DURSUN DELEN

Spears and Patterson Endowed Chairs in Business Analytics
Professor of Management Science and Information Systems
Spears School of Business, Oklahoma State University

REAL-WORLD DATA MINING

APPLIED BUSINESS ANALYTICS AND DECISION MAKING

Real-World Data Mining

female shoppers who purchase seasonal clothes based on their demographics, credit card transactions, and socioeconomic attributes. Furthermore, the analyst should gain an intimate understanding of the data sources—for example, where the relevant data is stored and in what form, whether data collection is automated or happens manually, who collects the data, and how often the data are updated. The analyst should also understand variables by asking questions such as "What are the most relevant variables?" "Are there any synonymous and/or homonymous variables?" and "Are the variables independent of each other—that is, do they stand as a complete information source without overlapping or conflicting information?"

In order to better understand the data, the analyst often uses a variety of statistical and graphical techniques, such as simple statistical descriptors/summaries of each variable (e.g., for numeric variables, the average, minimum/maximum, median, and standard deviation are among the calculated measures, whereas for categorical variables the mode and frequency tables are calculated), correlation analysis, scatterplots, histograms, and box plots. Careful identification and selection of data sources and the most relevant variables can make it easier for data mining algorithms to quickly discover useful knowledge patterns.

Data sources for data selection can vary. Normally, data sources for business applications include demographic data (e.g., income, education, number of households, age), sociographic data (e.g., hobbies, club memberships, entertainment), and transactional data (e.g., sales record, credit card spending, issued checks), among others.

Data can be categorized as quantitative and qualitative. Quantitative data is measured using numeric values. It can be discrete (e.g., integers) or continuous (e.g., real numbers). Qualitative data, also known as categorical data, contains both nominal and ordinal data. Nominal data has finite non-ordered values. For example, gender data has two values: male and female. Ordinal data has finite ordered values. For example, customer credit ratings are considered ordinal data because the ratings can be excellent, fair, and bad. Quantitative data can be readily represented by some sort of probability distribution. A probability distribution describes how the data is dispersed and shaped. For instance, normally distributed data is symmetric and is commonly referred to as being a bell-shaped curve. Qualitative data may be coded to numbers and then described by frequency distributions. Once the relevant data is selected according to the data mining business objective, data preprocessing should be conducted. (For more details on data in data mining, see Chapter 4, "Data and Methods in Data Mining.")

Step 3: Data Preparation

The purpose of data preparation (commonly called data preprocessing) is to prepare the data identified in the previous step for analysis using data mining methods. Compared to the other steps in CRISP-DM, data preprocessing consumes the most time and effort—roughly 80% of the total time spent on a data mining project. This step requires such enormous effort because real-world data is generally incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data), noisy (containing errors or outliers), and inconsistent (containing discrepancies in codes or names).

Some parts of the data may have different formats because they are taken from different data sources. The selected data may be from flat files, voice message, images, and Web pages, and it needs to be converted to a consistent and unified format. In general, data cleaning means filtering, aggregating, and filling in missing values (a.k.a. imputation). By filtering the data, an analyst examines the selected variables for outliers and redundancies. Outliers differ greatly from the majority of data, or data that is clearly out of range of the selected data groups. For example, if the age of a customer included in the data is 190, this must be a data entry error and should be identified and fixed (perhaps taken out of a data mining project that examines the various aspects of customers, since age is perceived to be a critical customer characteristic). Outliers may occur for many reasons, such as human errors or technical errors, or may naturally occur in a data set due to extreme events. Suppose the age of a credit card holder is recorded as “12.” This is likely a data entry error, most likely made by a human. However, there might actually be an independently wealthy preteen with important purchasing habits. Arbitrarily deleting this outlier could dismiss valuable information.

Data may also be redundant, with the same information recorded in several different ways. Daily sales of a particular product are redundant to seasonal sales of the same product because an analyst can derive the sales from either daily data or seasonal data. Aggregating data reduces data dimensions. Note that although an aggregated data set has a small volume, the information remains. If a marketing promotion for furniture sales is considered in the next three or four years, the available daily sales data can be aggregated as annual sales data. The size of the sales data is then dramatically reduced. By smoothing data, missing values of the selected data are found and new or reasonable values are then added. These added values could be the average number of the variable (mean) or the mode. A missing value often causes no solution when a data mining algorithm is applied to discover the knowledge patterns.

