ePDH online.com

Credit: 1 PDH

Course Title:

Computational Tools for Data Processing in Smart Cities

Approved for Credit in All 50 States

Visit epdhonline.com for state specific information including Ohio's required timing feature.

3 Easy Steps to Complete the Course:

1. Read the Course PDF
2. Purchase the Course Online & Take the Final Exam
3. Print Your Certificate

epdh.com is a division of Certified Training Institute

Chapter 3

Computational Tools for Data Processing in Smart Cities

Danilo Hernane Spatti and
Luisa Helena Bartocci Liboni

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65208

Abstract

Smart Grids provide many benefits for society. Reliability, observability across the energy distribution system and the exchange of information between devices are just some of the features that make Smart Grids so attractive. One of the main products of a Smart Grid is to data. The amount of data available nowadays increases fast and carries several kinds of information. Smart metres allow engineers to perform multiple measurements and analyse such data. For example, information about consumption, power quality and digital protection, among others, can be extracted. However, the main challenge in extracting information from data arises from the data quality. In fact, many sectors of the society can benefit from such data. Hence, this information needs to be properly stored and readily available. In this chapter, we will address the main concepts involving Technology Information, Data Mining, Big Data and clustering for deploying information on Smart Grids.

Keywords: Big Data, Data Mining, clustering

1. Introduction to Big Data concepts

Since the 1970s, companies search for efficient schemes to use their data. In this area of knowledge, concepts of database systems are very common, with the large variety storage suppliers, companies that need large storage can make use of tools that best meet their needs without the premise of getting attached to a particular technology.

Database management systems (DMSs) integrate information systems of different companies, always efficiently making information available to users. In a contemporary context, we have seen a substantial growth of storage needs, particularly in the electric sector.

Also, it is possible to highlight the cheapening of storage technologies as well as the large offer of high-quality, decentralized professional services, such as those supported by cloud computing.

Therefore, with higher storage capacity and with greater demands for storing and registering all possible information for further analysis, the concept of Big Data emerges.

With increased volume of information, it becomes impossible to manage the database through conventional methods. The term Big Data stands for data that require special tools for information extraction since information of interest is immersed in a completely unobservable environment. A very simple example is shown in **Figure 1**.

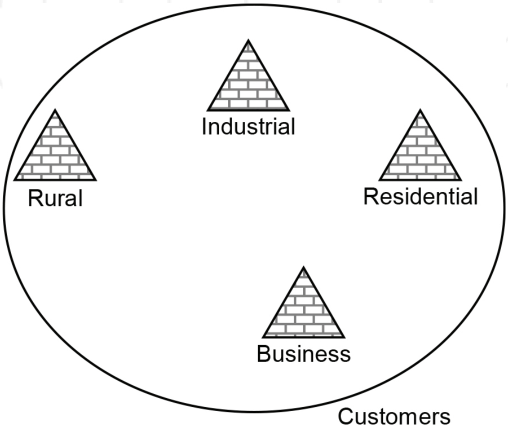

Figure 1. Example of databases.

In a set of customers stored in a naturalistic way, it is not possible to distinguish each of the internal components. In this case, the customer base has become a Big Data. Therefore, it is necessary to search for tools and construction rules that allow the correct distinction between the internal elements.

According to Manyika et al. [1], the implementation of information within the so-called Big Data environments account for a considerable share of the Information Technology market, with potential growth for years to come. To better understand the concept, four basic characteristics associated with Big Data systems can be defined:

1.1. Data volume

Consists in specifying the density of information that a database must support. Notably, we can highlight this as the most important characteristic in the DMS project. It depends on the nature of the information to be stored and also the type of communication channel to be used. The volume of Big Data systems is considered large and with exponential scalability.

1.2. Access speed

A robust DMS, able to store a multitude of data entities, must also possess a compatible access response. The evolution of database technology allowed the suppression of slow access, but this is still a crucial issue when specifying a DMS, which will be used in Big Data.

1.3. Data diversity

Big Data not only relates to pure information stored in tables. In some scenarios, DMS should be able to operate in environments containing several layers of information such as device status and keys, records, oscillography, spreadsheets, text documents, images, videos, etc. This is a prospect in the Smart Grid environment, considering that the integration of information is the basis for its operation.

