Applied Data Mining for Business and Industry

Applied Data Mining for Business and Industry, Second Edition Paolo Giudici and Silvia Figini
 ISBN: 978-0-470-05886-2

Applied Data Mining for Business and Industry

Second Edition

PAOLO GIUDICI
Department of Economics, University of Pavia, Italy

SILVIA FIGINI
Faculty of Economics, University of Pavia, Italy

A John Wiley and Sons, Ltd., Publication

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Giudici, Paolo.
Applied data mining for business and industry / Paolo Giudici, Silvia Figini. – 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-05886-2 (cloth) – ISBN 978-0-470-05887-9 (pbk.)
1. Data mining. 2. Business–Data processing. 3. Commercial statistics. I. Figini, Silvia. II. Title.
QA76.9.D343G75 2009
005.74068—dc22

2009008334

A catalogue record for this book is available from the British Library

ISBN: 978-0-470-05886-2 (Hbk)
ISBN: 978-0-470-05887-9 (Pbk)

Typeset in 10/12 Times-Roman by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall, UK

Contents

CHAPTER 1

Introduction

From an operational point of view, data mining is an integrated process of data analysis that consists of a series of activities that go from the definition of the objectives to be analysed, to the analysis of the data up to the interpretation and evaluation of the results. The various phases of the process are as follows:

Definition of the objectives for analysis. It is not always easy to define statistically the phenomenon we want to analyse. In fact, while the company objectives that we are aiming for are usually clear, they can be difficult to formalise. A clear statement of the problem and the objectives to be achieved is is of the utmost importance in setting up the analysis correctly. This is certainly one of the most difficult parts of the process since it determines the methods to be employed. Therefore the objectives must be clear and there must be no room for doubt or uncertainty.

Selection, organisation and pre-treatment of the data. Once the objectives of the analysis have been identified it is then necessary to collect or select the data needed for the analysis. First of all, it is necessary to identify the data sources. Usually data is taken from internal sources that are cheaper and more reliable. This data also has the advantage of being the result of the experiences and procedures of the company itself. The ideal data source is the company data warehouse, a 'store room' of historical data that is no longer subject to changes and from which it is easy to extract topic databases (data marts) of interest. If there is no data warehouse then the data marts must be created by overlapping the different sources of company data.

In general, the creation of data marts to be analysed provides the fundamental input for the subsequent data analysis. It leads to a representation of the data, usually in table form, known as a data matrix that is based on the analytical needs and the previously established aims.

Once a data matrix is available it is often necessary to carry out a process of preliminary cleaning of the data. In other words, a quality control exercise is carried out on the data available. This is a formal process used to find or select variables that cannot be used, that is, variables that exist but are not suitable for analysis. It is also an important check on the contents of the variables and

Applied Data Mining for Business and Industry, Second Edition Paolo Giudici and Silvia Figini
 ISBN: 978-0-470-05886-2

the possible presence of missing or incorrect data. If any essential information is missing it will then be necessary to supply further data. (See Agresti (1990).

Exploratory analysis of the data and their transformation. This phase involves a preliminary exploratory analysis of the data, very similar to on-line analytical process (OLAP) techniques. It involves an initial evaluation of the importance of the collected data. This phase might lead to a transformation of the original variables in order to better understand the phenomenon or which statistical methods to use. An exploratory analysis can highlight any anomalous data, data that is different from the rest. This data will not necessarily be eliminated because it might contain information that is important in achieving the objectives of the analysis. We think that an exploratory analysis of the data is essential because it allows the analyst to select the most appropriate statistical methods for the next phase of the analysis. This choice must consider the quality of the available data. The exploratory analysis might also suggest the need for new data extraction, if the collected data is considered insufficient for the aims of the analysis.

Specification of statistical methods. There are various statistical methods that can be used, and thus many algorithms available, so it is important to have a classification of the existing methods. The choice of which method to use in the analysis depends on the problem being studied or on the type of data available. The data mining process is guided by the application. For this reason, the classification of the statistical methods depends on the analysis's aim. Therefore, we group the methods into two main classes corresponding to distinct/different phases of the data analysis.

