Alberto Fernández · Salvador García
Mikel Galar · Ronaldo C. Prati
Bartosz Krawczyk · Francisco Herrera

Learning from Imbalanced Data Sets

Springer

Learning from Imbalanced Data Sets

Alberto Fernández • Salvador García • Mikel Galar
Ronaldo C. Prati • Bartosz Krawczyk
Francisco Herrera

Learning from Imbalanced Data Sets

Springer

Alberto Fernández
Department of Computer Science and AI
University of Granada
Granada, Granada, Spain

Salvador García
Department of Computer Science and AI
University of Granada
Granada, Granada, Spain

Mikel Galar
Institute of Smart Cities
Public University of Navarre
Pamplona, Spain

Ronaldo C. Prati
Department of Computer Science
Universidade Federal do ABC
Santo Andre, Brazil

Bartosz Krawczyk
Department of Computer Science
Virginia Commonwealth University
Richmond, VA, USA

Francisco Herrera
Department of Computer Science and AI
University of Granada
Granada, Spain

ISBN 978-3-319-98073-7 ISBN 978-3-319-98074-4 (eBook)
https://doi.org/10.1007/978-3-319-98074-4

Library of Congress Control Number: 2018955498

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my beloved wife, my family, friends, and close colleagues. For their support since the very beginning.

Alberto Fernández

To my family
Salvador García

To my family
Mikel Galar

To my family
Ronalo C. Prati

To my family
Bartosz Krawczyk

To my family
Francisco Herrera

Preface

Learning with imbalanced data refers to the scenario in which the amounts of instances that represent the concepts in a given problem follow a different distribution. The main issue when addressing such a learning problem is when the accuracy achieved for each class is also different. This situation occurs since the learning process of most classification algorithm is often biased toward the majority class examples, so that minority ones are not well modeled into the final system. Being a very common scenario in real-life applications, the interest of researchers and practitioners on the topic has grown significantly during these years.

Based on the experience of the authors after several years focused on imbalanced classification, this book aims at offering a general and comprehensible overview for anyone interested in this area of study. It contains a formal description of the problem and focuses on its main features and the most relevant proposed solutions. Additionally, it considers the different scenarios in Data Science for which the imbalanced classification can suppose a real challenge.

After a gentle introduction to the KDD process and current state of Data Science in the first chapter, the book then stresses the gap with standard classification tasks by establishing the foundations and reviewing the case studies with a direct application in this area in Chap. 2. Then, Chap. 3 introduces the main ad hoc evaluation metrics to be considered in this area of study. The book also covers the different approaches that have been traditionally applied to address the binary skewed class distribution. Specifically, it reviews cost-sensitive learning (Chap. 4), data-level preprocessing methods (Chap. 5), and algorithm-level solutions (Chap. 6), taking also into account those ensemble-learning solutions that embed any of the former alternatives (Chap. 7). Furthermore, it focuses in Chap. 8 on the extension of the problem for multi-class problems, where the former classical methods are no longer to be applied in a straightforward way.

The book includes in Chap. 9 some notes on data reduction, being provided in order to understand the advantages related to the use of this type of approaches. Then, Chap. 10 focuses on the data intrinsic characteristics that are the main causes which, added to the uneven class distribution, truly hinders the performance of

classification algorithms in this scenario. Finally, this book introduces some novel areas of study that are gathering a deeper attention on the imbalanced data issue. Specifically, Chap. 11 considers the classification of data streams, Chap. 12 the non-classical classification problems, and finally Chap. 13 discusses the scalability related to Big Data. To sum up, some examples of software libraries and modules to address imbalanced classification are given in Chap. 14.

This thorough review on the current and future state of imbalanced classification aims giving this topic the significance it deserves. In particular, the interest of research and academia is clearly shown by the rising number of publications and citations year by year. In the foreseeable future, it predictably will continue expanding with novel significant developments, as many contemporary real-world applications must be addressed from the viewpoint of imbalanced classification.

The intended audience of this book are developers and engineers aiming to apply imbalance-learning techniques to solve different kinds of real-world problems, as well as researchers and students needing a comprehensive review on techniques, methodologies, and tools for learning from imbalanced data.

We wish to thank all our collaborators of the research group "Soft Computing and Intelligent Information Systems." We are also thankful to our families for their helpful support.