Step 4: Model Building

In the fourth step of the CRISP-DM process, various modeling techniques are selected and applied to an already prepared data set in order to address the specific business need. The modeling step also encompasses the assessment and comparative analysis of the various types of models that can address the same type of data mining tasks (e.g., clustering, classification). Because there is not a universally known best method or algorithm for a specific data mining task, an analyst should use a variety of viable model types, along with a well-defined experimentation and assessment strategy, to identify the "best" method for a given data mining problem. Even for a single method or algorithm, a number of parameters need to be calibrated to obtain optimal results. Some methods may have specific requirements for the way the data is to be formatted; thus returning to the data preparation step is often necessary.

Depending on the business need, the data mining task can be a prediction (either classification or regression), an association, or a clustering/segmentation type. Each of these data mining tasks can use a variety of data mining methods and algorithms. For instance, classification-type data mining tasks can be accomplished by developing neural networks or by using decision trees, support vector machines, or logistic regression. These data mining methods and their respective algorithms are explained in Chapter 4 and Chapter 5, "Data Mining Algorithms."

The standard procedure for modeling in data mining is to divide a large preprocessed data set into subsets for training and validation or testing. Then the analyst can use a portion of the data (the training set) to develop the models (no matter what modeling technique and/or algorithm is used) and use the other portion of the data (the test set) for testing the model that was just built. The principle is that if you build a model on a particular set of data, it will of course test quite well on the data on which it was built. By dividing the data and using part of it for model development and testing it on a separate set of data, an analyst can create convincing and reliable results for the accuracy and reliability of the model. The idea of splitting the data into components is often carried to additional levels, with multiple splits in the practice of data mining. For further details about data splitting and other evaluation methods, see Chapter 4.

Step 5: Testing and Evaluation

In the fifth step of the CRISP-DM process, the developed models are assessed and evaluated for accuracy and generality. This step assesses the degree to which the selected model (or models) meets the business objectives and whether more models need to be developed and assessed. Another option is to test the developed model(s) in a real-world scenario if time and budget constraints permit. Even though the outcome of the developed models is expected to relate to the original business objectives, other findings that are not necessarily related to the original business objectives but that might also unveil additional information or hints for future directions are often discovered.

Testing and evaluation is a critical and challenging step. No value is added by the data mining task until the business value obtained from discovered knowledge patterns is identified and recognized. Determining the business value from discovered knowledge patterns is somewhat similar to playing with puzzles. The extracted knowledge patterns are pieces of the puzzle that need to be put together in the context of the specific business purpose. The success of this identification operation depends on the interaction among data analysts, business analysts, and decision makers (e.g., business managers). Data analysts may fully understand the data mining objectives and what they mean to the business, and business analysts and decision makers may not have the technical knowledge to interpret the results of sophisticated mathematical solutions; therefore, interaction among them is necessary. In order to properly interpret knowledge patterns, it is often necessary to use a variety of tabulation and visualization techniques (e.g., pivot tables, cross-tabulation of findings, pie charts, histograms, box plots, scatterplots).

Step 6: Deployment

Development and assessment of models is not the end of a data mining project. Even if the purpose of a model is to do a simple exploration of the data, the knowledge gained from such exploration needs to be organized and presented in a way that the end user can understand and benefit from. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if an analyst will not carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

The deployment step of the CRISP-DM process may also include maintenance activities for the deployed models. Because the business is constantly changing, the data that reflects the business activities is also changing. Over time, the models (and the patterns embedded within them) built on the old data may become obsolete, irrelevant, or misleading. Therefore, monitoring and maintaining the models are important if the data mining results are to become a part of the day-to-day business and its environment. Careful preparation of a maintenance strategy helps avoid unnecessarily long periods of incorrect use of data mining results. In order to monitor the deployment of the data mining result(s), a project needs a detailed plan for the monitoring process, which may not be trivial with complex data mining models.

The CRISP-DM process is the most complete and most popular data mining methodology practice in industry as well as in academia. Rather than use it as is, practitioners add their own insights to make it specific to their style of practice.