- **Descriptive methods.** The main objective of this class of methods (also called symmetrical, unsupervised or indirect) is to describe groups of data in a succinct way. This can concern both the observations, which are classified into groups not known beforehand (cluster analysis, Kohonen maps) as well as the variables that are connected among themselves according to links unknown beforehand (association methods, log-linear models, graphical models). In descriptive methods there are no hypotheses of causality among the available variables.
- **Predictive methods.** In this class of methods (also called asymmetrical, supervised or direct) the aim is to describe one or more of the variables in relation to all the others. This is done by looking for rules of classification or prediction based on the data. These rules help predict or classify the future result of one or more response or target variables in relation to what happens to the explanatory or input variables. The main methods of this type are those developed in the field of machine learning such as neural networks (multilayer perceptrons) and decision trees, but also classic statistical models such as linear and logistic regression models.

Analysis of the data based on the chosen methods. Once the statistical methods have been specified they must be translated into appropriate algorithms for computing the results we need from the available data. Given the wide range of specialised and non-specialised software available for data mining, it is not necessary to develop ad hoc calculation algorithms for the most 'standard'

applications. However, it is important that those managing the data mining process have a good understanding of the different available methods as well as of the different software solutions, so that they can adapt the process to the specific needs of the company and can correctly interpret the results of the analysis.

Evaluation and comparison of the methods used and choice of the final model for analysis. To produce a final decision it is necessary to choose the best 'model' from the various statistical methods available. The choice of model is based on the comparison of the results obtained. It may be that none of the methods used satisfactorily achieves the analysis aims. In this case it is necessary to specify a more appropriate method for the analysis. When evaluating the performance of a specific method, as well as diagnostic measures of a statistical type, other things must be considered such as the constraints on the business both in terms of time and resources, as well as the quality and the availability of data. In data mining it is not usually a good idea to use just one statistical method to analyse data. Each method has the potential to highlight aspects that may be ignored by other methods.

Interpretation of the chosen model and its use in the decision process. Data mining is not only data analysis, but also the integration of the results into the company decision process. Business knowledge, the extraction of rules and their use in the decision process allow us to move from the analytical phase to the production of a decision engine. Once the model has been chosen and tested with a data set, the classification rule can be generalised. For example, we will be able to distinguish which customers will be more profitable or to calibrate differentiated commercial policies for different target consumer groups, thereby increasing the profits of the company.

Having seen the benefits we can get from data mining, it is crucial to implement the process correctly in order to exploit it to its full potential. The inclusion of the data mining process in the company organisation must be done gradually, setting out realistic aims and looking at the results along the way. The final aim is for data mining to be fully integrated with the other activities that are used to support company decisions. This process of integration can be divided into four phases:

- **Strategic phase.** In this first phase we study the business procedures in order to identify where data mining could be more beneficial. The results at the end of this phase are the definition of the business objectives for a pilot data mining project and the definition of criteria to evaluate the project itself.
- **Training phase.** This phase allows us to evaluate the data mining activity more carefully. A pilot project is set up and the results are assessed using the objectives and the criteria established in the previous phase. A fundamental aspect of the implementation of a data mining procedure is the choice of the pilot project. It must be easy to use but also important enough to create interest.
- **Creation phase.** If the positive evaluation of the pilot project results in implementing a complete data mining system it will then be necessary to

establish a detailed plan to reorganise the business procedure in order to include the data mining activity. More specifically, it will be necessary to reorganise the business database with the possible creation of a data warehouse; to develop the previous data mining prototype until we have an initial operational version and to allocate personnel and time to follow the project.

- **Migration phase.** At this stage all we need to do is to prepare the organisation appropriately so that the data mining process can be successfully integrated. This means teaching likely users the potential of the new system and increasing their trust in the benefits that the system will bring to the company. This means constantly evaluating (and communicating) the results obtained from the data mining process.