Granada, Spain — Alberto Fernández
Granada, Spain — Salvador García
Pamplona, Spain — Mikel Galar
Santo Andre, Brazil — Ronalo C. Prati
Richmond, VA, USA — Bartosz Krawczyk
Granada, Spain — Francisco Herrera
June 2018

Contents

Acronyms

ADASYN	Adaptive synthetic sampling
AL	Active learning
ANN	Artificial neural network
AUC	Area under the curve
AUC_{ROC}	Area under the ROC curve
AUC_{PR}	Area under the precision-recall curve
CV	Cross-validation
CNN	Condensed nearest rule
DM	Data mining
DR	Dimensionality reduction
EC	Error concentration
EM	Expectation-maximization
FCV	Fold cross-validation
FS	Feature selection
IS	Instance selection
KDD	Knowledge discovery in data
KEEL	Knowledge extraction based on evolutionary learning
KNN	K-Nearest neighbors
LLE	Locally linear embedding
LVQ	Learning vector quantization
MCC	Matthews correlation coefficient
MDS	Multidimensional scaling
MI	Mutual information
MIL	Multi-instance learning
ML	Machine learning
MLL	Multilabel learning
MLP	Multilayer perceptron
MV	Missing value
NCL	Neighborhood cleaning rule
NN	Nearest neighbor
OSS	One-sided selection

PCA	Principal components analysis
PU-learning	Positive and unlabeled learning
RBFN	Radial basis function network
ROC	Receiver operating characteristic curve
SMOTE	Synthetic minority over-sampling technique
SONN	Self-organizing neural network
SSL	Semi-supervised learning
SVM	Support vector machine

Chapter 1
Introduction to KDD and Data Science

Abstract Nowadays, the availability of large volumes of data and the widespread use of tools for the proper extraction of knowledge information has become very frequent, especially in large corporations. This fact has transformed the data analysis by orienting it towards certain specialized techniques included under the umbrella of Data Science. In summary, Data Science can be considered as a discipline for discovering new and significant relationships, patterns and trends in the examination of large amounts of data. Therefore, Data Science techniques pursue the automatic discovery of the knowledge contained in the information stored in large databases. These techniques aim to uncover patterns, profiles and trends through the analysis of data using reconnaissance technologies, such as clustering, classification, predictive analysis, association mining, among others. For this reason, we are witnessing the development of multiple software solutions for the treatment of data and integrating lots of Data Science algorithms. In order to better understand the nature of Data Science, this chapter is organized as follows. Sections 1.2 and 1.3 defines the Data Science terms and its workflow. Then, in Sect. 1.4 the standard problems in Data Science are introduced. Section 1.5 describes some standard data mining algorithms. Finally, in Sect. 1.6 some of the non-standard problems in Data Science are mentioned.

1.1 Introduction

Recent technological advances imply that the capacities to generate and store data are increased everyday. Among the factors that influence this reality we can highlight the widespread use of bar codes and QR reading, the automation of all type of transactions (commercial, business, economic, scientific) and the advances in data collection, among others. In addition, the Internet has rapid access to information, where both data and results can be easily obtained by others equipment. In this sense, current organizations although distant in the space, are very close in the cyberspace. All of these has led to strong economies of scale through the pooling of databases, theoretical knowledge and successful results. Furthermore, in the last decades there has been a change in the organizational environment that has caused

A. Fernández et al., *Learning from Imbalanced Data Sets*,
https://doi.org/10.1007/978-3-319-98074-4_1

a strong competition. This implies a need for organizations of all kinds to be able to survive in such changing environments.

Besides, the evolution of mass storage devices (in relation to price – storage capacity), such as hard disks that can store gigabytes of information at a reduced price, has led to companies and organizations to store all kinds of information. Citing some examples, we may refer to the data of customers and their transactions, to telemetry data, patients, price evolution in markets, among others. In the beginning, this information was stored in files that were difficult to handle, but with the advent of database management systems this difficulty was reduced. With the time, the amount of data that was stored began to grow and, although the tools to perform the data management was suitable, the significant relationships existing between them, began to surpass the human capacities for analysis. At the same time, database systems had begun to be decentralized, hence the decisions lacked credibility, inefficiency and lack of productivity.

All this explosive data growth generated, in the late 1980s, the emergence of a new field of research called KDD [15]. Under these acronyms hides, as suggested by Fayyad et al. [9], "the non-trivial process of discovering valid, new patterns, potentially useful and understandable in large volumes of data". The KDD process has served to unite researchers from areas in principle dispersed as Artificial Intelligence, Statistics, Visualization Techniques, Mathematics, Automatic Learning or Databases in the search for efficient techniques and that help to find the potential knowledge that is immersed in the large volumes of data stored by organizations on a daily basis [10].

