5

Data mining

As observed in previous chapters, the evolving technologies of information gathering and storage have made available huge amounts of data within most application domains, such as the business world, the scientific and medical community, and public administration. The set of activities involved in the analysis of these large databases, usually with the purpose of extracting useful knowledge to support decision making, has been referred to in different ways, such as *data mining*, *knowledge discovery*, *pattern recognition* and *machine learning*.

In particular, the term *data mining* indicates the process of exploration and analysis of a dataset, usually of large size, in order to find regular patterns, to extract relevant knowledge and to obtain meaningful recurring rules. Data mining plays an ever-growing role in both theoretical studies and applications.

In this chapter we wish to describe and characterize data mining activities with respect to investigation purposes and analysis methodologies. The relevant properties of input data will also be discussed. Finally, we will describe the data mining process and its articulation in distinct phases.

5.1 Definition of data mining

Data mining activities constitute an iterative process aimed at the analysis of large databases, with the purpose of extracting information and knowledge that may prove accurate and potentially useful for knowledge workers engaged in decision making and problem solving.

As described in Section 5.3, the analysis process is iterative in nature since there are distinct phases that might imply feedback and subsequent revisions. Usually such a process represents a cooperative activity between experts in the application domain and data analysts, who use mathematical models

Business Intelligence: *Data Mining and Optimization for Decision Making* *Carlo Vercellis*
 ISBN: 978-0-470-51138-1

for inductive learning. Indeed, experience indicates that a data mining study requires frequent interventions by the analysts across the different investigation phases and therefore cannot easily be automated. It is also necessary that the knowledge extracted be accurate, in the sense that it must be confirmed by data and not lead to misleading conclusions.

The term *data mining* refers therefore to the overall process consisting of data gathering and analysis, development of inductive learning models and adoption of practical decisions and consequent actions based on the knowledge acquired. The term *mathematical learning theory* is reserved for the variety of mathematical models and methods that can be found at the core of each data mining analysis and that are used to generate new knowledge.

The data mining process is based on inductive learning methods, whose main purpose is to derive general rules starting from a set of available examples, consisting of past observations recorded in one or more databases. In other words, the purpose of a data mining analysis is to draw some conclusions starting from a sample of past observations and to generalize these conclusions with reference to the entire population, in such a way that they are as accurate as possible. The models and patterns identified in this way may take on different forms, which will be described in the following chapters, such as linear equations, sets of rules in *if–then–else* form, clusters, charts and trees.

A further characteristic of data mining depends on the procedure for collecting past observations and inserting them into a database. Indeed, these records are usually stored for purposes that are not primarily driven by data mining analysis. For instance, information on purchases from a retail company, or on the usage of each telephone number stored by a mobile phone provider, will basically be recorded for administrative purposes, even if the data may be later used to perform some useful data mining analysis. The data gathering procedure is therefore largely independent and unaware of the data mining objectives, so that it substantially differs from data gathering activities carried out according to predetermined sampling schemes, typical of classical statistics. In this respect, data mining represent a *secondary* form of data analysis.

Data mining activities can be subdivided into two major investigation streams, according to the main purpose of the analysis: *interpretation* and *prediction*.

Interpretation. The purpose of interpretation is to identify regular patterns in the data and to express them through rules and criteria that can be easily understood by experts in the application domain. The rules generated must be original and non-trivial in order to actually increase the level of knowledge and understanding of the system of interest. For example, for a company in the retail industry it might be advantageous to cluster those customers who have taken out loyalty cards according to their purchasing profile. The segments

generated in this way might prove useful in identifying new market niches and directing future marketing campaigns.

Prediction. The purpose of prediction is to anticipate the value that a random variable will assume in the future or to estimate the likelihood of future events. For example, a mobile phone provider may develop a data mining analysis to estimate for its customers the probability of churning in favor of some competitor. In a different context, a retail company might predict the sales of a given product during the subsequent weeks. Actually, most data mining techniques derive their predictions from the value of a set of variables associated with the entities in a database. For example, a data mining model may indicate that the likelihood of future churning for a customer depends on features such as age, duration of the contract and percentage of calls to subscribers of other phone providers. There are, however, time series models, described in Chapter 9, which make predictions based only on the past values of the variable of interest.

