Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 51-11 (2018) 820–825

Clarifying Data Analytics Concepts for Industrial Engineering

Laura Cattaneo, Luca Fumagalli, Marco Macchi, Elisa Negri

Department of Management, Economics and Industrial Engineering, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano, Italy - Contact author's email: laura1.cattaneo@polimi.it

Abstract: **In the last decade manufacturing experienced a shift towards digitalization. Cost decrease of sensors, wireless connectivity, and the opportunity to store big amounts of data pushed a process towards a next generation of IT industry. Manufacturing now has the opportunity to gather large quantities of data, coming from different areas, such as product and process design, assembly, material planning, quality control, scheduling, maintenance, fault detection and cover all the product life cycle phases. The extraction of value from data is a new challenge that companies are now experiencing. Therefore, the need for analytical information system is growing, in order to explore datasets and discover useful and often hidden information. Data analytics became a keyword in this context, but sometimes it is not clear how different methods or tools are defined and could be effectively used to analyze data in manufacturing. The paper aims to present and clarify the meaning of terms that are currently and frequently used in the context of analytics. The paper also provides an overview of the data analysis techniques that could be used to extract knowledge from data along the manufacturing process.**

Keywords: Industry 4.0; data analytics; Big Data; data mining; artificial intelligence.

1. INTRODUCTION

In the past decades, a large amount of manufacturing data was collected in database management systems and data warehouses, due to the wide use of distributed control systems. Data come from different areas, such as product and process design, assembly, material planning, quality control, scheduling, maintenance, fault detection and cover all the product life cycle phases.

More recently, many countries announced a new wave of development plans in manufacturing. Among these, Germany proposed the concept of Industry 4.0 (Shrouf et al. 2014; Negri et al. 2017), whose the main goal is to develop smart factories for higher competitiveness and flexibility. In this context "smart" is strictly connected with "information", since the ability to extract information from data is needed in order to make the manufacturing process more intelligent.

At the same time, the technical development of data storages, computational power and analysis algorithms experienced a fast improvement, and research in data analysis for smart manufacturing has become a mandatory trend.

Data analytics wants to analyse raw data in order to discover hidden patterns and relationships among different variables. In a word, the main aim of data analysis is to extract useful information from rough data and transfer it to effective knowledge to improve product and process understanding and to support decisions (Ge et al. 2017).

In this context, some interesting research statements arise and need to be carefully detailed:

- to understand how to capture data along the production process.
- to understand how to analyse data in order to extract useful knowledge about product, production process and product life cycle;
- to understand how to share information among all the different stakeholders, along all the product life cycle.

In order to start along this direction, the first aim of this work is to clarify the meaning of some terms that are currently and frequently used in the context of data analytics.

The second aim is to provide an overview of the machine learning techniques that could be used to extract knowledge from data along the manufacturing process.

Other works in this direction were presented during the last decade. Pham and Afify's work reviewed data mining techniques in the manufacturing domain. (Pham and Afify 2005). Choudhary made a literature review on data mining applications in manufacturing (Choudhary et al. 2009). Harding compiled a survey on data mining systems in different areas of manufacturing, such as manufacturing planning and shop floor control (Harding et al. 2006). Recently, a new wave of works regarding data mining and manufacturing has been pushed under the influence of the "big data" phenomenon. Ge recently published a state of the art of data mining and analytics, reviewing several learning methods, as well as the application in the context of process industry (Ge et al. 2017).

The paper is organized as follows: Chapter 2 illustrates the main concepts in the field of data analytics, trying to clarify the meaning of the different terms that are used in this context. Chapter 3 summarizes the modern manufacturing context and highlights the importance of extracting value from data. Chapter 4 presents the list of the main data mining techniques, and illustrates where these techniques are already developed and used in manufacturing. Chapter 5 ends with some comments and future research directions.

2. BACKGROUND ON MAIN CONCEPTS

This work focuses on statistics and how some of its algorithms are used to investigate datasets of huge dimensions, the so called big data problems. The field of statistics is always in evolution, due to the arising problems that science and industry are bringing to the general attention. With the advent of computers, statistical problems exploded

2405-8963
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2018.08.440

both in size and complexity and some new areas of research were developed. Nowadays it is common to hear about data mining, machine learning and data analytics, even if the real meaning of these emerging terms is not clear enough. This paragraph tries to clear out the differences among the main concepts related to these topics, as shown in Fig. 1.

