1

Optimization and Big Data

The term *Big Data* refers to vast amounts of information that come from different sources. Hence *Big Data* refers not only to this huge data volume but also to the diversity of data types, delivered at various speeds and frequencies. This chapter attempts to provide definitions of *Big Data*, the main challenges induced by this context, and focuses on Big Data analytics.

1.1. Context of Big Data

As depicted in Figure 1.1, the evolution of Google requests on the term "Big Data" has grown exponentially since 2011.

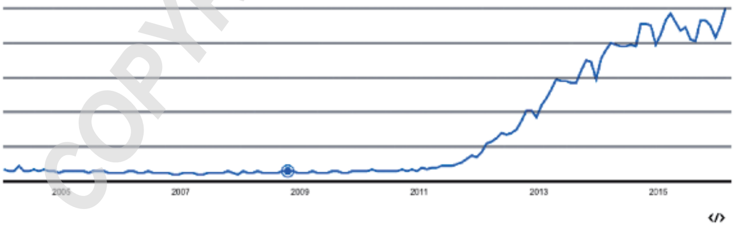

Figure 1.1. *Evolution of Google requests for "Big Data" (Google source)*

How can we explain the increasing interest in this subject? Some responses may be formulated, when we know that everyday 2.5 quintillion bytes of data are generated – such that 90% of the data in the world today have been created in the last two years. These data come from everywhere, depending on the industry and organization: sensors are used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records and cellphone GPS signals, to name but a few [IBM 16b]. Such data are recorded, stored and analyzed.

1.1.1. *Examples of situations*

Big Data appears in a lot of situations where large amounts of complex data are generated. Each situation presents challenges to handle. We may cite some examples of such situations:

– *Social networks*: the quantity of data generated in social networks is huge. Indeed, monthly estimations indicate that 12 billion tweets are sent by about 200 million active users, 4 billion hours of video are watched on YouTube and 30 billion pieces of content are shared on Facebook [IBM 16a]. Moreover, such data are of different formats/types.

– *Traffic management*: in the context of creation of smart cities, the traffic within cities is an important issue. This becomes feasible, as the widespread adoption in recent years of technologies such as smartphones, smartcards and various sensors has made it possible to collect, store and visualize information on urban activities such as people and traffic flows. However, this also represents a huge amount of data collected that need to be managed.

– *Healthcare:* in 2011, the global size of data in healthcare was estimated as 150 exabytes. Such data are unique and difficult to deal with because: 1) data are in multiple places (different source systems in different formats including text as well as images); 2) data are structured and unstructured; 3) data may be inconsistent (they may have different definitions according to the person in charge of filling data); 4) data are complex (it is difficult to identify standard processes); 5) data are subject to regulatory requirement changes [LES 16].

– *Genomic studies:* with the rapid progress of DNA sequencing techniques that now allows us to identify more than 1 million SNPs (genetic variations), large-scale genome-wide association studies (GWAS) have become practical. The aim is to track genetic variations that may, for example, explain genetic susceptibility for a disease. In their analysis on the new challenges induced by these new massive data, Moore *et al.* first indicate the necessity of the development on new biostatistical methods for quality control, imputation and analysis issues [MOO 10]. They also indicate the challenge of recognizing the complexity of the genotype–phenotype relationship that is characterized by significant heterogeneity.

In all these contexts, the term *Big Data* is now become a widely used term. Thus, this term needs to be defined clearly. Hence, some definitions are proposed.

1.1.2. *Definitions*

Many definitions of the term *Big Data* have been proposed. Ward and Baker propose a survey on these definitions [WAR 13]. As a common aspect, all these definitions indicate that size is not the only characteristic.

A historical definition was given by Laney from Meta Group in 2001 [LAN 01]. Indeed, even if he did not mention the term “Big Data”, he identified, mostly for the context of e-commerce, new data management challenges along three dimensions – the three “Vs”: volume, velocity and variety:

– *Data volume:* as illustrated earlier, the number of data created and collected is huge and the growth of information size is exponential. It is estimated that 40 zettabytes (40 trillion gigabytes) will be created by 2020.

– *Data velocity:* data collected from connected devices, websites and sensors require specific data management not only because of real-time analytics needs (analysis of streaming data) but also to deal with data obtained at different speeds.

– *Data variety:* there is a variety of data coming from several types of sources. Dealing simultaneously with such different data is also a difficult challenge.