PART I

Methodology

Applied Data Mining for Business and Industry, Second Edition Paolo Giudici and Silvia Figini
 ISBN: 978-0-470-05886-2

CHAPTER 2

Organisation of the data

Data analysis requires the data to be organised into an ordered database. We will not discuss how to create a database in this book. The way in which the data is analysed depends on how the data is organised within the database. In our information society there is an abundance of data which calls for an efficient statistical analysis. However, an efficient analysis assumes and requires a valid organisation of the data.

It is of strategic importance for all medium-sized and large companies to have a unified information system, called a data warehouse, that integrates, for example, the accounting data with the data arising from the production process, the contacts with the suppliers (supply chain management), the sales trends and the contacts with the customers (customer relationship management). This system provides precious information for business management. Another example is the increasing diffusion of electronic trade and commerce and, consequently, the abundance of data about web sites visited together with payment transactions. In this case it is essential for the service supplier to understand who the customers are in order to plan offers. This can be done if the transactions (which correspond to clicks on the web) are transferred to an ordered database that can later be analysed.

Furthermore, since the information which can be extracted from a data mining process (data analysis) depends on how the data is organised it is very important that the data analysts are also involved in setting up the database itself. However, frequently the analyst finds himself with a database that has already been prepared. It is then his/her job to understand how it has been set up and how it can be used to achieve the stated objectives. When faced with poorly set-up databases it is a good idea to ask for these to be reviewed rather than trying to laboriously extract information that might ultimately be of little use.

In the remainder of this chapter we will describe how to transform the database so that it can be analysed. A common structure is the so-called data matrix. We will then consider how sometimes it is a good idea to transform a data matrix in terms of binary variables, frequency distributions, or in other ways. Finally, we will consider examples of more complex data structures.

2.1 Statistical units and statistical variables

From a statistical point of view, a database should be organised according to two principles: the statistical units, the elements in the reference population that

Applied Data Mining for Business and Industry, Second Edition Paolo Giudici and Silvia Figini
 ISBN: 978-0-470-05886-2

are considered important for the aims of the analysis (for example, the supply companies, the customers, or the people who visit the site); and the statistical variables, characteristics measured for each statistical unit (for example, if the customer is the statistical unit, customer characteristics might include the amounts spent, methods of payment and socio-demographic profiles).

The statistical units may be the entire reference population (for example, all the customers of the company) or just a sample. There is a large body of work on the statistical theory of sampling and sampling strategies, but we will not go into details here (see Barnett, 1974).

Working with a representative sample rather than the entire population may have several advantages. On the one hand it can be expensive to collect complete information on the entire population, while on the other hand the analysis of large data sets can be time-consuming, in terms of analysing and interpreting the results (think, for example, about the enormous databases of daily telephone calls which are available to mobile phone companies).

The statistical variables are the main source of information for drawing conclusions about the observed units which can then be extended to a wider population. It is important to have a large number of statistical variables; however, such variables should not duplicate information. For example, the presence of the customers' annual income may make the monthly income variable superfluous.

Once the units and the variables have been established, each observation is related to a statistical unit, and, correspondingly, a distinct value (level) for each variable is assigned. This process leads to a data matrix.

Two different types of variables arise in a data matrix: qualitative and quantitative. Qualitative variables are typically expressed verbally, leading to distinct categories. Some examples of qualitative variables include sex, postal codes, and brand preference.

Qualitative variables can be sub-classified into nominal, if their distinct categories appear without any particular order, or ordinal, if the different categories are ordered. Measurement at a nominal level allows us to establish a relation of equality or inequality between the different levels ($=, \neq$). Examples of nominal measurements are the colour of a person's eyes and the legal status of a company. The use of ordinal measurements allows us to establish an ordered relation between the different categories. More precisely, we can affirm which category is bigger or better ($=, >, <$) but we cannot say by how much. Examples of ordinal measurements are the computing skills of a person and the credit rate of a company.

Quantitative variables, on the other hand, are numerical – for example age or income. For these it is also possible to establish connections and numerical relations among their levels. They can be classified into discrete quantitative variables, when they have a finite number of levels (for example, the number of telephone calls received in a day), and continuous quantitative variables, if the levels cannot be counted (for example, the annual revenues of a company).