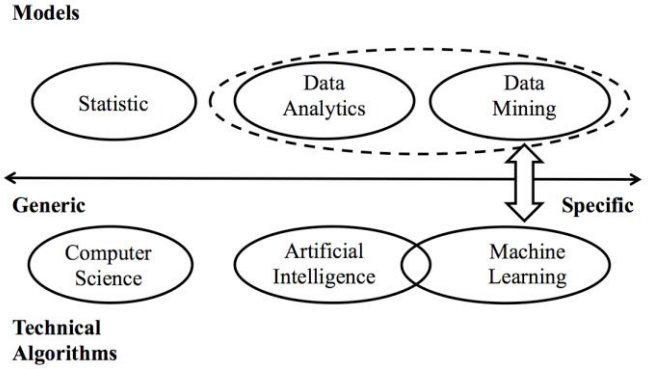

Figure 1: Classification of the main concepts.

Data mining is the extraction of implicit, previously unknown and potentially useful information from data. The idea is to construct numerical algorithms that investigate datasets to extract hidden patterns and regularities. Patterns that reveal strong structures can be used to predict future outcomes. In data mining, the data is stored electronically and the search is automated through computers.

Machine learning is the field that developed most of the algorithms to find and describe structural patterns in data. So, we can say that machine learning represents the *technology instruments* of data mining.

Actually, we can say that a dividing line between machine learning and data mining does not exist because data analysis algorithms lie on a continuum, and there is a direct connection between these two fields, as we can observe from Fig. 1. At the same time, we can conclude that there is not a clear separation between the modelling aspects and the technical algorithms. Some mining algorithms indeed arise from methods taught in standard statistical courses, others are derived from computer science. During the years, very similar methods were developed in parallel in computer science and statistics. At the end, we can say that machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence (AI). In 1959, Arthur Samuel defined machine learning as a "field of study that gives computers the ability to learn without being explicitly programmed" (Simon, 2013).

More recently, the term machine learning has been standardized by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) [ISO/IEC 2382-28:1995; ISO/IEC 2382-31:1997] in such a way: "automatic learning, process by which a functional unit improves its performance by acquiring new knowledge or skills, or by reorganizing existing knowledge or skills".

For the scope of this work, data mining and related machine learning techniques can be grouped into two different sets: the supervised and the unsupervised learning, focusing on two of the four categories proposed by (Ge et al. 2017).

1) *Supervised learning* deals with predicting an outcome value based on the measures of other input variables. In this case, we have a quantitative (such as the pressure in a process) or categorical (such as working/not working) output measure that we want to predict based on a set of features or attributes that we are able to measure (such as energy consumption, flux speed etc.). In this case we need a training dataset, in which we are able to measure and register the outcome and the features for a specific number of entities. Using these training data we are looking for a prediction model, or a learner, which will enable us to predict the outcome for new instances.
2) *Unsupervised learning* deals with description of associations and patterns among a set of input data. In this case, we observe only the features and we are not able to define an output. The final aim is to find hidden patterns, to describe data and to find out how they may be organized or clustered (Hastie et al. 2008).

Data analytics, or equivalently **data analysis,** is a trend word, since it is strictly connected with the buzzword "Big Data". Actually we can say that data analytics is a more comprehensive term respect to data mining, since incorporates a great deal of statistical thinking. In fact, as showed by the doted oval in Fig. 1, data analytics is composed both of data mining (and its technical algorithmic counterpart in machine learning) and statistical models to clean data at the beginning and to validate rules at the end. Data analytics is therefore composed of: initial standard statistical techniques, such as visualization of data, selection of attributes, discarding outliers etc. in order to construct the initial example set; then, machine learning algorithms construct rules and statistical tests are used to validate them (Hastie et al. 2008).