The former definition has been extended. First, a fourth “V” has been proposed: *veracity*. Indeed, another important challenge is the uncertainty of data. Hence around 1 out of 3 business leaders do not trust the information they use to make decisions [IBM 16a]. In addition, a fifth “V” may also be associated with Big Data: *value*, in a sense that the main interest to deal with data is to produce additional value from information collected [NUN 14].

More recently, following the line of “V” definitions, Laney and colleagues from Gartner [BEY 12] propose the following definition:

> *“Big data” is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making*.

This definition has been reinforced and completed by the work of DeMauro *et al*. who analyzed recent corpus of industry and academic articles [DEM 16]. They found that the main themes of Big Data are: information, technology, methods and impact. They propose a new definition:

> *Big Data is the Information asset characterized by such a high-volume, -velocity and -variety to require specific technology and analytical methods for its transformation into value.*

Even if these definitions of 3Vs, 4Vs or 5Vs are the more widely used to explain to a general public the context of Big Data, some other attempts of definitions have been proposed. The common point of these definitions is to mainly reduce the importance of the size characteristic for the benefit of the complexity one.

For example, in the definition proposed by MIKE2.0 [1], it is indicated that elements of *Big Data* include [MIK 15]:

– the degree of complexity within the dataset;

– the amount of value that can be derived from innovative versus non-innovative analysis techniques;

– the use of longitudinal information supplements the analysis.

They indicate that “big” refers to big complexity rather than big volume. Of course, valuable and complex datasets of this sort naturally tend to grow rapidly and so Big Data quickly becomes truly massive. *Big Data can be very small and not all large datasets are big.* As an example, they consider that the

[1] MIKE2.0 (Method for an Integrated Knowledge Environment) is an open source delivery framework for Enterprise Information Management.

data streaming from a hundred thousand sensors on an aircraft is Big Data. However, the size of the dataset is not as large as might be expected. Even a hundred thousand sensors, each producing an eight byte reading every second, would produce less than 3GB of data in an hour of flying (100,000 sensors × 60 minutes × 60 seconds × 60 bytes).

1.1.3. *Big Data challenges*

Rather, it is a combination of data management technologies that have evolved over time. Big Data enables organizations to store, manage and manipulate vast amounts of data at the right speed and at the right time to get the right insights.

Hence the different steps of the value chain of *Big Data* may be organized in three stages:

1) data generation and acquisition;

2) data storage and management;

3) data analysis.

Each of these stages leads to challenges from the highest importance. Many books and articles are dedicated to this subject (see, for example, [CHE 14, HU 14, JAG 14]).

1.1.3.1. *(Big) Data generation and acquisition*

The generation of data is not a problem anymore, due to the huge number of sources that can generate data. We may cite all kinds of sensors, customer purchasing, astronomical data and text messages. One of the challenges may be to *a priori* identify data that may be interesting to generate. What should be measured? This is directly linked with the analysis that needs to be realized. Much of these data, for example data generated by sensor networks that are highly redundant, can be filtered and compressed by orders of magnitude without compromising our ability to reason about the underlying activity of interest. One challenge is to define these *online* filters in such a way that they do not discard useful information, since the raw data is often too voluminous to even allow the option of storing it all [JAG 14]. On the contrary, generated data may offer a rich context for further analysis (but may lead to very complex ones).

Before being stored, an information extraction process that extracts the required information from the underlying sources and expresses it in a structured form suitable for storage and analysis is required. Indeed, most data sources are notoriously unreliable: sensors can be faulty, humans may provide biased opinions, remote websites might be stale and so on. Understanding and modeling these sources of error is a first step toward developing data cleaning techniques. Unfortunately, much of this is data source and application dependent and is still a technical challenge [JAG 14].

1.1.3.2. *(Big) Data storage and management*

Many companies use one or several relational database management systems to store their data. This allows them to identify what the data stored are and where they are stored. However, these systems are less adapted for a Big Data context and one of the challenges linked to Big Data is the development of efficient technologies to store available data.

These technologies must be able to deal with specificities of Big Data, such as scalability (limitations of the underlying physical infrastructure), variety of data (including unstructured data), velocity of data (taking into account non-synchronous acquisition), etc. Hence, non-relational database technologies, such as NoSQL, have been developed. These technologies do not rely on tables and may be more flexible.