Although the name with which this area of research appeared was that of KDD, other names have been used for this same concept. Some of them are *Knowledge Discovery, Data Discovery, Discovery Information, Knowledge Extraction, Data Extraction, Pattem Discovery, DM, Data Science*. At present the names that enjoy greater acceptance have been the DM and Data Science [2, 29]. Both processes need smart methods to extract information from data and to optimize the results. In the beginning, DM was only used to refer to the stage of the process in which they are applied techniques and pattern discovery algorithms. However, currently it is used to refer to the overall process of extracting knowledge from the data. Similarly, the term Data Science is currently used to generalize the DM and KDD terms into a new discipline which encompasses techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science.

The great increase of data that the organizations have to analyze not only resulted in the appearance of Data Science, but at the same time Big Data concept emerges. One of the great problems of Data Science is that the data was never stored thinking that as a consequence, prior to the analysis, is necessary a process of integration and cleaning of data that in many cases results more expensive than the analysis itself. However, the appearance of the Data Warehouses as repositories of centralized information allows the processes can not be performed on data sets that have been previously integrated and subjected to cleaning processes.

For researchers in the fields of knowledge named previously, these two recent areas of research pose a great challenge to find a new way of thinking, designing, and implementing both the basis as the data analysis.

1.2 A Definition of Data Science

The diagram in Fig. 1.1 illustrates the idea that the KDD is a process [22], that is, it is a set of tasks or stages, which will be analyzed in detail throughout this chapter, and among which include:

- Establishment of a relevant problem.
- Selection of the appropriate data to solve the problem.
- Exploration and cleaning of data.
- Processing and modification of data.
- Application of modeling techniques (algorithms for the discovery of patterns).
- Obtaining and interpreting the models obtained.
- Use of knowledge obtained.
- Generation of new data from your application in the real world.

But how does the Data Science process differ from the analysis that other disciplines perform? Traditional systems of data exploitation are fundamentally based on the existence of previous hypotheses or models. Once the hypothesis is formulated, it is analyzed empirically from the information in the available data and the results obtained are interpreted as a response to the initial hypothesis.

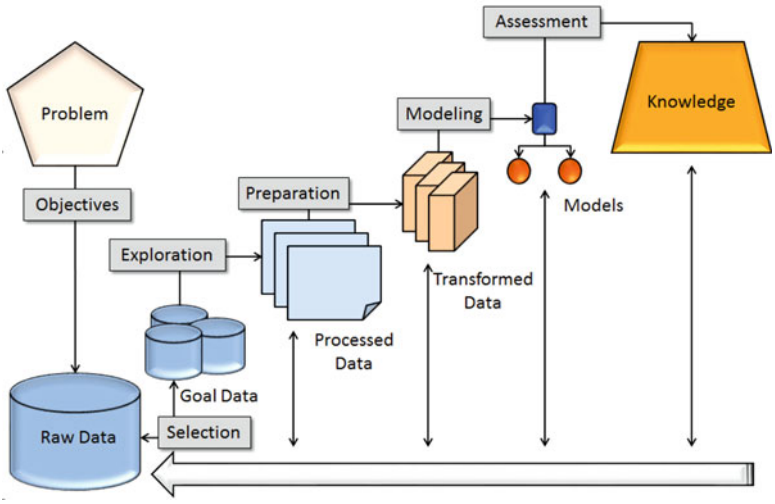

Fig. 1.1 KDD process

However, in common use, this methodology raises two problems. On the one hand, the individual who formulates the hypothesis must guess or know for certain what is the necessary information to accomplish the task. On the other hand, given the complexity of stored data and their interrelations, model verification is nowadays inadequate in many fields for decision making. The interpretation of results will thus be limited for its true quality.

Therefore, supplementing the above analysis with the possibility of discovering in an inductive manner for information and key hidden patterns in data is the main feature of Data Science. For instance, some examples are:

- *Automatic prediction of trends and behaviors*. Data Science automates the process of obtaining predictive information in big databases. Some issues that have traditionally required complex calculations can now be answered directly and quickly from the data. A typical example of a predictive problem is targeted marketing. Data Science uses data from past promotional campaigns to identify the objectives that are most likely to make future campaigns. Other examples of predictive problems include prediction of financial risk, identification of segments of the population likely to respond identically to certain events, etc.
- *Automatic discovery of previously unknown patterns*. The Data Science tools filter the data contained in large databases and identify previously hidden patterns. An example of pattern discovery is the analysis of sales data to identify apparently unrelated products that are often purchased together. Other problems of pattern discovery include detection of fraudulent transactions with credit cards and data identification anomalies that could represent input errors [7].