Sometimes, a model developed for the purpose of prediction may also turn out to be effective for interpretation. In the case of classification trees with parallel axis splitting rules, which will be described in Chapter 10, the models generated for predictive purposes may also prove useful in identifying recurrent explanatory phenomena.

5.1.1 Models and methods for data mining

There are several learning methods that are available to perform the different data mining tasks. A number of techniques originated in the field of computer science, such as classification trees or association rules, and are referred to as *machine learning* or *knowledge discovery in databases*. In most cases an empirically based approach tends to prevail within this class of techniques. Other methods belong to multivariate statistics, such as regression or Bayesian classifiers, and are often parametric in nature but appear more theoretically grounded. More recent developments include mathematical methods for learning, such as *statistical learning theory*, which are based on solid theoretical foundations and place themselves at the crossroads of various disciplines, among which probability theory, optimization theory and statistics.

Example 5.1 – Linear regression. Linear regression models, described in Chapter 8, are one of the best-known learning and predictive methodologies in classical statistics. In its simplest form, linear regression is used to relate a dependent response variable Y to an independent predictor X

through a linear regression in the form $Y = aX + b$, where a and b are parameters to be determined using the available past observations. For example, Y may represent the sales of a mass consumption product during a week and X the total advertisement cost during the same week. With respect to the development phases of a model, the selection of a linear function determines the type of relationship between the predictor and the response variable. A reasonable evaluation metric is the sum of the squared differences between the values of Y actually observed in the past and the values predicted by the linear model. An appropriate optimization algorithm calculates the value of the parameters a and b in order to minimize the sum of squared errors.

Irrespective of the specific learning method that one wishes to adopt, there are other recurrent steps in the development of a data mining model, as shown in Example 5.1:

- the selection of a class of models to be used for learning from the past and of a specific form for representing patterns in the data;
- the definition of a metric for evaluating the effectiveness and accuracy of the models being generated;
- the design of a computational algorithm in order to generate the models by optimizing the evaluation metric.

In the next chapters we will provide a description of the most popular classes of methods for prediction and pattern recognition. Some methodologies can be applied to perform several data mining tasks, among those described in Section 5.4. Since it is generally possible to relate each class of models to a prevailing activity, in the following chapters each technique will be associated with the most appropriate task.

5.1.2 Data mining, classical statistics and OLAP

Data mining projects differ in many respects from both classical statistics and OLAP analyses. Such differences are shown in Table 5.1, with reference to an example.

The main difference consists of the active orientation offered by inductive learning models, compared with the passive nature of statistical techniques and OLAP. Indeed, in statistical analyses decision makers formulate a hypothesis that then has to be confirmed on the basis of sample evidence. Similarly, in OLAP analyses knowledge workers express some intuition on which they base

Table 5.1 Differences between OLAP, statistics and data mining

OLAP	statistics	data mining
extraction of details and aggregate totals from data	verification of hypotheses formulated by analysts	identification of patterns and recurrences in data
information	validation	knowledge
distribution of incomes of home loan applicants	analysis of variance of incomes of home loan applicants	characterization of home loan applicants and prediction of future applicants

extraction, reporting and visualization criteria. Both methods – on one hand statistical validation techniques and on the other hand information tools to navigate through data cubes – only provide elements to confirm or disprove the hypotheses formulated by the decision maker, according to a *top-down* analysis flow. Conversely, learning models, which represent the core of data mining projects, are capable of playing an active role by generating predictions and interpretations which actually represent new knowledge available to the users. The analysis flow in the latter case has a *bottom-up* structure. In particular, when faced with large amounts of data, the use of models capable of playing an active role becomes a critical success factor, since it is hard for knowledge workers to formulate a priori meaningful and well-founded hypotheses.