Among these technologies, we may cite:

– key-value pair databases, based on the key-value pair model, where most of the data are stored as strings;

– document databases, a repository for full document-style content. In addition, the structure of the documents and their parts may be provided by JavaScript Object Notation (JSON) and/or Binary JSON (BSON);

– columnar databases or column-oriented database, where data are stored in across rows (e.g. HBase from Apache). This offers great flexibility, performance and scalability, in terms of volume and variety of data;

– graph databases, based on node relationships that have been proposed to deal with highly interconnected data;

– spatial databases that incorporate spatial data. Let us note that spatial data itself is standardized through the efforts of the Open Geospatial Consortium

(OGC), which establishes OpenGIS (geographic information system) and a number of other standards for spatial data.

Big Data management includes data transportation [CHE 14]: transportation of data from data sources to data centers or transportation of data within data centers. For both types of transportation, technical challenges arise:

– efficiency of the physical network infrastructure to deal with the growth of traffic demand (the physical network infrastructure in most regions around the world is constituted by high-volume, high-rate and cost-effective optic fiber transmission systems, but other technologies are under study);

– security of transmission to ensure the property of data as well as its provenance.

These technological challenges related to data acquisition, storage and management are crucial to obtain well-formed available data that may be used for analysis.

1.1.3.3. *(Big) Data analysis*

(Big) Data analysis aims at extracting knowledge from the data. Regarding the knowledge to be extracted, Maimon *et al*. identify three levels of analysis [MAI 07]:

– *Reports:* the simplest level deals with report generation. This may be obtained by descriptive statistics as well as simple database queries.

– *Multi-level analysis:* this requires advanced database organization to make such analysis (OLAP multi-level analysis).

– *Complex analysis:* this is used to discover unknown patterns. This concerns specifically data mining, as it will be defined later, and requires efficient approaches. This book focuses on this level of analysis.

In contrast to traditional data, Big Data varies in terms of volume, variety, velocity, veracity and value. Thus, it becomes difficult to analyze Big Data with traditional data analytics tools that are not designed for them. Developing adequate Big Data analytics techniques may help discover more valuable information. Let us note that *Big Data* brings not only new challenges, but also opportunities – the interconnected Big Data with complex and heterogeneous contents bear new sources of knowledge and insights.

We can observe that while *Big Data* has become a highlighted buzzword over the last few years, *Big Data mining*, i.e. mining from Big Data, has almost immediately followed up as an emerging interrelated research area [CHE 13].

Typically, the aim of data mining is to uncover interesting patterns and relationships hidden in a large volume of raw data. Applying existing data mining algorithms and techniques to real-world problems has recently been running into many challenges. Current data mining techniques and algorithms are not ready to meet the new challenges of Big Data. Mining Big Data requires highly scalable strategies and algorithms, more efficient preprocessing steps such as data filtering and integration, advanced parallel computing environments, and intelligent and effective user interaction.

Hence the goals of Big Data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numerous parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts [CHE 13]. In particular, the need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM) has increased remarkably, parallel to the emergence of powerful parallel and very-large-scale data processing platforms, e.g. Hadoop MapReduce [LAN 15].

In this book, we are mainly interested in this stage of the value chain of *Big Data*, that is to say how can we analyze Big Data and, in particular, how metaheuristics may be used for this. Hence the analysis stage is detailed in the following sections.

1.1.4. *Metaheuristics and Big Data*

A common definition of metaheuristics is:

> *Techniques and methods used for solving various optimization problems, especially large-scale ones.*

By this definition, metaheuristics seem to be good candidates to solve large-scale problems induced by the *Big Data* context.

However, metaheuristics are able to provide the answer not only to the large-scale characteristics, but also to the other ones:

– *Data volume:* metaheuristics are mostly developed for large-scale problems. Moreover, their ability to be parallelized gives opportunities to deal with very large ones.

– *Data velocity:* in a context where data are regularly updated and/or the response must be a real-time one, metaheuristics are any-time methods that may propose a good solution rapidly (even if it is not optimal).

– *Data variety:* working simultaneously with different types of data may be difficult for some standard methods, for example those coming from statistics. Metaheuristics propose encodings that are able to consider several types of data simultaneously. This will give the opportunity to jointly analyze data coming from different sources.

– *Data veracity:* working with uncertainty (or more precisely with unknown data) may also be difficult for classical methods. Metaheuristics can propose to integrate stochastic approaches or partial analysis to be able to extract information from these non-precise data.

– *Data value:* metaheuristics are optimization methods based on an objective function. Hence they enable us to evaluate the interest of the knowledge extracted, its value. Using different objective functions gives the opportunity to express the value in different ways according to the context to which it is applied, for example.

In the context of Big Data, metaheuristics have mostly been used within the data analysis step for solving data mining tasks. However, some of them have been proposed to solve other kinds of optimization problems that are related to the Big Data context.

For example, in the work of Stanimirovic and Miskovic, a problem of exploration of online social networks is studied [STA 13]. The goal is to choose locations for installing some control devices and to assign users to active control devices. Several objective functions are proposed. They formulate the problems as several optimization problems and propose a metaheuristic (a pure evolutionary algorithm; EA) and two hybrid metaheuristics (EA with a local search; EA with a Tabu Search) to solve the identified optimization problems (for more information about metaheuristics, see Chapter 2). Therefore, they define all the necessary components (encodings, operators, etc.). They compare their methods on large-scale problems (up to 20,000 user nodes and 500 potential locations) in terms of quality of the solution produced and time required to obtain a good solution. The results obtained are very convincing.

1.2. Knowledge discovery in Big Data

The relationships between metaheuristics and Big Data are linked strongly to the data analysis step, which consists of extracting knowledge from available data. Hence we will focus on this data mining aspect. This section will first situate the data mining in the whole context of knowledge discovery and then present the main data mining tasks briefly. These tasks will be discussed in detail in the following chapters of the book, as one chapter will be dedicated to each of them. Hence, each chapter will present an optimization point of view of the data mining task concerned and will present how metaheuristics have been used to deal with it.

1.2.1. *Data mining versus knowledge discovery*

Knowledge Discovery in Databases (KDD) has recently seen an explosion of interest in many application domains, thanks to the numerous data available that have to be deeply analyzed not only with simple reporting. KDD is the process of identifying valid, novel, useful and understandable patterns from large datasets. Data mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns (implicit or explicit) – which are the essence of valuable knowledge [MAI 10].

Hence KDD is an inductive (not deductive) process. Its aim is to infer knowledge that is generalized from the data in the database. This process is generally not supported by classical database manager systems.

Knowledge discovery problems raise interesting challenges for several research domains such as statistics, information theory, databases, machine learning, data visualization and also for operations research (OR) and optimization as very large search spaces of solutions have to be explored.

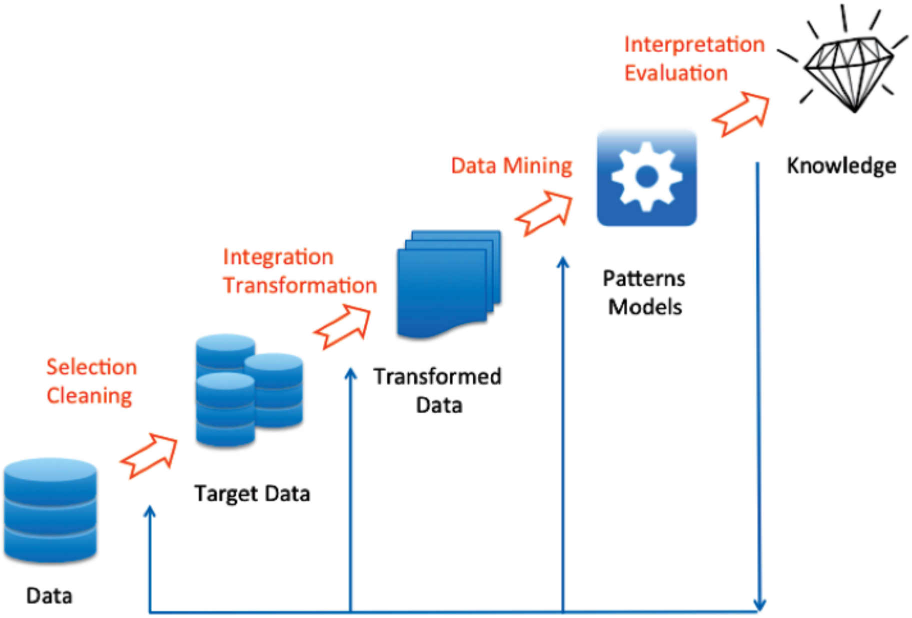

Figure 1.2. *Overview of the KDD process*

Given a context and intended knowledge that has to be extracted (that may be expressed by questions), a KDD project is identified. Then, the whole KDD process starts from raw data and applies different steps to produce knowledge from this data (see Figure 1.2):

– *Selection/cleaning:* starting from raw data, some information is selected to deal with the KDD goals that are identified. Then, the cleaning step may consist of handling missing values and removing noise or outliers, for example. Complex statistical methods as well as data mining algorithms have been proposed for this purpose. This is a crucial step as these data represent the raw material for the following steps.

– *Integration/transformation:* the aim of this second step is to prepare data to be exploited. It may include dimension reduction (feature selection, sampling) and attribute transformation (discretization of numerical attributes and functional transformation). This may also be a crucial step for the success of the KDD project as it is context dependent and is directly linked to the goals of the KDD project.

– *Data mining:* this is the heart of the knowledge discovery process. It allows the extraction of useful information from large datasets or databases. Several data mining tasks may be identified according to the type of patterns expected. Within this data mining step, which can be iterative, an important aspect deals with the evaluation of the extracted patterns. This data mining step is described hereafter.

– *Interpretation/evaluation:* patterns extracted from the data mining step are transformed into knowledge, thanks to interpretation. An evaluation is realized to determine whether the extracted knowledge has a real value (this is a new knowledge) and whether it answers the identified goals. If not, some adjustments have to be done and the process is reiterated either from the beginning or from an intermediate step.

1.2.2. *Main data mining tasks*

Data mining tasks can be classified into two categories: predictive or supervised and descriptive or unsupervised tasks. The supervised tasks learn on available data to make predictions for new data, whereas unsupervised tasks involve a description of the data and existing relationships. Main data mining tasks are (supervised) classification, clustering (also called unsupervised classification), association rule mining and feature selection, as depicted in Figure 1.3. Indeed, even if the feature selection may be used in the integration step, to prepare data, it can also be jointly used with other data mining tasks such as classification or clustering. Hence we decide to consider it within data mining tasks. To give a general overview, each of these tasks is briefly described hereafter. They will be detailed in the chapters dedicated to them.

1.2.2.1. *(Supervised) classification and regression*

The aim is to build a model to predict the unknown value of a target variable from the known values of other variables. In a classification problem, the variable being predicted, called the class, is categorical and the task becomes a regression problem when the predicted variable is quantitative. The model is constructed using available data (available observations), and then for new observations, the model is applied to determine the value of the target variable.

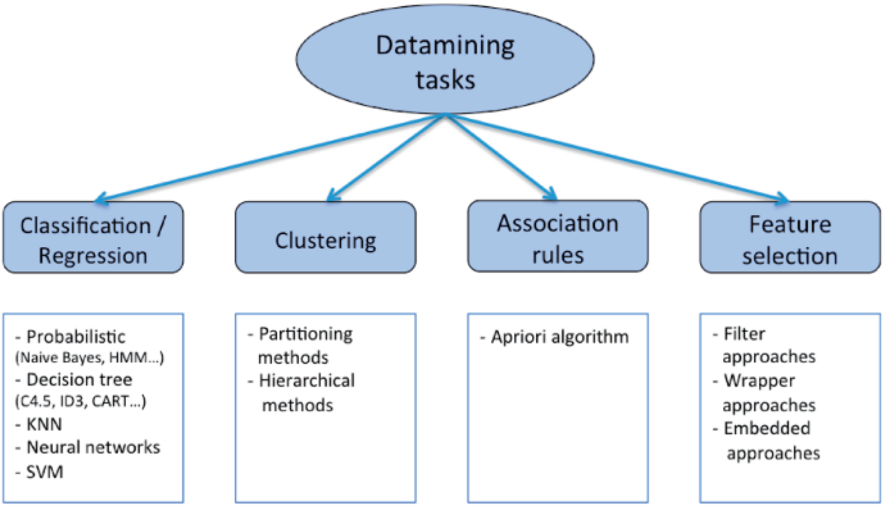

Figure 1.3. *Overview of main tasks and approaches in data mining*

There are numerous applications of classification. It can be used, for example:

– in fraud detection, to determine whether a particular credit card transaction is fraudulent;

– in medical disease diagnosis, to determine whether a patient may pick up a disease in the future;

– in marketing, to identify customers who may be interested in a given product;

– in social network analysis, to predict useful properties of actors in a social network.