AI has been developed in the field of computer science, as we can see from Fig. 1, and deals with computer programs that possess own decision making capability to solve a problem of interest. AI systems can perform actions such as perception, interpretation, reasoning, learning, communication and decision making to arrive to a solution for a given problem, imitating the intelligent behaviour of human expertise.

ISO/IEC [ISO/IEC 2382-1:1993] have defined AI as the "branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement". An update of this definition was made in 1995, where they state that AI is an "interdisciplinary field, usually regarded as a branch of computer science, dealing with models and systems for the performance of functions generally associated with human intelligence, such as reasoning and learning".

Fig. 1 shows that AI also includes some techniques belonging to machine learning, but is not confined to them: AI deals with pattern recognition, automation, computer vision, virtual reality, diagnosis, image processing, nonlinear control,

robotics, automated reasoning, process planning, intelligent agent and control (Kumar 2017).

Big Data has now become a ubiquitous term in many sectors in industry and academia. Recent works invested time and effort in the proposition and the acceptance of a standard definition of Big Data (Laney 2001, Beyer and Laney 2012; E Zikopoulos and Eaton 2011; Zaslavsky et al. 2013).
It is clear that Big Data refer to a type of datasets, in which variables are different in type and nature, for this reason, they compel to be analysed with a number of specific techniques, models and algorithms, that belong to the Data Analytics, Data Mining and Machine Learning, in this way being directly linked with the main concepts of this paper.
In 2012 Beyer and Laney proposed a quite cited definition of Big Data, known as "the 3-V's" definition, which introduces the framework of the 3-dimensional increase in data: Volume, Velocity, and Variety (Beyer and Laney 2012). Other authors have extended the definition adding further features such as Value, Veracity and Complexity (Schroeck et al. 2012, Dijcks 2013). More recently De Mauro has proposed a consensual definition, which states that "Big Data represents the Information assets characterized by such a High Volume, Variety and Velocity to require specific Technology and Analytical Methods for its transformation into Values" (De Mauro et al. 2015).
Much effort has been devoted to Big Data paradigm and vocabulary clarification also by ISO/ICE. In 2014 they have published a report in which working definition of Big Data is given. "Big Data is a data set(s) with characteristics (e.g. volume, velocity, variety, variability, veracity, etc.) that for a particular problem domain at a given point in time cannot be efficiently processed using current/existing/established/ traditional technologies and techniques in order to extract value", (ISO/IEC JTC 1 2014). More recently, NIST states that "Big Data consists of extensive datasets -primarily in the characteristics of volume, variety, velocity, and/or variability- that require a scalable architecture for efficient storage, manipulation, and analysis" (NIST 2017).

3. EXTRACTION OF VALUE FROM DATA

In the last decade manufacturing has experienced a tremendous shift towards digitalization. Wireless connectivity, sensors cost decrease and the opportunity to store big amount of data fuelled a process towards a next generation of IT industry. Manufacturing has now the possibility to gather a large quantity of data, which span from production status and utilization, to machinery condition monitoring and products quality detection. Manufacturing industry generates about a third of all world data and this is going to increase in the future years (Manyika et al. 2012).
The extraction of value from data is a new challenge that companies are now experiencing. Consequently, the need for analytical information systems is ever-growing to guide corporate decision-making, during the different phases of the product process, such as production, maintenance, logistic and quality control (Flath and Stein 2018). Manufacturing companies need to embrace data analytics to remain competitive in the global market (Lee et al. 2013). As we have already presented in the previous chapters, machine learning techniques represent the instruments to realize predictive models. By incorporating predictive analytics, it is possible to address unobservable problems, such as machine degradation and hidden defects when used in conjunction with manufacturing system data.
A survey presented by Harding and a special issue published on "data mining and applications in engineering design, manufacturing and logistics" (Feng and Kusiak 2006), clearly indicated the potential aim of data mining in these areas. Through data mining it is possible to achieve competitive advantages and the superiority of data mining and machine learning over other experimental techniques lies in the fact that data can be collected during the normal operations of the manufacturing process (Harding et al. 2006).

4. MACHINE LEARNING TECHNIQUES IN SMART MANUFACTURING

The diversity of data mining tools provides great opportunities, but the profusion of different options may cause some confusion. In the following we present a list of the machine learning techniques that are currently used in manufacturing.

4.1 Supervised Learning.

4.1.1 Linear Methods for Regression

Linear models for regression were largely developed in the pre-computer age of statistics, but even today there are still good reasons to study and use them: they are simple and often provide an adequate and interpretable description of how the inputs affect the output. A linear regression model assumes that the regression function is linear in the variable inputs X_1, … X_p. Multivariate linear regression is a generalization of linear regression, by considering more than one dependent variable (Rawlings et al, 1998). In the process industry there are many applications of linear model regression, such as process monitoring, process control and quality prediction (Noorossana et al. 2010, Amiri et al. 2014).

4.1.2 Linear Method for Classification

The linear classifier uses the variable's characteristics to identify which class (or group) the variable belongs to. Classification learning operates with a training example and an actual outcome, the class of the example. The learning scheme is presented as a set of classified examples from which it is expected to learn a way of classifying unseen examples. The antecedent, or precondition, is a series of rules and the consequent, or conclusion, gives the class (or classes) of the example according to the rules defined in the antecedent. The success of the classification learning can be tested on a new independent dataset, for which the true classification is known, but not made available to the machine.

4.1.3 Decision Tree Based Method

A "divide-and-conquer" approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree. Nodes in a decision tree involve testing a particular attribute. Usually, the test at a node compares an attribute value with a constant. However,

some trees compare two attributes with each other, or use some function of one or more attributes. Leaf nodes give a classification that applies to all instances that reach the leaf, or a set of classifications, or a probability distribution over all possible classifications. To classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf. Kwak presented a data mining based production control system for a testing and rework cell in a dynamic computer-integrated manufacturing system. Their system analyzes the present situation and suggests dispatching rules to be followed and evaluates the effect of those decisions. It uses a decision tree based module to generate classification rules (Kwak and Yih 2004). Machine fault diagnosis was investigated and automated using the decision tree method in different works (Jeong et al. 2006, Aydin et al. 2014, Karabadji et al. 2014).

4.1.4 Neural Networks

This method was developed at the same time both in statistics and AI, based on an essentially identical model (White 1989). The central idea is to extract linear combinations of the inputs as new features, and then model the target output as a nonlinear function of these features. By a trial and error method, the network learns the correlations between the input and output pairs and then applies them to predict unknown output for a new set of data (Hastie et al. 2008).

Neural Networks offer a number of advantages, including less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables. Disadvantages include its "black-box" nature, greater computational effort and the empirical nature of model development.

The Artificial Neural Networks (ANNs) are based on the human brain by making the right connections between neurons. It is a flexible mathematical structure and the structure is capable of identifying complex nonlinear relationships between input and output datasets, like other machine methods, ANNs were used to solve a wide variety of tasks (Munirathinam and Ramadoss 2014). In their work the authors highlight how ANNs can develop a predictive method for machine outliers and showed how this helps the overall enhancement of yield in semiconductor manufacturing.

Neural networks were used for manufacturing process monitoring and control by several authors (Braha and Shmilovici 2002, Hou and Huang 2004, Liu and Jolley 2015). Other relevant applications are focused on machine fault diagnosis and fault classification (Hou and Huang 2004, Kusiak and Verman 2012, Nagpal and Brar 2014, Loboda and Robles 2015) and quality control (Ghu 2005, Pian and Zhu 2015).

4.2 Unsupervised Learning

4.2.1. Principal Component Analysis

Principal Components Analysis (PCA) is a quite popular method to transform a high dimensional dataset into a lower dimension dataset. PCA converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. To this end, it finds the n principal axes in the original m-dimensional space where $m>>n$, and guarantees that the variance between the points is the highest. By selecting the axes that explain most of the variance, the number of variables could be reduced from m to n. In this way the core of information is preserved since new variables are combinations of old variables (Hotelling 1933).

Within manufacturing, PCA was used for process monitoring (Frigieri et al. 2017), to check product quality (Kao et al. 2017) and for machine fault detection (Perera et al. 2006, Pimental et al. 2014, Bakdi et al. 2017).

4.2.2 Association Rules

Association rules emerged as a tool for mining commercial databases. The goal is to find joint values of the input variables that appear most frequently in the dataset, i.e. to predict combinations of attributes.

Cunha applied the association rule mining for improving the quality of assembly operations. The computational results showed that the source of assembly faults can be detected using the association rule, even in presence of noise. The extracted associations can be used to improve production quality by avoiding 'risky' sequences (Cunha et al. 2006).

Vinodh used association rules to evaluate the agility in supply chain. They enabled the decision makers to make flexible decisions in the presence of attributes such as flexibility, quality, innovativeness, pro-activity and costs (Vinodh et al. 2011). Kao used association rules to extract features representing relationships between different workstations in order to predict final product quality and to analyse the causes of possible product defects (Kao et al. 2017).

4.2.3 Cluster Analysis

Cluster analysis implements the grouping or segmentation of different objects into subsets or clusters as primary goal, meaning that objects in the same cluster are more similar than other objects assigned to other clusters.

Cluster analysis is used to recognize whether the dataset consists of distinct subgroups, each group representing objects with substantially different properties. Central to the construction of the method is the definition of the degree of similarity (or dissimilarity) between the different objects. So depending on this notion, different procedures could be implemented such as the partitioning method (k-means or k-medoids) and the hierarchical method (Ward's method or single linkage) (Flath and Stein 2018).

Within the industrial decision making, Chen proposed an integrated model by combining k-means clustering, feature selection, and the decision tree method into a single evaluation model to address evaluation of suppliers in the supply chain (Chen et al. 2012). K-means clustering method was developed for machine fault detection (Khediri et al. 2012, Zhou et al. 2014). Zhao used the k-means clustering algorithm for process monitoring (Zhao et al. 2013).

4.2.4 Self-Organizing Maps (SOM)

This method can be viewed as a constrained version of k-means clustering, in which the original high-dimensional

observations can be mapped down onto the two-dimensional coordinate system (Hastie et al 2008). A SOM map consists of components called nodes or neurons. Each node is associated with a weight vector of the same dimension as the input data vectors and a position in the map space. The nature of SOM makes it possible to be utilized for various industrial application, such as dimensionality reduction, data visualization, process monitoring (Ge et al. 2017).
Among the most recent applications, an advanced monitoring platform was developed for industrial wastewater treatment, which is based on the SOM driven multivariate approach (Liukkonen et al. 2013). The SOM method was also used for exploring the information in soil database (Rivera et al. 2014). A SOM-based topological preservation technique was proposed for nonlinear process monitoring (Robertson et al. 2015).

Besides this list of machine learning methods, extensive references could be found in literature, in which data mining techniques are applied in manufacturing (Zhang et al. 2017, Ge et al. 2017).

5. CONCLUSION

Given the importance of extracting value from data, i.e. to send the right information to the right person in the right moment, which is boosted by the digitization trends that are permeating the manufacturing industry context, it is clear that a deeper knowledge of the data analytics tools and concepts is highly desired also in the industrial engineering field.
This paper poses itself as a preliminary work that aims at clarifying the main differences of the data analytics concepts and techniques and, among the latter, at summarizing what are the machine learning techniques that are currently most used in the industrial engineering research and practice.
This is a foundational work for a deeper investigation about the mathematical and statistical tools of data analytics in relation to industry. This will aim at giving industrial engineering researchers the right insights and competences for a wider and more technically robust investigation of how interesting recent trends, such as Industry 4.0 and IoT, could impact the manufacturing sector.

REFERENCES

Aydin, I., Karakose, M., Akin, E. (2014). An approach for automated fault diagnosis based on a fuzzy decision tree and boundary analysis of a reconstructed phase space. *ISA Trans*., 53(2), 220-229.

Amiri, A., Saghaei, A., Mohseni, M., Zarehsaz, Y. (2014). Diagnosis aids in multivariate multiple linear regression profiles monitoring. Commun. Stat.-Theory Methods, 43(14), 3057-3079.

Atzori, L., Iera, A., Morabito, G., (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787–2805.

Bakdi, A., Kouadri, A., Bensmail, A. (2017). Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme, in *Control Engineering Practice*, Volume 66, 64-75.

