Exploratory Data Mining and Data Cleaning

Tamraparni Dasu
Theodore Johnson
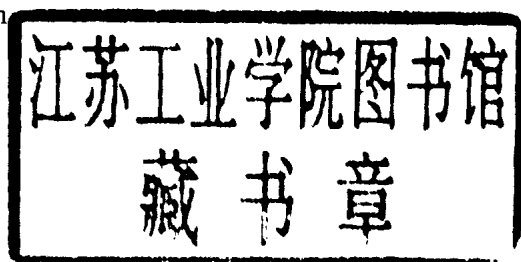
Exploratory Data Mining and Data Cleaning

TAMRAPARNI DASU

THEODORE JOHNSON

AT&T Labs, Research Division
Florham Park, NJ

WILEY-
INTERSCIENCE

Preface

As data analysts at a large information-intensive business, we often have been asked to analyze new (to us) data sets. This experience was the original motivation for our interest in the topics of exploratory data mining and data quality. Most data mining and analysis techniques assume that the data have been joined into a single table and cleaned, and that the analyst already knows what she or he is looking for. Unfortunately, the data set is usually dirty, composed of many tables, and has unknown properties. Before any results can be produced, the data must be cleaned and explored—often a long and difficult task.

Current books on data mining and analysis usually focus on the last stage of the analysis process (getting the results) and spend little time on how data exploration and cleaning is done. Usually, their primary aim is to discuss the efficient implementation of the data mining algorithms and the interpretation of the results. However, the true challenges in the task of data mining are:

- Creating a data set that contains the relevant and accurate information, and
- Determining the appropriate analysis techniques.

In our experience, the tasks of exploratory data mining and data cleaning constitute 80% of the effort that determines 80% of the value of the ultimate data mining results. Data mining books (a good one is [56]) provide a great amount of detail about the analytical process and advanced data mining techniques. However they assume that the data has already been gathered, cleaned, explored, and understood.

As we gained experience with exploratory data mining and data quality issues, we became involved in projects in which data quality improvement was the goal of the project (i.e., for operational databases) rather than a prerequisite. Several books recently have been published on the topic of ensuring data quality (e.g., the books by Loshin [84], by Redman [107]), and by English [41]). However, these books are written for managers and take a

managerial viewpoint. While the problem of ensuring data quality requires a significant managerial support, there is also a need for technical and analytic tools. At the time of this writing, we have not seen any organized exposition of the technical aspects of data quality management. The most closely related book is Pyle [102], which discusses data preparation for data mining. However, this text has little discussion of data quality issues or of exploratory data mining—pre-requisites even to preparing data for data mining.

Our focus in this book is to develop a systematic process of data exploration and data quality management. We have found these seemingly unrelated topics to be inseparable. The exploratory phase of any data analysis project inevitably involves sorting out data quality problems, and any data quality improvement project inevitably involves data exploration. As a further benefit, data exploration sheds light on appropriate analytic strategies.

Data quality is a notoriously messy problem that refuses to be put into a neat container, and therefore is often viewed as technically intractable. We have found that data quality problems can be addressed, but doing so requires that we draw on methods from many disciplines: statistics, exploratory data mining (EDM), databases, management, and metadata. Our focus in this book is to present an integrated approach to EDM and data quality. Because of the very broad nature of the subject, the exposition tends to be a summarization of material discussed in great detail elsewhere (for which we provide references), with an emphasis on how the techniques relate to each other and to EDM and data quality. Some topics (such as data quality metrics and certain aspects of EDM) have no other good source, so we discuss them in greater detail.

EXPLORATORY DATA MINING (EDM)

Data sets of the twenty-first century are different from the ones that motivated analytical techniques of statistics, machine learning and others. Earlier data sets were reasonably small and relatively homogeneous so that the structure in them could be captured with compact models that had large but a manageable number of parameters. Many researchers have focused on scaling the methods to run efficiently and quickly on the much larger data sets collected by automated devices. In addition, methods have been developed specifically for massive data (i.e., data mining techniques). However, there are two fundamental issues that need to be addressed before these methods can be applied.

