Knowledge Discovery in Databases and Data Mining

D

Petr Berka
University of Economics, Prague, Czech Republic & University of Finance and Administration, Prague, Czech Republic

INTRODUCTION

Knowledge discovery in databases (KDD) or data mining (DM) is aimed at acquiring implicit knowledge from data and using it to build classification, prediction, description, etc. models for decision support. As more data is gathered, with the amount of data doubling every three years, data mining becomes an increasingly important tool to transform this data into knowledge. While it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set. This article covers the following topics:

- Basic definitions of knowledge discovery in databases and data mining
- Tasks and application areas
- The process of knowledge discovery in databases
- Standardization effort in the area of data mining
- Data Mining tools
- Text mining and web mining as specific subfields of data mining
- Important research challenges

BACKGROUND

The rapid growth of data collected and stored in various application areas brings new problems and challenges in their processing and interpretation. While database technology provides tools for data storage and "simple" querying, and statistics offers methods for analyzing small sample data, new approaches are necessary to face these challenges. These approaches are usually called knowledge discovery in databases or data mining. These terms are often used interchangeably. We will support the view that knowledge discovery in databases is a broader concept covering the whole process in which data mining (also called modeling or analysis) is just one step applying machine learning or statistical algorithms to preprocessed data and building (classification or prediction) models or finding interesting patterns. Thus, we will understand knowledge discovery in databases as the

Non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data (Fayyad et al., 1996, p. 6),

or as an

Analysis of observational data sets to find unsuspected relationships and summarize data in novel ways that are both understandable and useful to the data owner (Hand et al., 2001, p. 1).

Similarly, *data mining refers to extracting knowledge from large amounts of data* (Han et al., 2011, p. 5).

DATA MINING TASKS AND APPLICATION AREAS

Knowledge discovery in databases is commonly used to perform the tasks of data description and summarization, segmentation, concept description, classification, prediction, dependency analysis, or deviation detection (Fayyad et al., 1996; Chapman et al., 2000).

Data Description and Summarization

The goal is a concise description of the data characteristics, typically in elementary and aggregated form. This gives the user an overview of the data structure.

DOI: 10.4018/978-1-4666-5888-2.ch174

Even a very simple and preliminary analysis of this kind is appreciated by data owners and users.

Segmentation

Segmentation (or clustering) aims at separation of the data into interesting and meaningful subgroups or classes where all members of a subgroup share common characteristics. Client profiling and clustering of gene expression data are two examples of this type of task.

Client profiling can be based on the purchase history or service usage history of customers or clients; similar behavior patterns can be used to divide clients into groups and to create profiles of these groups.

Clustering of gene expression data (data in the form of so-called DNA microarrays that are obtained by measuring mRNA levels in cells) can help us identify groups of genes with related expression patterns. Genes with a "close" expression pattern will tend to participate in a similar biological function. We thus can use these patterns, e.g., to group together normal cells belonging to various tissue types.

Classification

Classification assumes that there is a set of objects which belong to different classes. The objective is to build classification models, which assign correct class labels to previously unseen and unlabeled objects. Some examples of this type of task might be credit risk assessment and credit scoring, churn (retention) analysis, customer modeling (to evaluate the propensity to buy a product), or medical and technical diagnostics.

Credit risk assessment, *credit scoring* and loan application approvals are typical data mining tasks. The data for this type of application usually consists of socio-demographic characteristics (e.g., age, gender, region, job), economic characteristics (e.g., income, savings and investments on deposit), the characteristics of the loan (e.g., purpose, amount, monthly payments) and, of course, the loan approval decision. Loan application evaluation is a classification task, in which the final decision can be either a " crisp " yes/no decision about the loan or a numeric score expressing the financial standing of the applicant.

Churn (retention) analysis is performed in areas where a new customer can be acquired only by enticing him from the competitors. A typical application area is thus telecommunication services. Here service usage data can be used to predict which customers are liable to transfer to another provider. Offers of new services can then be customized to retain as many customers as possible.

In *customer modeling* (targeted marketing) the problem is to evaluate the response of a company's customers to a promotion campaign to find the most likely prospects to contact. This can be turned into a classification task to indicate a higher or lower likelihood of a given binary (yes or no) outcome that expresses the propensity to buy the offered product. *Medical diagnosis* of various diseases (leukemia, cancer) is another example of classification tasks. Besides classical data about the patients, gene expression data or microarray data have gained increasing popularity for this task in recent years.

Prediction

Prediction is very similar to classification. The only difference is that, within prediction, the target attribute (class) is not a qualitative discrete attribute but a continuous one. A prediction model is created using data from the past. The outcome of such prediction can be either a categorical (increase, decrease or stability of the next value of the target attribute), or numeric (the next value of the target attribute itself) value of the target attribute. Some examples might be exchange rate prediction, prediction of energy consumption or sales forecasting.

Concept Description

Concept description aims at an understandable description of concepts or classes. The purpose is not to develop complete models with a high prediction accuracy, but to gain insights. Examples of this type of task include description of loyal customers, bad loan applications and insurance claims frauds. Concept description is closely related to classification; if the classification model is interpretable by humans, it can be considered as concept description.

Dependency Analysis

Dependency analysis consists of finding a model that describes significant dependencies (or associations)