5.1.3 Applications of data mining

Data mining methodologies can be applied to a variety of domains, from marketing and manufacturing process control to the study of risk factors in medical diagnosis, from the evaluation of the effectiveness of new drugs to fraud detection.

Relational marketing. Data mining applications in the field of relational marketing, described in Chapter 13, have significantly contributed to the increase in the popularity of these methodologies. Some relevant applications within relational marketing are:

- identification of customer segments that are most likely to respond to targeted marketing campaigns, such as *cross-selling* and *up-selling*;
- identification of target customer segments for retention campaigns;
- prediction of the rate of positive responses to marketing campaigns;
- interpretation and understanding of the buying behavior of the customers;

- analysis of the products jointly purchased by customers, known as *market basket analysis*.

Fraud detection. Fraud detection is another relevant field of application of data mining. Fraud may affect different industries such as telephony, insurance (false claims) and banking (illegal use of credit cards and bank checks; illegal monetary transactions).

Risk evaluation. The purpose of risk analysis is to estimate the risk connected with future decisions, which often assume a dichotomous form. For example, using the past observations available, a bank may develop a predictive model to establish if it is appropriate to grant a monetary loan or a home loan, based on the characteristics of the applicant.

Text mining. Data mining can be applied to different kinds of texts, which represent unstructured data, in order to classify articles, books, documents, emails and web pages. Examples are web search engines or the automatic classification of press releases for storing purposes. Other text mining applications include the generation of filters for email messages and newsgroups.

Image recognition. The treatment and classification of digital images, both static and dynamic, is an exciting subject both for its theoretical interest and the great number of applications it offers. It is useful to recognize written characters, compare and identify human faces, apply correction filters to photographic equipment and detect suspicious behaviors through surveillance video cameras.

Web mining. Web mining applications, which will be briefly considered in section 13.1.9, are intended for the analysis of so-called *clickstreams* – the sequences of pages visited and the choices made by a web surfer. They may prove useful for the analysis of e-commerce sites, in offering flexible and customized pages to surfers, in caching the most popular pages or in evaluating the effectiveness of an e-learning training course.

Medical diagnosis. Learning models are an invaluable tool within the medical field for the early detection of diseases using clinical test results. Image analysis for diagnostic purpose is another field of investigation that is currently burgeoning.

5.2 Representation of input data

In most cases, the input to a data mining analysis takes the form of a two-dimensional table, called a *dataset*, irrespective of the actual logic and material representation adopted to store the information in files, databases, data

warehouses and data marts used as data sources. The rows in the dataset correspond to the *observations* recorded in the past and are also called *examples*, *cases*, *instances* or *records*. The columns represent the information available for each observation and are termed *attributes*, *variables*, *characteristics* or *features*.

The attributes contained in a dataset can be categorized as *categorical* or *numerical*, depending on the type of values they take on.

Categorical. Categorical attributes assume a finite number of distinct values, in most cases limited to less than a hundred, representing a qualitative property of an entity to which they refer. Examples of categorical attributes are the province of residence of an individual (which takes as values a series of names, which in turn may be represented by integers) or whether a customer has abandoned her service provider (expressed by the value 1) or remained loyal to it (expressed by the value 0). Arithmetic operations cannot be applied to categorical attributes even when the coding of their values is expressed by integer numbers.

Numerical. Numerical attributes assume a finite or infinite number of values and lend themselves to subtraction or division operations. For example, the amount of outgoing phone calls during a month for a generic customer represents a numerical variable. Regarding two customers A and B making phone calls in a week for €27 and €36 respectively, it makes sense to claim that the difference between the amounts spent by the two customers is equal to €9 and that A has spent three fourths of the amount spent by B.

Sometimes a more refined taxonomy of attributes can prove useful.

Counts. Counts are categorical attributes in relation to which a specific property can be true or false. These attributes can therefore be represented using Boolean variables {true, false} or binary variables {0,1}. For example, a bank's customers may or may not be holders of a credit card issued by the bank.

Nominal. Nominal attributes are categorical attributes without a natural ordering, such as the province of residence.

Ordinal. Ordinal attributes, such as education level, are categorical attributes that lend themselves to a natural ordering but for which it makes no sense to calculate differences or ratios between the values.

Discrete. Discrete attributes are numerical attributes that assume a finite number or a countable infinity of values.[1]

[1] If a set A has the same cardinality as the set $\mathbb{N}$ of natural numbers, then we say that A is *countable*. In other words, a set is countable if there is a bijection between that set and $\mathbb{N}$. There exist sets, such as the set $\mathbb{R}$ of real numbers, that are infinite and not countable, and are therefore called *uncountable*.

Continuous. Continuous attributes are numerical attributes that assume an uncountable infinity of values.

To represent a generic dataset $\mathcal{D}$, we will denote by m the number of observations, or rows, in the two-dimensional table containing the data and by n the number of attributes, or columns. Furthermore, we will denote by

$$\mathbf{X} = [x_{ij}], \quad i \in \mathcal{M} = \{1, 2, \ldots, m\}, \quad j \in \mathcal{N} = \{1, 2, \ldots, n\}, \tag{5.1}$$

the matrix of dimensions $m \times n$ that corresponds to the entries in the dataset $\mathcal{D}$. We will write

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \tag{5.2}$$

$$\mathbf{a}_j = (x_{1j}, x_{2j}, \ldots, x_{mj}) \tag{5.3}$$

for the n-dimensional row vector associated with the ith record of the dataset and the m-dimensional column vector representing the jth attribute in $\mathcal{D}$, respectively.

5.3 Data mining process

The definition of data mining given at the beginning of Section 5.1 refers to an iterative process, during which learning models and techniques play a key, though non-exhaustive, role. Figure 5.1 shows the main phases of a generic data mining process.

Definition of objectives. Data mining analyses are carried out in specific application domains and are intended to provide decision makers with useful knowledge. As a consequence, intuition and competence are required by the domain experts in order to formulate plausible and well-defined investigation objectives. If the problem at hand is not adequately identified and circumscribed one may run the risk of thwarting any future effort made during data mining activities. The definition of the goals will benefit from close cooperation between experts in the field of application and data mining analysts. With reference to Example 5.2, it is possible to define the problem and the goals of the investigation as the analysis of past data and identification of a model so as to express the propensity of customers to leave the service (churn) based on their characteristics, in order to understand the reasons for such disloyalty and predict the probability of churn.

In the hierarchy of infinities, countable sets, such as the set $\mathbb{Q}$ of rational numbers, are placed on the lowest step and are consequently less 'dense' than others.