Several approaches have been proposed. For an overview on Big Data classification, the reader may refer to [SUT 15]. Among the most widely used approaches, we may cite:

– *Probabilistic classification:* this uses statistical inference to compute a probability of an observation to belong to each of the possible class. The class with the highest probability is selected as the best class. Such an approach allows the computation of confidence value associated with its selected class label. Classical models and algorithms for probabilistic classification include Naive Bayes classifier, logistic regression, HMM, etc.

– *Decision trees* create a hierarchical partitioning of the data using a split criterion. Some of the methods for decision tree construction are C4.5, ID3 and CART.

– The *KNN (K-nearest neighbors)* method associates each new observation with the most represented class within its K-nearest neighbors. This method does not require any learning phase. To deal with data in a Big Data context, a MapReduce-based framework to distribute the partitioning methodology for prototype reduction techniques has been proposed [TRI 15].

– *Neural networks* simulate the human brain and its ability to perform learning.

– *SVM classifiers* use a linear condition to separate the classes from one another. Recently, an implementation of SVM on a quantum computer has been proposed to deal with very large size datasets [REB 13].

1.2.2.2. *Unsupervised classification or clustering*

The clustering task aims at decomposing or partitioning a dataset into groups so that points (here the observations) in a group are similar to each other (distance between points of the groups is minimized) and are as different as possible from points of other groups (distance between points of different groups is maximized). Clustering is directly linked to the definition of a distance between points.

The main classical approaches for clustering are:

– *partitioning methods* partition data into sets so that sets are as homogeneous as possible. The most famous method is k-means;

– *hierarchical methods* either merge or divide clusters to construct homogeneous ones.

Within the clustering task, we may also consider the *biclustering*, co-clustering or two-mode clustering that simultaneously cluster rows and columns of the matrix representing data [MIR 96, KRI 09]. This subspace clustering allows us to treat attributes and objects interchangeably, and to find relationships between elements with regard to these two directions. Different types of biclusters may be required: either biclusters with constant values, or constant values on lines or columns, or biclusters with coherent additive or multiplicative values. These approaches have been widely used in bioinformatics and many algorithms, mainly based on statistics, have been proposed. The complexity of biclustering problems depends on the exact formulation of the problem, but most of them are $\mathcal{NP}$-complete, which limits the use of exact approaches for large-scale problems.

1.2.2.3. *Association rule mining*

The association rules problem was first formulated by Agrawal *et al*. [AGR 93] and was called the market-basket problem. The initial formulation of this problem was: given a set of items and a large collection of sales records – a transaction date and a list of items bought in the transaction – the task is to find relationships between the items contained in the different transactions. Since this first application, many other problems have been studied with association rules that are defined in a more general way. Let us consider a database composed of transactions (records) described according to several – maybe many – attributes (columns). Association rules provide a very simple (but useful) way to present correlations or other relationships among attributes expressed in the form $A \Rightarrow C$, where A is the antecedent part (condition) and C the consequent part (prediction). A and C are sets of attributes that are disjoint. The best-known algorithm to mine association rules was *a priori* proposed by Agrawal and Srikant [AGR 94]. This two-phase algorithm first finds all frequent item sets and then generates high confidence rules from these sets. A lot of improvements of the initial method, as well as efficient implementations (including parallel implementations), have been proposed to enable us to deal with very large databases [BOR 03, YE 06, ZAK 01].

1.2.2.4. *Feature selection*

A difficulty in data mining is linked to the huge size of datasets and the presence of too many attributes. Including all the attributes could lead to a worse model in the classification procedure than if some of them were removed. For example, some attributes could be redundant or unrelated to the predictor variable. Hence the selection of some attributes could be necessary to reduce the computational time of data mining algorithms, to simplify the model obtained to have an accurate discrimination between observations. Then, the objective is to find a subset of p' relevant variables, where $p' << p$.

Therefore, the main goal of feature selection in supervised learning is to find a feature subset that produces higher classification accuracy. On the other hand, in unsupervised learning, feature selection aims to find a good subset of features that forms high-quality clusters for a given number of clusters.

In supervised learning, three approaches exist according to the interaction with the classification procedure:

– *filter approaches* evaluate features according to their characteristics to select (or not) them;

– *wrapper approaches* evaluate the quality of a subset of features using a learning algorithm, for example;

– *embedded approaches* combine the two aforementioned approaches by incorporating in a wrapper approach a deeper interaction between attribute selection and classifier construction.