- A "data set" is often a patchwork of data collected from many sources, which might not have been designed for integration. One example of this problem is when two corporate entities providing different services to a common customer base merge to become a single entity. Another is when different divisions of a "federation enterprise" need to merge their data

stores. In such situations, approximate matching heuristics are used to combine the data. The resulting patchwork data set will have many data quality issues that need to be addressed. The data are likely to contain many other data glitches, and these need to be treated as well.

- Data mining methods often do not focus on the "appropriateness of the model for the data," namely, goodness-of-fit. While finding the best model in a given class of models is desirable, it is equally important to determine the class of models that best fits the data.

There is no simple or single method for analyzing a complex, unfamiliar data set. The task typically requires the sequential application of disparate techniques, leveraging the additional information acquired at each stage to converge to a powerful, accurate and fast method. The end-product is often a "piecewise technique" where at each stage we might have had to adapt or extend, to improvise on an existing method. The importance of such an approach has been emphasized by statisticians such as John Tukey [123] and more recently in the machine learning community, for instance, in the Auto-Class project [19].

DATA QUALITY

A major confounding factor in EDM is the presence of data quality issues. These are often unearthed as "interesting patterns" but on closer examination prove to be artifacts. We emphasize this aspect in our case study, since typically data analysts spend a significant portion of their time weeding-out data quality problems. No matter how sophisticated the data mining techniques, bad data will lead to misleading findings.

While most practitioners of data analysis are aware of the pitfalls of data quality issues, it is only recently that there has been an emphasis on the systematic detection and removal of data problems. There have been efforts directed at managing processes that generate the data, at cleaning up databases (e.g. merging/purging of duplicates), and at finding tools and algorithms for the automatic detection of data glitches. Statistical methods for process control (predominantly univariate) that date back to quality control charts developed for detecting batches of poorly produced lots in industrial manufacturing are often adapted to monitor fluctuations in variables that populate databases.

For operations databases, data quality is an end in itself. Most business (and governmental, etc.) processes involve complex interactions between many databases. Data quality problems can have very expensive manifestations (e.g., "losing" a cross-country cable, forgetting to bill customers). In this electronic age, many businesses (and governmental organizations, etc.) would like to "e-enable" their customers—that is, let them examine the relevant parts of the

operational databases to manage their own accounts. Depending on the state of the underlying databases, this can be embarrassing or even impossible.

SUMMARY

In this book, we intend to:

- Focus on developing a modeling strategy through an iterative data exploration loop and incorporation of domain knowledge;
- Address methods for dealing with data quality issues that can have a significant impact on findings and decisions, using commercially available tools as well as new algorithmic approaches;
- Emphasize application in real-life scenarios throughout the narrative with examples;
- Highlight new approaches and methodologies, such as the *DataSphere* space partitioning and summary-based analysis techniques, and approaches to developing data quality metrics.

The book is intended for serious data analysts everywhere that need to analyze large amounts of unfamiliar, potentially noisy data, and for managers of operations databases. It can also serve as a text on data quality to supplement an advanced undergraduate or graduate level course in large-scale data analysis and data mining. The book is especially appropriate for a cross-disciplinary course in statistics and computer science.

ACKNOWLEDGMENTS

Contents

CHAPTER 1

Exploratory Data Mining and Data Cleaning: An Overview

1.1 INTRODUCTION

Every data analysis task starts by gathering, characterizing, and cleaning a new, unfamiliar data set. After this process, the data can be analyzed and the results delivered. In our experience, the first step is far more difficult and time consuming than the second. To start with, data gathering is a challenging task complicated by problems both sociological (such as turf sensitivity) and technological (different software and hardware platforms make transferring and sharing data very difficult). Once the data are in place, acquiring the metadata (data descriptions, business rules) is another challenge. Very often the metadata are poorly documented. When we finally are ready to analyze the data, its quality is suspect. Furthermore, the data set is usually too large and complex for manual inspection.