Note that very often the levels of ordinal variables are 'labelled' with numbers. However, this labelling does not make the variables into quantitative ones.

Once the data and the variables have been classified into the four main types (qualitative nominal and ordinal, quantitative discrete and continuous), the database must be transformed into a structure which is ready for a statistical analysis, the data matrix. The data matrix is a table that is usually two-dimensional, where the rows represent the n statistical units considered and the columns represent the p statistical variables considered. Therefore the generic element (i, j) of the matrix $i = 1, \ldots, n$ and $j = 1, \ldots, p$ is a classification of the statistical unit i according to the level of the jth variable.

The data matrix is where data mining starts. In some cases, as in, for example, a joint analysis of quantitative variables, it acts as the input of the analysis phase. In other cases further pre-processing is necessary. This leads to tables derived from data matrices. For example, in the joint analysis of qualitative variables it is a good idea to transform the data matrix into a contingency table. This is a table with as many dimensions as the number of qualitative variables that are in the data set. We shall discuss this point in more detail in the context of the representation of the statistical variables in frequency distributions.

2.2 Data matrices and their transformations

The initial step of a good statistical data analysis has to be exploratory. This is particularly true of applied data mining, which essentially consists of searching for relationships in the data at hand, not known a priori. Exploratory data analysis is usually carried out through computationally intensive graphical representations and statistical summary measures, relevant for the aims of the analysis.

Exploratory data analysis might thus seem, on a number of levels, equivalent to data mining itself. There are two main differences, however. From a statistical point of view, exploratory data analysis essentially uses descriptive statistical techniques, while data mining, as we will see, can use both descriptive and inferential methods, the latter being based on probabilistic methods. Also there is a considerable difference in the purpose of the two analyses. The prevailing purpose of an exploratory analysis is to describe the structure and the relationships present in the data, perhaps for subsequent use in a statistical model. The purpose of a data mining analysis is the production of decision rules based on the structures and models that describe the data. This implies, for example, a considerable difference in the use of alternative techniques. An exploratory analysis often consists of several different exploratory techniques, each one capturing different potentially noteworthy aspects of the data. In data mining, on the other hand, the various techniques are evaluated and compared in order to choose one for later implementation as a decision rule. A further discussion of the differences between exploratory data analysis and data mining can be found in Coppi (2002).

The next chapter will explain exploratory data analysis. First, we will discuss univariate exploratory analysis, the examination of available variables one at a time. Even though the observed data is multidimensional and, therefore, we need to consider the relationships between the available variables, we can gain a great

deal of insight by examining each variable on its own. We will then consider multivariate aspects, starting with bivariate relationships.

Often it seems natural to summarise statistical variables with a frequency distribution. As it happens for all procedures of this kind, the summary makes the analysis and presentation of the results easier but it also naturally leads to a loss of information. In the case of qualitative variables the summary is justified by the need to be able to carry out quantitative analysis on the data. In other situations, such as in the case of quantitative variables, the summary is done essentially with the aim of simplifying the analysis.

We start with the analysis of a single variable (univariate analysis). It is easier to extract information from a database by starting with univariate analysis and then going on to a more complicated analysis of multivariate type. The determination of the univariate distribution frequency starting off from the data matrix is often the first step of a univariate exploratory analysis. To create a frequency distribution for a variable it is necessary to establish the number of times each level appears in the data. This number is called the absolute frequency. The levels and their frequency together give the frequency distribution.

Multivariate frequency distributions are represented in contingency tables. To make our explanation clearer we will consider a contingency table with two dimensions. Given such a data structure it is easy to calculate descriptive measures of association (odds ratios) or dependency (chi-square).

The transformation of the data matrix into univariate and multivariate frequency distributions is not the only possible transformation. Other transformations can also be very important in simplifying the statistical analysis and/or the interpretation of the results. For example, when the p variables of the data matrix are expressed in different units of measure it is a good idea to standardise the variables, subtracting the mean of each one and dividing it by the square root of its variance. The variable thus obtained has mean equal to zero and variance equal to unity.