Data Science techniques can provide the benefits of automation on existing software or hardware platforms and can be implemented over new systems as existing platforms. When Data Science processes are implemented on high-performance parallel process systems, they can analyze very large databases in a few seconds, achieving the category of Big Data analytics.

1.3 The Data Science Process

In the initial definition of Data Science, the data refer to a set of facts or cases that confound the database. A pattern does reference to an expression in some language that serves to describe a subset of the data or a model applicable to that data. That is, a pattern is an instance of a determined model. Therefore, the extraction of patterns is understood as the extraction of a model for some data, that is, any high-level description of the data.

This process implies that Data Science is a conjunction of steps, although it is not trivial, because it is assumed to require a complex analysis. The patterns must be valid, with some degree of certainty, and novel, at least for the system and, preferably, for the user, to which should report some kind of benefit.

Everything indicated above implies that some set of measures can be defined to evaluate the patterns obtained. These measures may for instance to evaluate the goodness, utility, simplicity and certainty of the patterns. The Data Science process, as discussed above, is the process to apply in some database the required operations of selection, exploration, sampling, transformation and modeling methods for extracting interesting patterns that will represent knowledge.

The KDD process is an iterative and interactive process due to the fact it includes numerous steps in which the user has to make decisions. It is iterative because it may be necessary to access from any of the above, and interactive because the process is monitored and controlled by the user in a direct way.

Even though there are different alternatives described in the literature, the process consists of the following four main stages:

- **Selection of Goals:** in this phase, the problem has to be studied and we have to decide what is the goal of the project. It is also desirable to have expectations of success or failure of the project given that these concepts are relative. For example, depending on the problem, a model capable of successfully predicting the 70% of cases can be considered as a failure in absolute accuracy, but a great success if the procedure used previously only achieved a correct prediction of the 60% of the cases. With a good approach of the problem, It is easier to discover the data sources and the most suitable DM algorithms to be applied. A bad approach to the problem can lead us to wrong results. At this stage, the costs and economic benefits of the project have to be also estimated for the sake of achieving the best solution as possible.
- **Data preprocessing:** this stage of the KDD process is the one that most effort requires [25]. This phase consists of four main steps, although there could be more interpretations:
  - *Selection of the data:* the internal or external data sources are identified and the necessary subset of data is selected, either relations of one database or text files.
  - *Preparation of the data:* once the data to be used have been identified, we must understand the meaning of the attributes to detect integration errors, such as the existence of repeated data with different names or same data with different formats. These problems arise because the data may come from different sources, and not all of them store the same information in the same way. After this preprocess, what we will have is a data set suitable for the correct functioning of the remaining phases of the Data Science process [13].
  - *Transformation of data:* once analyzed the type of problem and the type of available data, we have to choose the algorithm or set of algorithms to be applied. As each algorithm requires a different format in the input data, we must transform the data to adjust the requirements of the selected algorithm/s.
  - *Reduction of data:* these techniques can be applied to achieve a reduced representation of the data set which will be much smaller in volume and tries

to keep most of the integrity of the original data [11]. The goal is to provide to the later DM algorithms with a mechanism to produce the same (or almost the same) outcome when it is applied over reduced data instead of the original data, at the same time as when mining becomes efficient.

- **Construction of the model:** it is the main stage because it is where the different data analysis algorithms are apply to the data, which were transformed, prepared and possibly reduced in the previous stages. During this stage, the patterns present in the data are searched. Depending on the algorithm selected, a different form will be obtained at the output. At this stage it is possible to use several times the same algorithm or even we can use different kinds of algorithms.
- **Analysis of the results:** it is the moment to interpret and evaluate the results obtained in the previous stage. Different techniques for visualization are often used to display the results obtained. Once the results are visualized, the user must interpret them, and if they do not meet their expectations, she must reapply the algorithms with other parameters, and even to run other algorithms to try to obtain more desirable results. All this makes the process of Data Science iterative. At this stage, we have to specify how to use the obtained results. They may be either integrated into an expert system or implemented as procedures in a database management system to make decisions.

1.3.1 Selection of the Data

At this stage, it is first necessary to assess the present problem we want to address. We will thus have to study the antecedents on how the problem has been solved by other organizations and to point out the advantages and shortcomings of the procedure that it is currently applied. Then, the objectives we want to approach with a Data Science process should be posed. Among others characteristics, we can stress quantifiable, realistic, relevant, multiple objectives clearly defined with a list of priorities.