1.4. Data consistency

A non-reliable database is a problem that emerges from data analysis. The out-of-control growth of the population of entities in a DMS can lead to situations where information quality is jeopardized. In this case, indispensable tools must ensure the integrity and reliability of Big Data databases, considering that operations are not possible with naturalistic observation.

2. Data Mining and attributes selection

2.1. Knowledge database discovery

Knowledge discovery in a database (KDD) is a field in Information Technology that integrates tools for automatically and intelligently analyse large repositories of information. It is an iterative process, whose main task is Data Mining [2]. **Figure 2** shows, in a simplified way, the process of KDD.

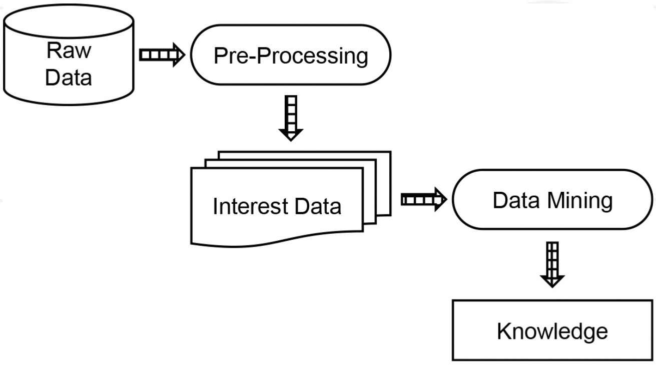

Figure 2. Simplified KDD process.

Data Mining is inserted in this context as one of the stages of KDD, responsible for metrics and methods that will extract the characteristics of interest. We can also highlight the pre-processing of raw data as another key step to reaching information knowledge.

It can be seen in **Figure 2** that the main concept involving Data Mining is to reduce the search space, resulting in a subset of data that is significantly smaller, but that substantially represents the main data set up [3].

KDD can play key roles in various industries, such as searches on financial information for loans, diagnosis for preventive maintenance systems, pattern detection in satellite images, among others [4]. One of the main units handled in KDD processes are the attributes, which are responsible for characterizing information.

2.2. Attribute selection

In Ref. [4] authors define an attribute as "Any instance which is given as input to a machine learning a technique that has fixed values for each of its characteristics". Correlating this concept with databases, attributes are the columns of a DMS, representing the different characteristics of the several instances contained in the database. Instances are often referred records or tuples.

Considering this concept of "attribute", authors of Ref. [5] define the attribute selection process as the determination of a subset of good attributes that will be responsible for generalizing the information contained in the database. The subset of attributes obtained must necessarily have a size equal to or less than the original set of attributes.

Regarding the operation of algorithms for selecting attributes, they usually choose the attributes by an assessment of individual or subsets of attributes. The individual assessment orders the attributes with respect to their relevance. Thus, these methods remove irrelevant attributes, however, do not eliminate redundant attributes. In the case of subset assessments, attributes that are redundant, as well as irrelevant, are removed from the set of attributes [6].

Methods for selecting attributes can be divided into two classes: wrappers [7] and filters [8, 9]. Filters differ from wrappers just on the independence of the algorithm used for the attribute selection [9, 10]. There are a wide diversity of algorithms that perform these tasks. However, only three of these are explained in greater details.

2.2.1. Attribute selection based on wrapper

Wrapper methods are commonly used when one wants to select attributes in supervised learning problems. The methodology consists of the input of a attribute set in which the attributes pass through a predetermined search method and evaluation algorithm. Wrapper methods are called this way, because it essentiality wraps up an evaluation algorithm.

The search method is an algorithm that yields different subset of attributes. There are various methods for searching subsets: forward search, where the initial subset starts with a single attribute and then have attributes inserted; backward search: where the initial subset is formed by all attributes and then have attributes removed: bi-directional search: an initial subset will have attributes removed and inserted; exhaustive search: all the possible subsets are obtained.