The transformation of data is also a way of solving quality problems because some data may be missing or may have anomalous values (outliers). Two main approaches are used with missing data: (a) it may be removed; (b) it may be substituted it by means of an appropriate function of the remaining data. A further problem occurs with outliers. Their identification is often itself a reason for data mining. Unlike what happens with missing data, the discovery of an anomalous value requires a formal statistical analysis, and usually it cannot be eliminated. For example, in the analysis of fraud detection (related to telephone calls or credit cards, for example), the aim of the analysis is to identify suspicious behaviour. For more information about the problems related to data quality, see Han and Kamber (2001).

2.3 Complex data structures

The application aims of data mining may require a database not expressible in terms of the data matrix we have used up to now. For example, there are often

other aspects of data collection to consider, such as time and/or space. In this kind of application the data is often presented aggregated or divided (for example, into periods or regions) and this is an important aspect that must be considered (on this topic see Diggle *et al*., 1994).

The most important case refers to longitudinal data – for example, the comparison in n companies of the p budget variables in q subsequent years. In this case there will be a three-way matrix that can be described by three dimensions: n statistical units, p statistical variables and q time periods. Another important example of data matrices with more than two dimensions concerns the presence of data related to different geographic areas. In this case, as in the previous one, there is a three-way matrix with space as the third dimension – for example, the sales of a company in different regions or the satellite surveys of the environmental characteristics of different regions. In such cases, data mining should use times series methods (for an introduction see Chatfield, 1996) or spatial statistics (for an introduction see Cressie, 1991).

However, more complex data structures may arise. Three important examples are text data, web data, and multimedia data. In the first case the available database consists of a library of text documents, usually related to each other. In the second case, the data is contained in log files that describe what each visitor does at a web site during a session. In the third case, the data can be made up of texts, images, sounds and other forms of audio-visual information that is typically downloaded from the internet and that describes an interaction with the web site more complex than the previous example. Obviously this type of data implies a more complex analysis. The first challenge in analysing this kind of data is how to organize it. This has become an important research topic in recent years (see Han and Kamber, 2001). In Chapter 6 we will show how to analyze web data contained in a log file.

Another important type of complex data structure arises from the integration of different databases. In modern applications of data mining it is often necessary to combine data that come from different sources, for example internal and external data about operational losses, as well as perceived expert opinions (as in Chapter 12). For further discussion about this problem, also known as data fusion, see Han and Kamber (2001).

Finally, let us mention that some data are now observable in continuous rather than discrete time. In this case the observations for each variable on each unit are a function rather than a point value. Important examples include monitoring the presence of polluting atmospheric agents over time and surveys on the quotation of various financial shares. These are examples of continuous time stochastic processes which are described, for instance, in Hoel *et al*. (1972).

2.4 Summary

In this chapter we have given an introduction to the organisation and structure of the databases that are the object of the data mining analysis. The most important point is that the planning and creation of the database cannot be ignored but it is

one of the most important data mining phases. We see data mining as a process consisting of design, collection and data analysis. The main objectives of the data mining process are to provide companies with useful/new knowledge in the sphere of business intelligence. The elements that are part of the creation of the database or databases and the subsequent analysis are closely interconnected. Although the chapter summarises the important aspects given the statistical rather than computing nature of the book, we have tried to provide an introductory overview.

We conclude this chapter with some useful references for the topics introduced in this chapter. The chapter started with a description of the various ways in which we can structure databases. For more details on these topics, see Han and Kamber (2001), from a computational point of view; and Berry and Linoff (1997, 2000) from a business-oriented point of view. We also discussed fundamental classical topics, such as measurement scales. This leads to an important taxonomy of the statistical variables that is the basis of the operational distinction of data mining methods that we adopt here. Then we introduced the concept of data matrices. The data matrix allows the definition of the objectives of the subsequent analysis according to the formal language of statistics. For an introduction to these concepts, see Hand *et al*. (2001). We also introduced some transformations on the data matrix, such as the calculation of frequency distributions, variable transformations and the treatment of anomalous or missing data. For all these topics, which belong the preliminary phase of data mining, we refer the reader to Hand *et al*. (2001), from a statistical point of view, and Han and Kamber (2001), from a computational point of view. Finally, we briefly described complex data structures; for more details the reader can also consult Hand *et al*. (2001) and Han and Kamber (2001).