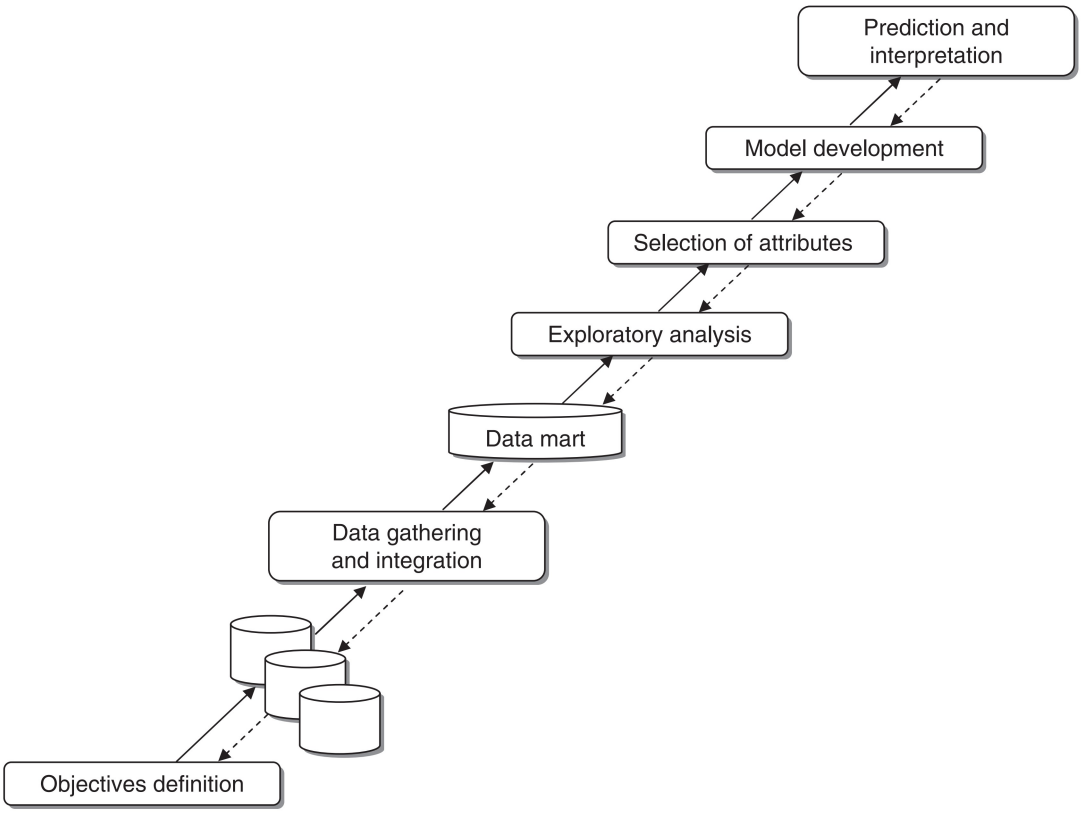

Figure 5.1 Data mining process

Example 5.2 – Retention in the mobile phone industry. Table 5.2 shows the two-dimensional structure of input data from an example of the analysis of customer loyalty. Suppose that a mobile phone company carries out a data mining analysis with both prediction and interpretation goals. On the one hand, the company wishes to assess the likelihood of future churning by each customer, in order to target marketing actions for retention purposes. On the other hand, the intent is to understand the reasons why customers churn, with the purpose of improving the service level and reducing future churning. Table 5.2 contains 23 observations and 12 attributes, whose meaning is indicated in Table 5.3.

The first 11 attributes represent *explanatory* variables, while the last attribute represents the *target* variable, expressing the class of each record in relation to the objectives of the data mining analysis. The first explanatory variable gives personal demographic information while the rest refer to the use of the service. Observed values are relative to time period of index $t - 2$ for the explanatory attributes, whereas for the target variable they refer to period t. The difference in time placement is required in

Table 5.2 An example of input data for a data mining model

area	numin	timein	numout	Pothers	Pmob	Pland	numsms	numserv	numcall	diropt	churner
3	32	8093	45	0.14	0.75	0.12	18	1	0	0	0
3	277	157842	450	0.26	0.35	0.38	9	3	0	1	0
1	17	15023	20	0.37	0.23	0.40	1	1	0	0	0
1	46	22459	69	0.10	0.39	0.51	33	1	0	0	0
1	19	8640	9	0.00	0.00	1.00	0	0	0	0	0
2	17	7652	66	0.16	0.42	0.43	1	3	0	1	0
3	47	17768	11	0.45	0.00	0.55	0	0	0	0	0
3	19	9492	42	0.18	0.34	0.48	3	1	0	0	1
1	1	84	9	0.09	0.54	0.37	0	0	0	0	1
2	119	87605	126	0.84	0.02	0.14	12	1	0	0	0
4	24	6902	47	0.25	0.26	0.48	4	1	0	0	0
1	32	28072	43	0.28	0.66	0.06	0	1	0	0	0
3	103	112120	24	0.61	0.28	0.11	24	2	0	0	0
3	45	21921	94	0.34	0.47	0.19	45	2	0	1	0
1	8	25117	89	0.02	0.89	0.09	189	1	3	0	0
3	4	945	16	0.00	0.00	1.00	0	0	0	0	1
2	83	44263	83	0.00	0.00	0.67	0	0	0	0	1
2	22	15979	59	0.05	0.53	0.41	5	2	0	1	1
2	0	0	57	0.00	1.00	0.00	15	1	1	0	1
4	162	114108	273	0.18	0.15	0.41	2	3	0	1	1
4	21	4141	70	0.14	0.58	0.28	0	1	0	1	1
4	33	10066	45	0.12	0.21	0.67	0	0	0	0	1
4	5	965	40	0.41	0.27	0.32	64	1	0	0	1