Once the objectives have been defined, we must draw up an implementation plan specifying: the temporal duration, a budget, an analysis of monetary and opportunity costs as well as expectations of benefits, elaboration of a schedule and identification of possible external factors that are key to the organization.

It is worth to recall that this phase is key to the success of the Data Science process. Frequently, researchers and inexperienced analysts tend to think that the data are the origin of a Data Science process. This error usually ends with unproductive results and therefore a waste of time and resources. A thorough knowledge of the problem and the formulation of objectives are therefore vital in any Data Science task.

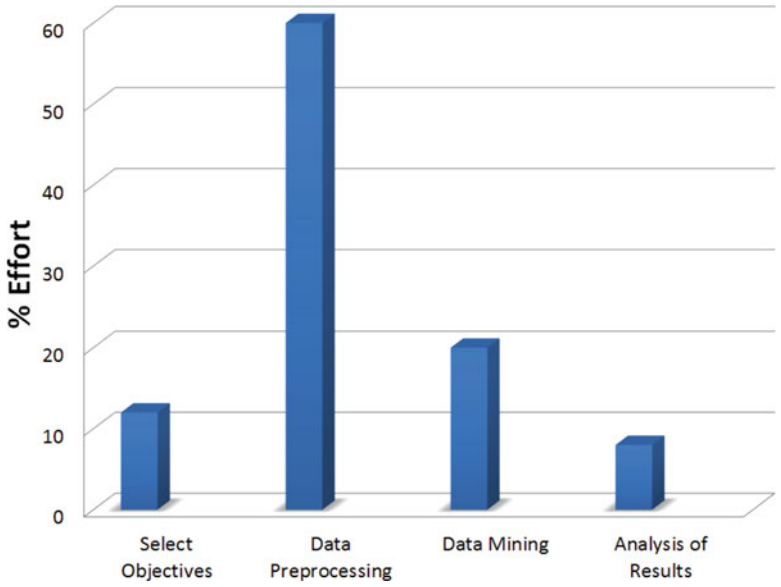

Fig. 1.2 Effort required for each stage in KDD

1.3.2 Data Preprocessing

Often, in organizations involved in Data Science projects, there is an excessive rush in the application of powerful analytical techniques for the extraction of hidden knowledge in the data. To use such techniques and tools, as explained above, it is necessary to develop one of the key parts of the project and of the longest time, which is the phase of Preprocessing prior to the application of the analysis algorithms [13]. Figure 1.2 shows the effort required at each stage of the KDD.

1.3.2.1 Why Is Preprocessing Required?

Inconsistencies, null values, extreme values and noise are properties of all data sets and relations in data bases. Incomplete data are generated for different reasons, for example, the attributes of interest are not always available or the information you have is erroneous. Other data are not stored because at the time of entering the data they were thought to be of no interest. Noise is again available for different reasons such as a simple problem in data collection instruments and personnel, other times it is due to transmission mechanisms or simple inconsistencies in code naming and assignment policies. In this way, data cleansing routines (Fig. 1.2) will help fill in the null values, identifying outliers and solving inconsistencies. Uncleaned data creates confusion for the scanning procedures and although some algorithms include

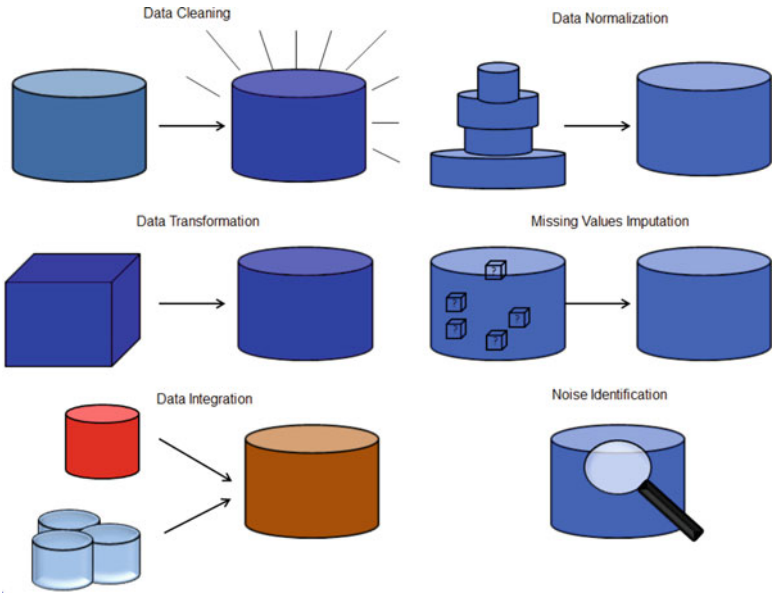

Fig. 1.3 Forms of data preprocessing

routines for cleaning these mechanisms are often not robust, so it is preferable to clean them beforehand.