CHAPTER 3

Summary statistics

In this chapter we introduce univariate summary statistics used to summarize the distribution of univariate variables. We then consider multivariate distributions, starting with summary statistics for bivariate distributions and then moving on to multivariate exploratory analysis of qualitative data. In particular, we compare some of the numerous summary measures available in the statistical literature. Finally, in consideration of the difficulty in representing and displaying high-dimensional data and results, we discuss a popular statistical method for reducing dimensionality, principal components analysis.

3.1 Univariate exploratory analysis

3.1.1 Measures of location

The most common measure of location is the (arithmetic) mean, which can be computed only for quantitative variables. The mean of a set $x_1, x_2, \ldots, x_N$ of N observations is given by

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \sum \frac{x_i}{N}.$$

We note that, in the calculation of the arithmetic mean, the biggest observations, can counterbalance and even overpower the smallest ones. Since, all the observations are used in the computation of the mean, its value can be influenced by outlying observations In financial data where extreme observations are common, this happens often and, therefore, alternatives to the mean are probably preferable as measures of location.

The previous expression for the arithmetic mean can be calculated on the data matrix. Table 3.1 shows the structure of a data matrix and Table 3.2 an example. When univariate variables are summarised with the frequency distribution, the arithmetic mean can also be calculated directly from the frequency distribution. This computation leads, of course, to the same mean value and saves computing time. The formula for computing the arithmetic mean from the frequency distribution is given by

$$\overline{x} = \sum x_i^* p_i.$$

Applied Data Mining for Business and Industry, Second Edition Paolo Giudici and Silvia Figini
 ISBN: 978-0-470-05886-2

Table 3.1 Data matrix.

	1	j	p	
1	$X_{1,1}$	$X_{1,j}$	$X_{1,p}$	
$\vdots$				
i	$X_{i,1}$	$X_{i,j}$	$X_{i,p}$	
$\vdots$				
n	$X_{n,1}$	$X_{n,j}$	$X_{n,p}$	

Table 3.2 Example of a data matrix.

	Y	X_1	X_2	...	X_5	...	...	...	...	...	X_{20}
N 1	1	1	18	...	1049			...			1
...											
N 34	1	4	24	...	1376			...			1
...											
...											
N 1000	0	1	30	...	6350			...			1

This formula is known as the weighted arithmetic mean, where the x_i^* indicate the distinct levels that the variable can take on and p_i is the relative frequency of each of these levels.

We list below the most important properties of the arithmetic mean:

- The sum of the deviations from the mean is zero: $\sum(x_i - \overline{x}) = 0$.
- The arithmetic mean is the constant that minimises the sum of the squares of the deviations of each observation from the constant itself: $\min_a \sum(x_i - a)^2 = \overline{x}$.
- The arithmetic mean is a linear operator: $N^{-1} \sum(a + bx_i) = a + b\overline{x}$.

A second simple measure of position or location is the modal value or mode. The mode is a measure of location computable for all kinds of variables, including qualitative nominal ones. For qualitative or discrete quantitative characters, the mode is the level associated with the greatest frequency. To estimate the mode of a continuous variable, we generally discretize the values that the variables assumes in intervals and compute the mode as the interval with the maximum density (corresponding to the maximum height of the histogram). To obtain a unique mode the convention is to use the middle value of the mode's interval.

Finally, another important measure of position is the median. Given an ordered sequence of observations, the median is the value such that half of the observations are greater than and half are smaller than it. The median can be computed for quantitative variables and ordinal qualitative variables. Given N observations in non-decreasing order the median is: