previous values or normative values; *evolution analysis*, i.e., detecting and describing regular patterns in data whose behavior changes over time. On the