However, returning to the problem that we wanted to solve, it is important to remember that we can have data from different sources, branches or organizations and that in each one of them possibly have data in more than one database. As a consequence we will need to integrate the data from multiple databases before proceeding with the analysis (Fig. 1.3), otherwise we will encounter redundancies and inconsistencies due to integration.

Finally, once we have the data ready for the analysis, we will find that we have the algorithm that we were going to apply requires categorical input of data and that our variables are given as numerical, which will lead us to transform the data before proceeding (Fig. 1.3).

1.3.3 Stages of the Data Preprocessing Phase

Each of these subphases is explained in detail below, analyzing the techniques that are applied to correct the defects that the data can present.

1.3.3.1 Selection of Data

The target of the selection stage is the identification of the available data sources and the extraction of the necessary data for a preliminary analysis, thus at the end of the phase you have the data that will be prepared to be submitted to Data Science techniques. It is obvious that the selection of the data depends on the type of problem to be solved and the goal pursued. Assuming that the data are gathered, the first task is to check the quantity and quality of the data. A good amount of data is needed to build robust models. But having large amounts of data is not enough, you will have to study each field, data types, maximum and minimum values in order to have large amounts of data with the highest quality.

Since the set of data to be selected will consist of a series of samples that will be described by means of a series of variables, it will be necessary to analyze the metadata (data about the data) associated with each variable to understand what each one means. Metadata not only provides a definition of the data or variable from the perspective of the business, but must provide information about data types, potential values, original source, format and other characteristics that have to do with the definition of the variable.

All this information is very important to take into account as it will be fundamental in later stages. Note that the type of algorithm to be applied will not only depend on the type of problem to solve but will depend on the type of variables used to describe the data. These types are usually divided into:

- Quantitative. They are subdivided into:
  - Discrete (number of persons, number of vehicles . . .).
  - Continues (salary, length, benefits . . .).
- Qualitative. You can distinguish:
  - Nominal. Name the object to which they refer without being able to establish an order (civil status, gender, colour, race, . . .)
  - Ordinal. An order can be established in its values (high, medium, low).

In this way it is usual to speak of qualitative variables and quantitative variables or of categorical and continuous variables. The nominal variables are at the most qualitative end, while the continuous variables stand out on the more quantitative side. Note that certain variables such as scales or rankings can be treated as discrete or ordinal variables depending on the case and therefore their definition may be less clear. When selecting the data, another important consideration is the life time of the variable, that is, to establish the period of time from which the variable will have lost its semantics or will no longer be significant.

1.3.3.2 Exploration of Data

The purpose of this subphase is to ensure the quality of the data that has been selected. As discussed above, the fact that data is clean and free from inconsistencies is a prerequisite for a successful Data Science project. On the other hand, the more and better the data is known, the easier it will be to know where to look in the modeling phase. The first task that must be done is a supervision of the structure of the data to be able to provide a first measure of the quality of the same.

In order to carry out this task, the visualization tools and statistical methods are usually employed. For categorical variables, the estimation of the frequency distributions of the values is the best way to understand the content. Simple tools like histograms or pie charts can help you visualize the distribution of each variable identifying null values and values out of range. When dealing with quantitative variables, it will be necessary to analyze measures such as the minimum and maximum values, the mean, the variance, the mode (value that occurs more frequently), the median (mean value), among others. Combining all these estimated, it will be possible to establish if a variable should be analyzed before continuing. Other useful tools are box-plots, histograms, or QQ-plot charts to study the distribution of variables and to show the distribution of one variable agains another different variable to analyze their relationship.

Once the data have been analyzed with the available tools, two of the tasks that are most frequently performed for each variable are the elimination of noise values, the processing of MVs and the detection of inconsistencies.