After the search for a subset of attributes, the wrapper method calculates a performance estimate by inducing the learning algorithm, which will evaluate only those selected attributes.

The search is terminated after a number of performances do not improvement. This parameter is used to escape from local maxima and seek global maxima.

After all the subsets have passed through the same process, the best subsets are selected and will again pass through the learning algorithm. that will extract the most important attributes within the subset, which are evaluated together with a validation set [7]. Finally, with the selected attributes, a classification can be made using the learning algorithm.

In **Figure 3**, it can be seen a brief presentation that generalizes the wrapper methods, to illustrate the procedures discussed above.

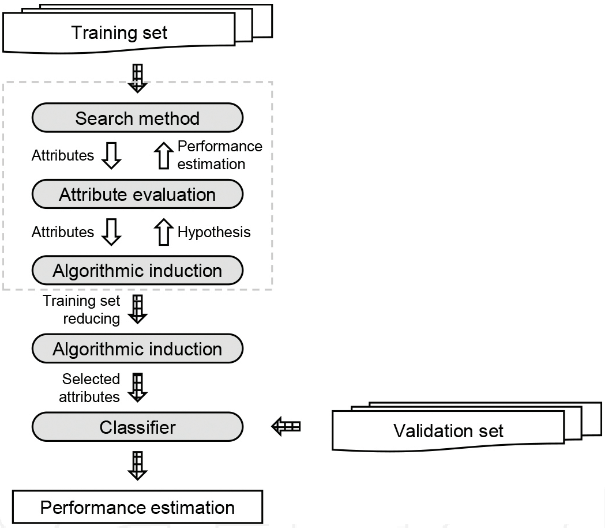

Figure 3. Overall representation of the wrapper attributes selectors.

This algorithm is useful, however, even providing better results than filters; it is slow, i.e., because the learning algorithm is called repeatedly.

2.2.2. Correlation-based feature selection

The correlation-based feature selection (CFS) method has been proposed in Ref. [12], where it can be applied to sets of continuous and discrete data, as shown in Ref. [10]. The method makes use of correlation to estimate the cost of attributes; however, a big difference presented by CFS, when compared to other filters, is that the selection starts by assessing attribute subsets. For each subset, the addition/removal of attributes is then evaluated [11].

The CFS algorithm was proposed according to the following hypothesis: "Good attribute subsets contain attributes highly correlated with the class; however, not greatly correlated to other attributes".

The search for the best subset of attributes ends only when the stopping criterion is satisfied, and this is considered when a predefined number of iterations return the same subset of attributes. Some advantages presented by the CFS algorithm are its rapid implementation, the possibility of application in any set of attributes.

2.2.3. Consistency-based filter

The consistency-based filter (CF) method proposed in Ref. [9] evaluates the attributes of subsets according to their consistency with the classes comprising the data set, and unlike most methods for selecting attributes, it does not use search heuristics, but a probabilistic search algorithm based on Las Vegas algorithm.

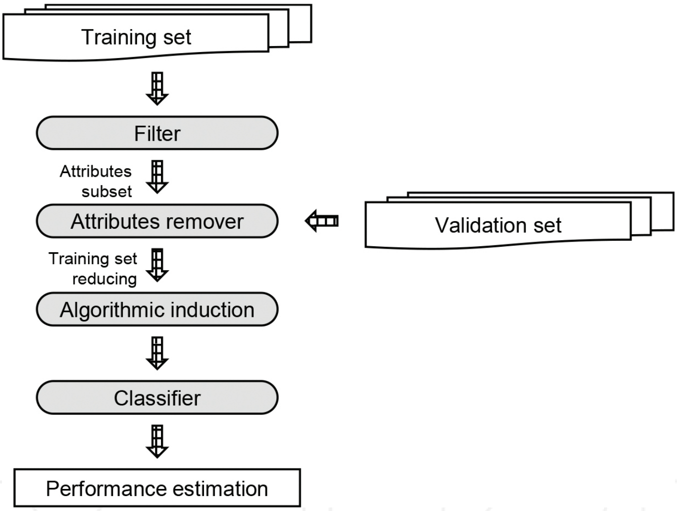

Figure 4. Overall representation of the filter selectors.

From the experimental results obtained by Ref. [9], it can be seen that this method provides a quick response, ensures the location of the optimal attribute subset and is easily implemented.

Las Vegas algorithm makes probabilistic choices to assist the search process. Thus it results quickly in the best attribute subset. This search is performed until a maximum number of attempts is reached. In the end, both the size of the attribute subset and their inconsistency with respect to the class are evaluated. The selected subset is smaller, and consistent since the consistency of an attribute subset is inversely proportional to its inconsistency.

A disadvantage presented by the probabilistic search compared to a heuristic search lies on the computational cost, which is a little higher. However, its biggest advantage is that it does not have the same vulnerability presented by the heuristic search when subjected to data sets with many related attributes.

As an illustration, a block diagram representing the general operation of filters employed in the feature selection task can be seen in **Figure 4**.

2.3. Data Mining

As mentioned before, Data Mining can be considered a step in the KDD [13]. This concept was originated from the need to extract patterns in databases in an attempt to extract valuable information, which could cause companies to become even more competitive within its sector of activity. In **Figure 5**, it can be seen that the Data Mining system is divided into four steps summarized below:

- Step 1 – use of the corporate database or a test database;
- Step 2 – use of techniques for pre-processing data, for example, data normalization (commonly used) and the techniques for selecting attributes of the database;
- Step 3 – from the pre-processed database, Data Mining techniques can then be applied, such as artificial neural networks and fuzzy inference system;
- Step 4 – the response obtained by the Data Mining process, as well as its analysis should be assessed to verify the performance of computational tools employed.

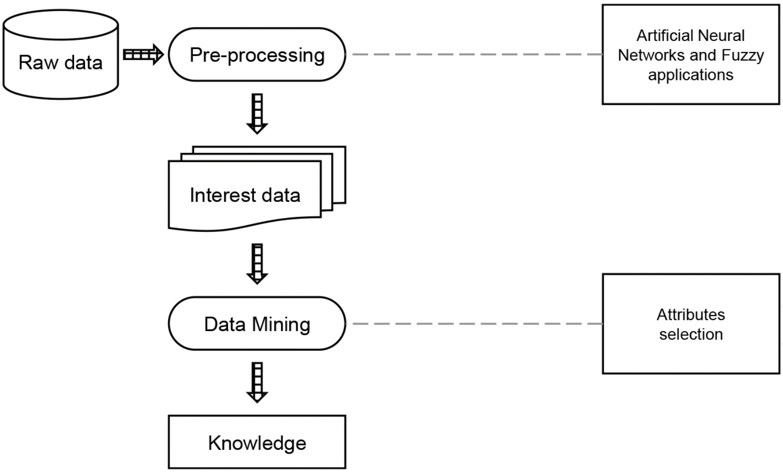

Figure 5. Representation of the KDD process.

Therefore, Data Mining is often used to specify groups of patterns that can be found in databases, without any prior knowledge of which pattern may provide information of interest [13–15].

There are cases where some idea of what kind of information one wants to extract from the database exists. However, the complexity of the extraction task prevents that the analyses are carried out by specialists [14].

The tasks performed by Data Mining are classified as predictive or descriptive [14]. When applying the predictive mining, data representing previous situations of a process are analysed to obtain patterns that may represent the current or future situations of the process.

Since the descriptive mining is used to characterize data contained in the actual database, only current conditions of the process can be determined by this type of mining.

2.4. Data Mining phases

According to Larose [16], Data Mining is considered a cyclical network, because success depends heavily on adjustments made in an ordered set of six phases, totally interrelated, as illustrated in **Figure 6**.

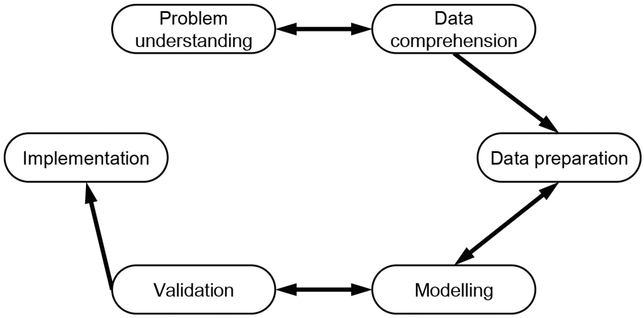

Figure 6. Data mining phases.