Table 5.3 Meaning of the attributes in Table 5.2

attribute	meaning
area	residence area
numin	number of calls received in period $t-2$
timein	duration in seconds of calls received in period $t-2$
numout	number of calls placed in the period $t-2$
Pothers	percentage of calls placed to other mobile telephone companies in period $t-2$
Pmob	percentage of calls placed to the same mobile telephone company in period $t-2$
Pland	percentage of calls placed to land numbers in period $t-2$
numsms	number of messages sent in period $t-2$
numserv	number of calls placed to special services in period $t-2$
numcall	number of calls placed to the call center in period $t-2$
diropt	binary variable indicating whether the customer corresponding to the record has subscribed to a special rate plan for calls placed to selected numbers
churner	binary variable indicating whether the customer corresponding to the record has left the service in period t

order to use the model for predictive purposes. Indeed, it is necessary to predict during the current period which customers will leave the service within 2 periods, based on the available information, in order to develop timely and effective retention actions; see also the discussion on time latency for predictive models in Section 13.1.4. Of course, in a real application the number of available attributes is much higher, of the order of hundreds or thousands, and the number of rows representing the customers is far greater, of the order of hundreds of thousands or millions of records. The purpose here is to obtain an inductive model that is capable of learning from past available observations and identifying a plausible relationship between the target variable and the explanatory attributes. Once the model has been created based on past records, it is possible to use it to predict the target class of new records or to understand common characteristics of customers who churn compared to those who remain loyal.

Data gathering and integration. Once the objectives of the investigation have been identified, the gathering of data begins. Data may come from different sources and therefore may require integration. Data sources may be internal, external or a combination of the two. The integration of distinct data sources may be suggested by the need to enrich the data with new descriptive

dimensions, such as geomarketing variables, or with lists of names of potential customers, termed *prospects*, not yet existing in the company information system. In some instances, data sources are already structured in data warehouses and data marts for OLAP analyses and more generally for decision support activities. These are favorable situations where it is sufficient to select the attributes deemed relevant for the purpose of a data mining analysis. There is a risk, however, that, in order to limit memory uptake, the information stored in a data warehouse has been aggregated and consolidated to such an extent as to render useless any subsequent analysis. For example, if a company in the retail industry stores for each customer the total amount of every receipt, without keeping track of each individual purchased item, a future data mining analysis aimed at investigating the actual purchasing behavior may be compromised. In other situations, the original data have a heterogeneous format with no predefined structure. In this case, the process of data gathering and integration becomes more arduous and therefore more prone to errors. Regardless of the original structure, input datasets of data mining analyses almost always take the form of two-dimensional tables, as observed above. Unlike many standard sampling procedures of classical statistics, datasets for data mining represent samples extracted in accordance with an unknown distribution, with the analysts not being able to influence and affect the data gathering process. Chapter 6 will discuss data preparation issues in more detail.

Exploratory analysis. In the third phase of the data mining process, a preliminary analysis of the data is carried out with the purpose of getting acquainted with the available information and carrying out *data cleansing*. Usually, the data stored in a data warehouse are processed at loading time in such a way as to remove any *syntactical* inconsistencies. For example, dates of birth that fall outside admissible ranges and negative sales charges are detected and corrected. In the data mining process, data cleansing occurs at a *semantic* level. First of all, the distribution of the values for each attribute is studied, using histograms for categorical attributes and basic summary statistics for numerical variables. In this way, any abnormal values (*outliers*) and missing values are also highlighted. These are studied by experts in the application domain who may consider excluding the corresponding records from the investigation. Chapter 7 will discuss the techniques used to develop exploratory data analysis.