Sometimes, improved data quality is itself the goal of the analysis, usually to improve processes in a production database (e.g., see the case study in Section 5.5.1). Although the goal seems different than that of making an analysis, the methods and procedures are quite similar—in both cases we need to understand the data, then take steps to improve data quality.

Fortunately, automated techniques can be applied to help understand the data (**Exploratory Data Mining**, or **EDM**), and to help ensure data quality (by **data cleaning** and applying **data quality metrics**). In this book we present these techniques and show how they can be applied to prepare a data set for analysis. This chapter will briefly outline the challenges posed to the analysis of massive data, the strategies for taming the data, and an overview of data exploration and cleaning methods, including developing meaningful data quality definitions and metrics.

1.2   CAUTIONARY TALES

A first question to ask is, why are data exploration and data preparation needed? Why not just go ahead and analyze the data? The answer is that the results are almost guaranteed to be flawed. More specifically, some of the problems that occur are:

- **Spurious results**: Data sets usually contain artifacts generated by external sources that are of no interest to us but get mixed up with genuine patterns of interest. For example, a study of traffic on a large telecommunications company's data network revealed interesting behavior over time. We were able to detect glitches caused by delays in gathering and transmitting traffic characteristics (e.g., number of packets) and remove such delays from inherent bursty patterns in the traffic. If we had not cleaned the data, we would have included the glitches caused by delays in the "signature usage pattern" of the customer, and would have detected misleading deviations from the glitched signatures in future time series.

- **Misplaced faith in black boxes**: Data mining is sometimes perceived as a black box, where you feed the data in and interesting results and patterns emerge. Such an approach is particularly misleading when no prior knowledge or experience is used to validate the results of the mining exercise. Consider the case of clustering, a method often used to find hidden groupings in the data for tasks such as target marketing. It is very hard to find good clusters without a reasonable estimate of the number of groups, the relative sizes of these groups (e.g., cluster 1 is 10 times larger than cluster 2) and the logic used by the clustering algorithm. For example, if we use a k-means algorithm that initializes cluster centers at random from the data, we need to choose at least 10 starting clusters to detect two clusters that constitute 10% and 90% of the total data set. Starting with fewer clusters would result in the algorithm finding one big cluster containing most of the points, with a few outliers constituting the other clusters.

  Log-linear models (e.g., logistic regression) are another common example of misplaced faith. The models are successful when the appropriate number of parameters and the correct explanatory variables are included. The model will not fit well if too few parameters and irrelevant variables are included in it, even if in reality the logistic regression model is the correct choice. It is important to explore the data to arrive at an appropriate analytical model.

- **Limitations of Popular Models**: Very often, a model is chosen because it is well understood or because the software is available, irrespective of the nature of the data. Analysts rely on the robustness of the models, even when underlying assumptions about the distribution (often the Normal

density) *do not hold. However, it is important to recognize that, although* classical parametric methods based on distributional and model assumptions are compact, powerful and accurate when used in the right conditions, they have limited applicability. They are not suitable for scenarios where not enough is known about the data or its distribution, to validate the assumptions of the classical methods. A good example is linear regression, which is often used inappropriately, because it is easy to use and interpret. The underlying assumptions of linear effect of variables and the form of error distributions *are rarely verified. A random data set might* yield a linear regression model with a "reasonable" R-square goodness-of-fit measure, leading to a false confidence in the model.

Even if a model is applicable, it may be difficult to implement because of the scale of the data. Many nonparametric methods, such as clustering, machine learning, neural networks and others, are iterative and require multiple passes over all the data. On very large data sets, they may be too slow.