It can be seen that the phase of understanding the problem and comprehending the data are bi-directionally connected for information exchange. According to MSDN Microsoft [17], these two phases comprise analysing business requirements, defining the scope of the problem, defining the metrics for evaluating the model and, ultimately, identifying specific data targets for the mining project.

Data can be spread and stored in different formats and may contain inconsistencies, such as missing or incorrect entries. With an appropriate cleaning process and filters, it is possible to fill the empty records. Such actions are carried out during the data preparation phase.

As pointed out in Ref. [16], it is in the modelling phase that algorithms will perform the mining process. The choice of which source to use is dependent on specific application goals.

In the validation phase, models developed in the previous phase will have their performance evaluated as to their performance across a set of data out of the set initially proposed for learning purposes. This stage has a bi-directional connection with the modelling so that fine adjustments can be tuned in mining algorithms.

Finally, after the validating the models, the tools developed in the previous stages can be implemented.

3. Clustering

Currently, tools focused on Data Mining can perform various tasks such as:

- Description provides information that enables the interpretation of the data, therefore, models need to be as transparent as possible to the user. In this case, the quality of information should be the highest possible.
- Estimation provides feature values for the process mapped to the database.
- Prediction allows finding future values of time series from historical information available, even if fitting the function is not possible.
- Classification provides data classification results with respect to previously known classes or categories of data. It is one of the tasks best performed by Data Mining algorithms.
- Association provides relationship rules between attributes, which can have no clearly observable similarities.
- Clustering provides grouping patterns of classes or attributes, which can have clearly observable similarities.

The grouping of standards, records, classes, etc., with similarities, is notably the first approach for managing a mass of unobservable naturalistic data. Clustering is considered as the first task to be performed in order to find common characteristics in databases without knowing the relationship between its entities.

After completing the data assembly process and looking for similarities between instances, further studies involving other tasks associated with Data Mining becomes possible.

Instances spatially distributed without any clearly observable common feature can be clustered by an algorithm that finds elements in common. The algorithms that perform this task are called unsupervised. This is because as this process is performed on data with little or no clear observability, the-the relationship between the entities is unknown.

Most algorithms are based on distance metrics such as the Euclidean distance. These algorithms seek similarities between records through monitoring a parameter that indicates the degree of similarity between the tested standards.

Currently, great efficiency in clustering processes has been achieved with intelligent algorithms, especially with the ones based on Artificial Neural Networks. The Adaptive Resonance Theory (ART) neural network architecture, setup with the recurring features the property to learn new patterns without destroying information learned previously. Therefore, it has a high clustering power and converges very quickly.

4. Database systems

The databases are susceptible to errors caused by a multitude of factors. The data may be incomplete when part of the information is missing, or inconsistent, when the stored information does not match reality [14].

The vast majority of these problems can be exacerbated when the incorrect choice of database, considering the factors listed previously.

Broadly, if it is found that the variety and volume of data currently generated quickly exceed the capabilities of a conventional DMS, the need for tools that can deal with such features is made clear.

We assume that a DMS able to act in Big Data environments must consist of heterogeneous information-storage bases, as exemplified in **Figure 7**.

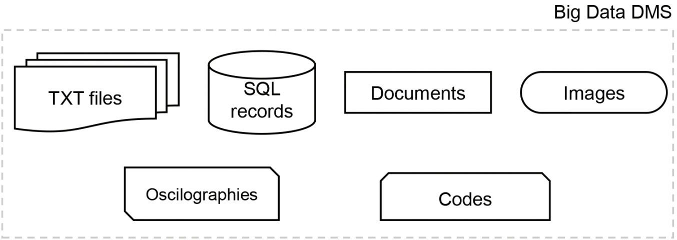

Figure 7. Big Data DMS.

TXT files can contain short reports or even status or device configurations. SQL records can document all kinds of information that can be tabulated, such as customer data, assets, etc. Documents and images have a multitude of applications in companies, which should be properly stored and managed.