1.2.3. *Data mining tasks as optimization problems*

As discussed previously, data mining tasks deal with operations such as the affectation of an object to a class, the grouping of objects, the selection of features, etc. All of these problems may be formulated as combinatorial optimization problems. Hence several works using optimization methods to solve data mining problems have been proposed [KAR 06, OLA 06, OLA 08, MEI 10, COR 12].

The context of Big Data makes difficult to solve those problems using exact approaches. Hence metaheuristics will be an interesting answer. In their book, Maimon *et al*. focus on soft computing for knowledge discovery. Although the chapters in that book present various approaches, the majority relate to metaheuristics, particularly evolutionary algorithms and swarm intelligence [MAI 07]. Moreover, Freitas focuses in his book on data mining and knowledge discovery with evolutionary algorithms, which represent one part of metaheuristics [FRE 08, FRE 13]. In particular, this book reveals how evolutionary algorithms may also be used for data preparation, rule discovery (included fuzzy rules) or clustering. Let us remark that one chapter is dedicated to the scaling of such algorithms to deal with large datasets.

1.3. Performance analysis of data mining algorithms

1.3.1. *Context*

There is no consensus on how to define and measure the quality of the extracted knowledge. However, three important properties may be mentioned: an extracted knowledge must be accurate, comprehensible and interesting [FRE 13]. The relative importance of these three aspects is highly dependent on the application context and must be defined at the beginning of the KDD process. In the same manner, the way of measuring each of these properties may also vary from one application context to another.

We consider in this part that the quality measure has been determined, and we focus on the methodology used to evaluate the performance of algorithms (which may be stochastic algorithms and may require several executions) and particularly to make comparisons between each one. Difficulties in the context of data mining are of several types:

1) *learning context:* the quality of the knowledge extracted – for example, the classification model or the clusters constructed – depends on its ability to be used on future unknown data. To evaluate this ability, a specific methodology is used to divide data into data used to learn (training dataset) and data used to evaluate the quality (validation dataset);

2) *supervised/unsupervised context:* in the supervised context, the ideal solution is known and may be used for the evaluation (for example, in a classification context, classes are known and it is possible to evaluate errors made by a classification model). In the unsupervised context, no information is *a priori* known on the knowledge extracted. Hence it may be difficult to evaluate quality as no reference exists. Most of the time indicators to measure the quality are proposed;

3) *specific versus generic:* as explained before, many steps in the KDD process are problem-specific. Hence designing a data mining method that is efficient in several application contexts may be difficult (and often useless). Therefore, regarding the interest of designing specific methods or generic ones, comparisons of algorithms may be realized either on specific datasets or not. When using several types of datasets, it may then be difficult to identify a

method that outperforms the others. Statistical tests must be done to examine the efficiency of the compared approaches, as explained hereafter.

Performance analysis of data mining algorithms is also a difficult task, which may have an impact on the choice of the method to use and in turn the quality of the results obtained. However, let us recall that whatever the quality of the results obtained, ultimately, a decision maker has to interpret the results. For these reasons, no responsible user will cede authority to the computer system. Rather, the decision maker will try to understand and verify the results produced by the data mining method. The data mining method must provide efficient visualization tools to make easy to the decision maker the analysis of results. This is an additional challenge with Big Data due to its complexity [JAG 14].

1.3.2. *Evaluation among one or several dataset(s)*

Regarding whether one or several dataset(s) are used to evaluate the performance of algorithms, some statistical tests have to be conducted to determine whether algorithms reach significant performance differences. Figure 1.4 shows a methodology proposed in [JAC 13a, JAC 13b] that identifies the statistical test to use, according to the number of datasets used for the comparison and the number of algorithms compared. It can be explained as follows.

First, while evaluating algorithms that have been executed several times, producing a set of results, among a single dataset (to identify the best algorithm for it, mostly a real-life one), the statistical test to use will depend on the number of algorithms compared. If only two algorithms are compared, then the Mann–Whitney test [MAN 47] on the set of results may be used. If more than two algorithms are compared, first the Kruskal–Wallis test [KRU 52] is used to determine whether algorithms are equivalent. However, additional tests may be performed to compare algorithms pairwise (the Mann–Whitney test, for example).