- *Noise Data:* noise is a random error or variance in a variable. Consequently, the variables affected by the noise will have values that fall outside the expected values for those variables. If these extreme values that are out of range are called outliers. Outliers can represent an opportunity to continue searching or simply be incorrect data. There are different types of and each one should be treated differently. Thus, for example, an outlier possibility is due to human errors in data collection. In this way an individual can appear with age over 1,000 years or with the negative salary. This error must be corrected. Another type of outliers is the one that is generated because certain operational changes have not been reflected in the Data Science environment. Clearly in this case the only action that has to be carried out is to update the metadata. Nevertheless, most of the noise involves little changes in data and advanced techniques must be applied in order to identify, remove or fix it [27].
- *Missing Values:* In most of the Data Science projects, when we face data analysis we find that many of the tuples (samples) have no value for certain attributes. Hence, two questions arise: what to do with that tuple? and, how can we fill in those values that we do not have? For this there are techniques ranging from ignoring the existence of these MVs, manually fill the data or use simple statistics such as average or correlations to obtain new values. The most advanced techniques and possibly the best approaches are the MV imputation based on predictive DM techniques. However, it should be remembered that no technique is perfect and that one has to be careful to avoid introducing more noise when eliminating MVs [21].

1.3.3.3 Transformation of Data

It represents a crucial phase because the success and accuracy of the models that will be obtained in the DM phase depends on how the data analyst decides to structure and present the input to the next phase. On the other hand, in this phase is when the data have to be codified to be a suitable input for the DM algorithms that will be used. In this way, if the algorithm to be used requires numerical input and the data selected are categorical or vice versa, it will be at this stage when the data is transformed so that they acquire the appropriate format. In addition, it is very common for new variables to be derived at this stage.

1.4 Standard Data Science Problems

The modeling phase is the central stage of the discovery process in which knowledge extraction algorithms are applied to previously preprocessed data. Actually, this step is inseparable from the next step in the chain of results analysis. In fact, often the analysis of the results obtained causes that it goes back again to the preprocessing phase in order to obtain more data or more attributes. In order for the process to be correct, it is essential that the analyst has the pre-processed data set, the corresponding metadata and all the data information that has been previously extracted in the previous steps of the analysis. What will happen in this stage will depend on the type of goal to be achieved. That is, it is not the same if the final result is a characterization of data or if the goal pursued is a predictive model where possibly the process will be longer and more complicated. The analysis of the results is one of the most important steps of the process.

For those who are first approaching a Data Science process, the number of existing algorithms to solve the same type of problem can lead to many confusions. Although it is difficult to establish a classification of the possible complications we can find in Data Science, it is even more difficult to find a procedure that establishes the algorithm suitable for each type of problem. In spite of all this, we will try to establish a guide that will help us to find the best types of algorithms to apply depending on the problem to be solved (goal) and the type of data that we are dealing with in each moment.

A first and general categorization of the problems will lead us to distinguish between descriptive problems (unsupervised learning) and predictive problems (supervised learning). Nevertheless, there are more complex problems assumed to be hybridizations, derivations or restricted formulations of the two mentioned basic problems. More details about them can be found in Sect. 1.6 of this chapter.

1.4.1 Descriptive Problems

In this context, we understand as a descriptive problem that whose goal is simply to find a description of the study data. These types of problems belong to the example

of knowing the clients of an organization (characteristics of the customers), or finding the products that are often bought together, or the symptoms of diseases presented together. The goal of all these problems is a description of the source data set. Nevertheless, analyzing these examples more in detail we observe that although both try to discover characteristics of the set origin, in the first case, (description of the clients) what is intended is to organize the clients into groups more or less homogeneous and extract the characteristics of these objects. However, in the second type of queries (products that are bought together or symptoms of diseases presented together), although the problem remains descriptive, the type of description required is different, since what is sought is to find associations between the values of attributes or properties of these objects. This causes a more detailed division of the descriptive problem into:

- *Clustering Analysis:* It refers to problems where the goal is to find homogeneous groups in the source population. These problems are also called profile segmentation. The typical example of segmentation is to segment customers.
- *Association Analysis:* It refers to the problems in which it is sought to obtain relationships between the attribute values of a database. The most typical example is to analyze the shopping cart.

1.4.2 Predictive Problems

On the other hand, there are problems of Data Science whose goal is to obtain a model that in the future can be applied to predict behaviors. These types of problems are called predictive or, in Artificial Intelligence environments, they are called supervised learning problems because the analyst provides the system with the desired response. However, once again we can analyze these problems with more attention to observe that the variable to be predicted can be a categorical variable (whether or not to buy a product). However, in the case of loans, the variable to be predicted is the probability of delay in payment, which is a numerical variable.