Codes of different applications can be stored, either to record update on devices, computers, component firmware, etc. Oscilographies can be responsible not only for protection tasks but also for billing and power quality.

Edgar F. Codd in 1969 proposed a structured model of data storage, which enabled the success of relational models. However, to treat the large volumes, diversity and scale of existing databases, there are new models able to act upon both the structured and unstructured environment.

In this context, it is clear that the use of relational databases for this purpose becomes impractical, giving space to other databases. Among these, we highlight the NoSQL, which is a whole new generation of DMS able to act not only with SQL.

Originally developed to keep up with the web growth in 2009, NoSQL is configured today as a safe and viable option for developers on Big Data as it offers free scheme for development and support, being able to work in environments with massive amounts of data.

We can highlight systems that use the core NoSQL [18]:

- Hadoop/HBase: Written in Java, which allows communication with any protocol.
- MapR, Hortonworks, Cloudera: Professional Hadoop distribution.
- Big Table: Distributed structured data storage system. Used by Google application systems [19].
- Amazon DynamoDB!: Provided by Amazon in its suite of Web Services, being flexible and having consistent latency below 10 ms at any scale [20].
- IBM Informix: Horizontally and vertically scalable relational model from IBM, made of C and contains enhancements to access speed.
- Neo4j: A model based on graphs and represents an evolution of models initially introduced by Facebook and Google.

The functioning of a synthesis Database Manage System (DMS) in Big Data environments does not reflect in the replacement of the traditional models. On the contrary, there should be coexistence between the various forms of information storage that companies already use.

The database mechanism shall function as an integrator of solutions and capable of reliably and quickly supporting high scalable information that is currently generated one of the major works involving the integration of databases can be found in Ref. [21], which contains a study by Harvard Business School.

According to a report conducted in 2010 by the Brazilian Ministry of Mines and Energy, as the related entities, as databases have differently, there is a tendency to adopt models such as XML, for the IEC 61850 defines models for data acquisition based on this language.

5. Data security

5.1. Internal security

Internal security ensures the integrity of the information stored. Every write operation on databases may incur errors, such as conversion errors from A/D converters, syntax error in the symbolic dictionary and physical errors in the recording media.

Ensure that the saved information will be available for future access depends on the technology used but also on the policy of information preservation. It is necessary to ensure user processes efficiently and methodically in data storage or make this process transparent, with a high level of automation. This second option is the one that has been used, and the burden of responsibility transferred to the team that manages the Information Technology.

As the mining process usually requires the integration and transformation of databases, the data internal security depend on several factors, such as [14]:

- Integration schemes: Cohesion of real-world entities;
- Redundancy attributes: Statistical analyses to check correlation between attributes;
- Resolving conflicting data: Differences involving scale, encoding, representation, etc.

Treatment of inconsistencies in the databases can be performed manually, automatically, or with both methods. About 70% of all of the processing time of a DMS is spent searching and correcting errors to ensure the internal security of the information.

5.2. External security

The external security concerns the vulnerability of the information stored in databases. With the introduction of the concept of "Internet of Things", it is assumed that a very large number of IP addresses are available, with the use of IPv6 designation. Indeed the penetration of Smart Grid devices in the daily lives of users is also one of the appeals of these technologies. However, caution is needed with which information will be available to users.

The number of known malicious codes grows day by day, representing a major challenge for the containment team to attack.

The vast majority of databases are currently working with encryption keys, getting the information confined only between devices able to exchange valid keys. Access attempts with invalid keys or even force to obtain valid keys require much computational resource from malicious users. However, this is a real security threat. Mainly because along with the concept of Big Data implemented in the Smart Grid, it is emerging the concept of Smart Cities, where there is the integration of systems and services.

In August 2015, it was held in São Paulo, Brazil, the event "Connected Smart Cities" [22]. One of the topics discussed was the security aspects in the so-called Smart Cities.

Once the idea is the integration of resources, networks, services, etc., such systems need to be separated into physical networks and into different levels of protection.