Attribute Selection. In the subsequent phase, the relevance of the different attributes is evaluated in relation to the goals of the analysis. Attributes that prove to be of little use are removed, in order to cleanse irrelevant information from the dataset. Furthermore, new attributes obtained from the original variables through appropriate transformations are included into the dataset. For example, in most cases it is helpful to introduce new attributes that reflect

the trends inherent in the data through the calculation of ratios and differences between original variables. Exploratory analysis and attribute selection are critical and often challenging stages of the data mining process and may influence to a great extent the level of success of the subsequent stages. The methods described in Chapters 6 and 7 may be useful for transforming and selecting the attributes.

Model development and validation. Once a high quality dataset has been assembled and possibly enriched with newly defined attributes, pattern recognition and predictive models can be developed. Usually the *training* of the models is carried out using a sample of records extracted from the original dataset. Then, the predictive accuracy of each model generated can be assessed using the rest of the data. More precisely, the available dataset is split into two subsets. The first constitutes the *training set* and is used to identify a specific learning model within the selected class of models. Usually the sample size of the training set is chosen to be relatively small, although significant from a statistical standpoint – say, a few thousands observations. The second subset is the *test set* and is used to assess the accuracy of the alternative models generated during the training phase, in order to identify the best model for actual future predictions. The most popular classes of learning models will be discussed in detail in the following chapters.

Prediction and interpretation. Upon conclusion of the data mining process, the model selected among those generated during the development phase should be implemented and used to achieve the goals that were originally identified. Moreover, it should be incorporated into the procedures supporting decision-making processes so that knowledge workers may be able to use it to draw predictions and acquire a more in-depth knowledge of the phenomenon of interest.

The data mining process includes feedback cycles, represented by the dotted arrows in Figure 5.1, which may indicate a return to some previous phase, depending on the outcome of the subsequent phases.

Finally, we should emphasize the importance of the involvement and interaction of several professional roles in order to achieve an effective data mining process:

- an expert in the application domain, expected to define the original objectives of the analysis, to provide appropriate understanding during the subsequent data mining activities and to contribute to the selection of the most effective and accurate model;

- an expert in the company information systems, expected to supervise the access to the information sources;

- an expert in the mathematical theory of learning and statistics, for exploratory data analysis and for the generation of predictive models.

Figure 5.2 illustrates the competencies and the involvement in the different activities for each actor in the data mining process.

5.4 Analysis methodologies

Data mining activities can be subdivided into a few major categories, based on the tasks and the objectives of the analysis. Depending on the possible existence of a target variable, one can draw a first fundamental distinction between *supervised* and *unsupervised* learning processes.

Supervised learning. In a supervised (or *direct*) learning analysis, a target attribute either represents the class to which each record belongs, as shown in Example 5.2 on loyalty in the mobile phone industry, or expresses a measurable quantity, such as the total value of calls that will be placed by a customer in a future period. As a second example of the supervised perspective, consider an investment management company wishing to predict the balance sheet of its customers based on their demographic characteristics and past investment transactions. Supervised learning processes are therefore oriented toward prediction and interpretation with respect to a target attribute.

Unsupervised learning. Unsupervised (or *indirect*) learning analyses are not guided by a target attribute. Therefore, data mining tasks in this case are aimed at discovering recurring patterns and affinities in the dataset. As an example, consider an investment management company wishing to identify clusters of customers who exhibit homogeneous investment behavior, based on data on past transactions. In most unsupervised learning analyses, one is interested in identifying *clusters* of records that are similar within each cluster and different from members of other clusters.

Taking the distinction even further, seven basic data mining tasks can be identified:

- characterization and discrimination;
- classification;
- regression;
- time series analysis;
- association rules;