IDEA GROUP PUBLISHING
701 E Chocolate Avenue, Hershey PA17033, USA
Tel: 717/533-8845; Fax: 717/533-8661; URL: www.idea-group.com
ITB7250

Chapter I

Introducing Data Mining and Knowledge Discovery

R. Sarker, H. Abbass and C. Newton
University of New South Wales, Australia

The terms Data Mining (DM) and Knowledge Discovery in Databases (KDD) have been used interchangeably in practice. Strictly speaking, KDD is the umbrella of the mining process and DM is only a step in KDD. We will follow this distinction in this chapter and present a simple introduction to the Knowledge Discovery in Databases process from an optimization perspective.

INTRODUCTION

Our present information age society thrives and evolves on knowledge. Knowledge is derived from information gleaned from a wide variety of reservoirs of data (databases). Not only does the data itself directly contribute to information and knowledge, but also the trends, patterns and regularities existing in the data files. So it is important to be able to, in an efficient manner, extract useful information from the data and the associated properties of the data, i.e., patterns and similarities.

A new area of research, data extraction or data mining, has evolved to enable the identification of useful information existing in the data reservoirs. To understand and recognize the major initiatives in this research area, we will briefly describe the terminology and approaches to data mining and knowledge discovery in databases.

Knowledge discovery in databases (KDD) is the process of extracting models and patterns from large databases. The term *data mining* (DM) is often used as a synonym for the KDD process although strictly speaking it is just a step within KDD. DM refers to the process of applying the discovery algorithm to the data. We

This chapter appears in the book, *Heuristic and Optimization for Knowledge Discovery*, edited by Ruhul Sarker, Hussein Abbass and Charles Newton.

define the KDD process as:

> KDD is the process of model *abstraction* from large databases and *searching* for valid, novel, and nontrivial *patterns* and *symptoms* within the abstracted model.

There are four keywords in the definition: abstraction, search, patterns, and symptoms. The database is a conventional element in KDD.

Abstraction: Abstraction is the process of mapping a system language Λ_1 to approximately an equivalent language Λ_2. The mapping is strong when it maps the system while neither losing existing patterns (completeness) nor introducing new patterns (soundness). Formally, Giunchiglia and Walsh (1992) define an abstraction, written f: $\Sigma_1 \rightarrow \Sigma_2$, as a pair of formal systems (Σ_1, Σ_2) with languages Λ_1 and Λ_2, respectively, and an effective total function f_Λ: $\Lambda_1 \rightarrow \Lambda_2$.

Search: It is more convenient to visualize the discovery process in terms of searching. One can measure the complexity and in turn the feasibility, of the discovery process by studying the search space. Most KDD steps can be seen as search-space reduction techniques. For example, the main goal for creating a target data set, data cleaning, and data reduction and projection is to reduce the noise in the data and to select a representative sample which then reduces the search space. The mining algorithm is the search technique used to achieve the overall process's objective.

Patterns: A pattern is an expression, η, in a language, Λ, describing a subset of facts, $\varphi_\eta \subseteq \varphi$, from all the facts, φ, which exist in the database (Fayyad, Piatetsky-Shapiro & Smyth, 1996c).

Symptoms: Although symptoms can be seen as patterns, the process of discovering the symptoms has more dimensions than finding simple descriptive patterns. Identification of symptoms is a major task for KDD if it is the intention to use it for decision support. The KDD process's role is to clear the noise within the system and discover abnormal signals (symptoms) that may contribute to potential problems.

In this definition, the term *process* implies that KDD consists of different steps such as: data preparation, search for patterns, knowledge evaluation and refinement. The discovered patterns should be *valid* with some degree of certainty, and *novel* (at least to the system and preferably to the users). *Nontrivial* delineates that the discovered patterns should not be obvious in the domain knowledge. They should, however, represent a substantial discovery to the user; otherwise the cost of the KDD process will not be justified.

THE KDD STEPS

In the literature, there have been some variations of the different steps involved in the KDD process. Although these variations do not differ in their entirety, some of them are more descriptive than others. Before we present the proposed steps, it