As communications occur between different entities, there is a targeting component searches for open protocols are adopted, such as based on the TCP/IP family.

NIST considers research in the area of information security still very incipient for the effective implementation of Smart Cities.

The challenges are reflected in the form of encryption key management systems with volumetric applied to Big Data, with also exponential scalability. As the integration will arrive at various levels not previously experienced by companies, secure protection alternatives eventually must face the balance between performance and redundancy.

Systems with great safety tend to be slow and thus provide numerous other types of problems such as overloading of communication channels, increased error rate, idle time processing system tasks. We can list a few critical factors for the protection of these databases:

- Segmentation of networks: An increased presence of schemes that correctly isolates network is fundamental for service integration of services.

- Redundancy: The information, which will travel between the entities, requires authentication mechanisms with low vulnerability to fraud.
- Enforcement of countermeasures: The use of security measures needs to be in accordance with all kinds of communications that will be employed by the company, such as firewalls, authentication certificates for VPN networks and security keys.
- Investment in human assets: Constant training will be needed so that those responsible for information security are always able to work in the services integration environment of Smart Cities.

Author details

Danilo Hernane Spatti[1*] and Luisa Helena Bartocci Liboni[2]

*Address all correspondence to: danilospatti@utfpr.edu.brs

1 Federal Technological University of Paraná, Curitiba, Brazil

2 Federal Institute of Education, Science Technology of São Paulo, University of São Paulo, São Paulo, Brazil

References

[1] Mayer-Schönberger, Viktor; Cukier, Kenneth. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, 2013.

[2] Appel, A. P. (2010). Pre processing and data mining methods for great amount of data and complex networks. PHD Thesis, University of São Paulo.

[3] Consularo, L.A. (2000). Data mining techniques for image analysis. PHD Thesis, University of Sao Paulo.

[4] Witten, I., Frank, E., Data Mining – Practical Machine Learning Tools. Morgan Kaufmann, 2005.

[5] Liu, H., Motoda, H., Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, 1998.

[6] Lee, H.D. (2005). Major attribute selection for knowledge database extraction. PHD Thesis, University of Sao Paulo.

[7] Kohavi, R., John, G.H., “Wrappers for feature subset selection”. Artificial Intelligence, vol. 97, pp. 273–324, 1997.

[8] Almuallim, H., Dietterich, T.G., “Learning with many irrelevant features”. Proc. of the 9th National Conference on Artificial Intelligence, pp. 547–552, 1991.

[9] Liu, H., Setiono, R., “A probabilistic approach to feature selection: a filter solution”. Proc. of the 13th International Conference on Machine Learning, pp. 319–327, 1996.

[10] Hall, M.A., “Correlation-based feature selection for discrete and numeric class machine learning”. Proc. of the 17th International Conference on Machine Learning (ICML), pp. 359–366, 2000.

[11] Hall, M.A., Holmes, G., “Benchmarking attribute selection techniques for discrete class data mining”. IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, pp. 1–16, 2003.

[12] Hall, M.A., Correlation-based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999.

[13] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Advances in Knowledge Discovery and Data Mining. AAAI Press, 1996

[14] Han, J., Kamber, M., Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[15] Oliveira, S.R. (2004). From data to knowledge: Evolution and challenges. Associated Professor Thesis, University of Sao Paulo.

[16] Larose, D.T., Discovering Knowledge in Data: An Introduction to Data Mining. Wiley, 2005.

[17] MSDN Microsoft, “Data Mining Concepts”. https://msdn.microsoft.com/pt-br/library/ms174949(v=SQL.120).aspx

[18] List of NoSQL databases. Accessed in September 2016, available in http://nosql-database.org

[19] Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., Gruber, R.E., Bigtable: A Distributed Storage System for Structured Data. Google Inc, 2006.

[20] Amazon DynamoDB. Accessed in September 2016, available in http://aws.amazon.com/pt/dynamodb/

[21] Davenport, T.H., Harris, J.G., Competing on Analytics: The New Science of Winning. Harvard Business Press, 2007.

[22] Connected Smart Cities. Accessed in September 2016, available in http://www.connectedsmartcities.com.br