• **Buyer Beware—No Guarantees**: Many data mining techniques do not provide any goodness-of-fit guarantees. For example, a clustering mechanism might find the "best" clusters as defined by some distance metric, but does not answer the question of how well the clusters replicate the structure in the data. Testing the goodness-of-fit of clustering results with respect to the data can be time consuming, involving simulation techniques. As a result, validation of clustering in the context of appropriateness to the data is often not implemented. The best or optimal model could still be very poor at representing the underlying data. For example, many financial firms (such as Long Term Capital Management) have mined data sets to find similarities or differences in the prices of various securities. In the case of LTCM, the analysts searched for securities whose price tended to move in opposite directions and placed hedges by purchasing both. Unfortunately, these models proved to be inaccurate, and LTCM lost billions of dollars when the price of the securities suddenly moved in the same direction.

Another frequently encountered pitfall of casual data mining is spurious correlations. It is possible to find random time series that move together over a period of time (e.g., the NASDAQ index and rainfall in Bangladesh) but have no identifiable association, let alone causal relationship. An accompanying hazard is the tendency to tailor hypotheses to the findings of a data mining exercise. A classical example is the beer–diaper co-occurrence revealed by mining supermarket purchase data. However, its not likely that one can increase beer sales by stocking shelves with diapers.

We hope that the cautionary tales show that it is essential that the analyst must clean and understand the data before analyzing it.

1.3 TAMING THE DATA

There are many books that address data analysis and model fitting in which a
single approach (logistic regression, neural networks) stands out as the method
of choice. In our experience, however, getting to the point where the model-
ing strategy is clear requires skill, science, and the lion's share of the work. The
effectiveness of the later analysis strongly depends on the knowledge learned
during the earlier ground work. For an example, the analyst needs to know,
what are the variables that are relevant (e.g., for predicting probability of
recovery from a disease—vital statistics, past history, genetic propensity)? Of
these, how many variables can be measured and how many are a part of the
available data? How many are correlated and redundant? Which values are
suspicious and possibly inaccurate?

The work of identifying the final analysis strategy is an iterative (but com-
putationally inexpensive) process alternating between **exploratory data
mining (EDM)** and data cleaning (improving data quality **(DQ)**). EDM con-
sists of simple and fast summaries and analyses that reveal characteristics of
the data, such as typical values (averages, medians), variability (variance,
range), prevalence of different values (quantiles) and inter-relationships (cor-
relations). During the course of EDM, certain data points that seem to be
unlikely (e.g., an outlier such as an 80-year-old third grader, a sign-up date of
08-31-95 for a service launched in 1997) motivate further investigation. Closer
scrutiny often finds data quality issues (a mistyped value, a system default
date), which, when fixed, result in cleaner, better quality data. In a later
chapter, we discuss a case study related to a provisioning data base where
clearing up data problems unearthed by EDM allowed us to significantly sim-
plify the model needed to represent the structure in the data. We note that
addressing DQ issues involves consulting with domain experts and incorpo-
rating their knowledge into the next round of EDM. Therefore, EDM and DQ
have to be performed in conjunction.

1.4 CHALLENGES

Unfortunately, the analyst has to do considerable ground work before the
underlying structure in the data comes into focus. Some of the challenges of
EDM and DQ are:

- **Heterogeneity and Diversity**: The data are often collected from many
  sources and stitched together. This is particularly true of data gathered
  from different organizations of a single "federation enterprise", or of an
  enterprise resulting from corporate mergers. Often, it is a problem even
  for data gathered from different departments in the same organization.
  The data might also be gathered from outside vendors (e.g., demograph-
  ics). While the combined information is presented to the analyst as a

single data set, it usually contains a superposition of several statistical processes. Analyzing such data using a single method or a black box approach can produce misleading, if not totally incorrect results, as will be explained later.