This distinction in the type of variables that the model predicts leads us to distinguish the predictive problems in:

- *Classification Problems:* They refer to the problems in which the variable to be predicted has a finite number of values, i.e. the variable is categorical. An example of such problems would be to find a model that, in the light of a history of customers classified as “good”, “regular” and “bad”, establishes what type of customer is a new one.
- *Regression Problems:* They refer to the problems in which the variable to be predicted is numerical. As an example, we could have the case of finding a model that establishes the likelihood that a client who is asking for a loan will repay it or not, or the probability that some symptoms are described or may or may not present a disease.

1.5 Classical Data Mining Techniques

The techniques are specific implementations of the algorithms that are used to carry out the operations of construction of the model. Not all algorithms designed to solve a given DM problem are the same and each one will have a certain number of advantages and disadvantages.

The convenience of applying a particular algorithm depends not only on the type of problem we are facing but also to a great extent on the type of data being processed. In this sense, it is convenient to analyze the different approaches and algorithms that exist in the literature, because in real life we find that the publicly available tools offer a whole range of possible algorithms and the end user is who has to decide which one to use. So unless you have a knowledge of these algorithms and an experience in their use, it will be very difficult to find the best solution to a given problem.

The following is a brief list of the DM techniques that can be applied to solve the described Data Science problems.

- **Predictive Models. Classification:** The classical supervised learning is used in these models. Decision trees [26], rule induction [11], instance-based learning [4], logistic regressions [30], SVMs [28] and ANNs [5] are often used. These models use a set of training data to create the model, which is then used to classify unknown individuals.
- **Predictive Models. Regression:** For the prediction of numerical values, linear regression and non-linear regression are used, along with regression versions of the previous methods seen in classification.
- **Standard Clustering:** Here, each data example is compared to all clusters by using a certain distance measure among them with the clusters already created. Then each input data example is assigned to the corresponding cluster. The number of clusters can be either automatically adjusted or not. The K-means algorithm is the best representative technique belonging to this family [3].
- **Hiercharchical Clustering:** This type of DM technique is appropriate when we do not know or have any information about the groups in which the clusters are classified. Hierarchical algorithms such as agglomerative or divisive are often used. ANNs based on non-supervised learning, such as the Kohonen maps, are also used.
- **Analysis of relationships. Associations:** The objective of this technique of DM is to find elements that imply the presence of other elements within the same transaction. The result of this technique are rules of the type "if X then Y". In the rules, X is called the antecedent of the rule and Y is called the consequent. One of the most commonly used association algorithms is Apriori [1]. It is based on counting the occurrences of all possible combinations of elements. What it does is to count the occurrences of all the elements present in the transactions of the database and to create a vector where each of its elements carries an account of an element of the database. Those cells of the vector whose value is below the support level (threshold) are ignored.

- **Analysis of relationships. Sequential patterns:** It tries to discover patterns between transactions in which one set of elements is followed by another set of elements spaced apart for a given period of time [23].
- **Time Series Forecasting:** This technique is intended to discover occurrences or sequences similar to those data that stores information that represents a time series, such as the evolution of market prices or telemetry data from a sensor.

1.6 Non-standard Data Science Problems

Some Data Science problems clearly differ from the standard ones and even some can not even be categorized into one of two possibilities of descriptive or predictive problems. As a result, this section will provide a brief description of other important non-standard problems that are well known and pose a challenge in the Data Science community.

We establish a dichotomous division based on the nature of the Data Science problem. When the problem involves a clear extension on the acquisition or distribution of data, restrictions imposed on the models or the implication of more complex procedures to obtain the adequate knowledge, we refer to a derivative or more restrictive problem. On the other hand, when the problem can only be understood as a mixture of descriptive and predictive problems, we refer to the hybrid paradigm. Note that we only mention some learning paradigms of the universe of possibilities and their interpretations, assuming that this section only intends to introduce the theme.

1.6.1 Derivative Problems

This type of problems are those based on a extension or restriction of the original Data Science problem.

1.6.1.1 Imbalanced Learning

It is an extended supervised learning paradigm, a classification problem where the data has exceptional distribution on the target attribute [8, 16, 20]. This issue occurs when the number of examples representing the class of interest is much lower than that of the other classes. Its presence in many real-world applications has brought along a growth of attention from researchers. This book is thought to give all the insights on this topic and here is not the right moment to give more details. Maybe, the impatient readers could jump to the rest of chapters to get into this exciting field.