While evaluating algorithms among several datasets, it is unlikely that one algorithm will outperform others in all the datasets. In the context of data mining, Demsar proposes recommendations to compare multiple learning algorithms over multiple datasets [DEM 06]. These recommendations present a way to evaluate the general performance of an algorithm over several independent datasets or problems, instead of evaluating the performance on a single one. The methodology proposed by Jacques *et al.* follows these recommendations [JAC 13a, JAC 13b]. It may be described as follows. When the number of datasets is small, the Mann–Whitney test [MAN 47] may still be used. However, as it deals with each dataset/algorithm separately, it will be less efficient when the number of datasets/algorithms increases. In this case, *Friedman* [FRI 37] and *Iman-Davenport* [IMA 80] statistical tests are used to detect differences between multiple algorithms over several datasets. These tests are based on ranks obtained by the algorithms over the different datasets. Then, the average ranks are exploited to graphically draw the results. Finally, pairwise comparison of algorithms is performed using the *Wilcoxon* statistical test [WIL 45] and the *Bergmann and Hommel's* [BER 87] procedure, as recommended by Garcia and Herrera [GAR 08].

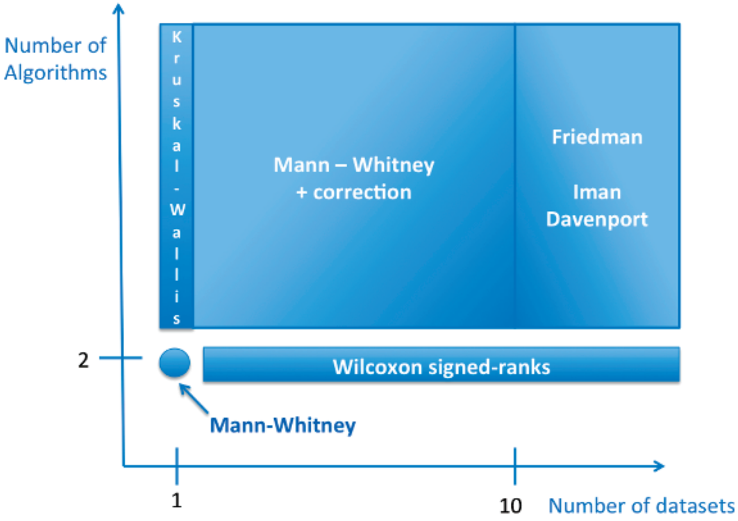

Figure 1.4. *Statistical test summary [JAC 13b]*

Following these recommendations allows us to determine whether significant differences in performance between algorithms may be found. In a Big Data context, where data may have changed over time, it is important to have a general view of the performance of the approaches used. Moreover, while applying optimization approaches and, in particular, metaheuristics that may be stochastic, a rigorous comparison approach is required.

1.3.3. *Repositories and datasets*

To compare performance of data mining approaches, classical datasets that are available on some repositories may be used. Among the most famous ones, we may cite:

– *UCI:* UC Irvine Machine Learning Repository – http://archive.ics.uci.edu/ml/ – that is the historical dataset repository. It provides 348 datasets to the machine learning community. To select datasets to use, several filters may be applied: data mining task, attribute type, data type, application area, size, etc. Let us remark that not many very large size datasets are available.

– *KDnuggets* – http://www.kdnuggets.com/datasets/ – provides links to 1) government and public data sites and portals, 2) data APIs from marketplaces, search engines, etc. and 3) data mining and data science competitions.

– *kaggle* – https://www.kaggle.com – is the leading platform for data prediction competitions and lists current data science competitions.

– *KEEL:* Knowledge Extraction based on Evolutionary Learning – http://sci2s.ugr.es/keel/ – is an open source java software for data mining and provides a dataset repository with more than 900 datasets.

– *RDataMining.com* – http://www.rdatamining.com/resources/data – provides links to other dataset repositories as well as specific tweets data (http://www.rdatamining.com/data).

– *Wikipedia* – https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research – provides datasets cited in peer-reviewed academic journals, ordered by applications.

Usually, the same datasets may be found on several repositories. The format of data may be different, as existing data mining frameworks have their own input formats. This will be discussed in Chapter 8, which is dedicated to frameworks.

1.4. Conclusion

Beyond the buzzword *Big Data*, real significant challenges exist. They are also linked to data acquisition, storage, management and analysis. While focusing on the analysis phase, several data mining tasks may be seen as combinatorial optimization problems and optimization approaches, and, in particular, metaheuristics have been widely used to deal with these problems. Indeed, metaheuristics are good candidates to solve large-scale problems induced by the Big Data context, as well as to deal with other characteristics such as velocity, variety, veracity or value of knowledge.