- **Data Quality**: Gathering data from different organizations, companies, and sources makes the information rich in content but poor in quality. It is hard to correlate data across sources since there are often no common keys to match on. For example, we might have information about Ms. X, who buys clothing from one business unit and books from another. If there is no common identifier in the two databases (such as customer ID, phone number, or social security number) it is hard to combine the information from the two business units. Keys like names and addresses are often used for the matching. However, there is no standard for names and addresses (Elizabeth, Liz; Street, St.; Saint, St.; other variants) so that matching databases using such **soft keys** is inexact (and time consuming), resulting in many data quality issues. Information related to the same customer might not be matched, whereas spurious matches might occur between similarly spelled names and addresses.

  Data quality issues abound in data sets generated automatically (telecommunication switches, Internet routers, e-transactions). Software, hardware and processing errors (reverting to defaults, truncating data, incomplete processing) are frequent.

  Other sources of data integrity issues are bad data models and inadequate documentation. The interpretation of an important attribute might depend on ancillary attributes that are not updated properly. For example, "Var A represents the current salary if Var B is populated. If not, it represents the salary upon termination. The termination date is represented by Variable C that is updated every three months." For Var A to be accurate, timely and complete, Var B and Var C should be maintained diligently. Furthermore, interpretation of Var A requires good documentation that is very rarely available. Such metadata reside in many places, often passed on through word-of-mouth or informal notes.

  Finally, there are the challenges of missing attributes, confusing default values (such as zero, i.e. zero revenue differs significantly from revenue whose value is not known that month) and good old-fashioned manual errors (data clerk entering elementary school student profile types age as 80 instead of 08). In the latter instance, if we did not know the data characteristics (typical ages of elementary school children) we would have no reason to suspect that the high value is corrupt, which would have significantly altered the results (e.g., average age of elementary school kids).

- **Scale**: Often the sheer volume of the data (e.g., an average of 60 Gbytes a day of packet flows on the network) is intimidating. Aside from the issues of collection, storage, and retrieval, the analyst has to worry about

summarizing the data meaningfully and accurately, trading-off storage constraints versus future analytical needs. Suppose, for example, that to perform a time series analysis we need at least 30 days worth of data. However, we can efficiently store and retrieve only a week's worth at the most. Therefore, computing and storing statistical summaries (averages, deviations, histograms) that will facilitate sophisticated analysis, as well as developing summary-based analyses, are a major part of the analyst's challenge.

- **New Data Paradigms**: The term "data" has taken on a broad meaning— any information that needs to be analyzed is considered "data". Nowadays, data come in all flavors. We have data that are scraped off the web, text documents, streaming data that accumulate very quickly, server logs from web servers and all kinds of audio and image data. It is a challenge to collect, store, integrate and manage such disparate types of data. There are no established methods for doing this as yet.

1.5   METHODS

In this section we give a brief outline of EDM and DQ methods. In subsequent chapters, we will explore these topics in detail.

A typical data set consists of data points, where each data point is defined by a set of variables or attributes. For example, a data point in a hypothetical data set of network traffic might be described by:

$$(source\_IP\_address, destination\_IP\_address, \\ number\_of\_packets\_sent, number\_of\_hops, time\_taken)$$

The above set of variables enclosed in parentheses is called a **vector of attributes**, where each item in the vector represents an aspect of the data point. Each data point differs from the other. Some attributes, such as the IP address, are assigned and are completely known. Variables such as packets_sent and time_taken vary from data point to data point depending on many observable and hidden factors such as network capacity, the speed of the connection, the load on the network and so on. The variability or uncertainty in the values of the attributes can be represented compactly using a probabilistic law or rule represented by f. A well-known example of f is the Gaussian, or Normal, distribution. In a way, f represents a complete description of the data, so that if we know f, we can easily infer any fact we want to derive from the data. We will discuss this aspect more in Section 2.2. Estimating the probabilistic rule f is important and valuable, however it is also difficult. Therefore we break it up into smaller sequential phases, where we leverage the information from each phase to make informed assumptions about some aspect of f. The assumptions are often pre-requisites for more sophisticated approaches to estimating f.