Beyer, M. A., Laney, D., (2012). The Importance of "Big Data": A Definition. *Gartner Publications*, 1–9.

Braha, D., Shmilovici, A. (2002). Data mining for improving a cleaning process in the semiconductor industry. *IEEE Transactions on Semiconductor Manuf.*, 15(1), 91-101.

Chen, Y.S., Cheng, C.H., Lai, C.J., (2012). Extracting performance rules of suppliers in the manufacturing industry: an empirical study. Journal of Intelligent Manufacturing, 23 (5), 2037-2045.

Choudhary, A.K., Harding, J.A., Tiwari, M.K. (2009). Data mining in manufacturing: a review based on the kind of knowledge. *J Intell Manuf* 20, 501-521.

Cunha, D., Agard, B., Kusiak, A. (2006). Data mining for improvement of product quality. *International Journal of Production Research*. 44(18-19), 4027-4041.

De Mauro, A., Greco, M., Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics, in *AIP Conference Proceeding,* Volume 1644, Number 1, 97-104.

Dijcks, J. (2013), *Oracle: Big data for the enterprise*, Oracle White Paper, Redwood Shores, CA.

Estrin, D., Culler, D., Pister, K. (2002). Connecting the physical world with pervasive networks. *IEEE Pervasive Computing*, 1(1), 59–69.

Flath, C. M., Stein, N., (2018). Towards a data science toolbox for industrial analytics application, in *Computers in Industry* 94, 16–25.

Feng, C.X., Kusiak, A. (2006). Special Issue on data mining and applications in engineering design, manufacturing and logistics. *International Journal of Production Research*, 44(14), 2689-2694.

Frigieri, E. P.,Ynoguti, C. A., Paiva, A. P. (2017). Correlation analysis among audible sound emissions and machining parameters in hardened steel turning, in *Journal of Intelligent Manufacturing*.

Ge, Z., Song, Z., Ding, S.X., Haung, A.B. (2017). Data mining and analytics in the process industry: the role of machine learning. *IEEE Special Section on data-driven monitoring, fault diagnosis and control of cyber-physical systems*, 5, 20590-20616.

Ghu, R.S. (2005). Real Time pattern recognition in statistical process control: A hybrid neural network/decision tree-based approach. *Proceedings of the Institution of Mechanical Engineers. Journal of Engineering Manufacture: Part B*. 219, 283-298.

Kao, H. A., Hsieh, Y-S., Chen, C-H., and Lee, J. (2017). Quality prediction modeling for multistage manufacturing based on classification and association rule mining, in *ICPMMT 2017*, MATEC Web of Conferences 123.

Karabadji, N.E.I., Seridi, H., Khelf, I., Azizi, N., Boulkroune, R. (2014). Improved decision tree construction based on attribute selection and data sampling for fault diagnosis in rotating machines. *Eng. Appl. Artif. Intell.*, 35,. 71-83.

Khediri, I.B., Weihs, C., Limam, M. (2012). Kernel k-means clustering based local support vector domain description fault detection of multimodal processes. *Expert Syst. Appl.* 39(2), 2166-2171.

Kumar, L.S.P. (2017). State of The Art-Intense Review on Artificial Intelligence Systems Application in Process Planning and Manufacturing. *Engineering Applications of Artificial Intelligence* 65, 294–329.

Kusiak, A., Verma, A. (2012). Analyzing bearing faults in wind

turbines: a data mining approach. *Renew. Energy* 48, 110-116.

Kwak, C., Yih, Y. (2004). Data mining approach to production control in the computer integrated testing cell. *IEEE Transactions on Industrial Electronics*, 15(2), 593-603.

Harding, J.A., Shahbaz, M., Srinivas, Kusiak, A. (2006). Data mining in manufacturing: A review. *American Society of Mechanical Engineering, Journal of Manufacturing Science and Engineering,* 128(4), 969-976.

Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, 2nd Edition.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol*. 24 417.

Hou, T.H., Huang, C.C. (2004). Application of fuzzy logic and variable precision rough set approach in a remote monitoring manufacturing process for diagnosis rules induction. *Journal of Intelligent Manuf.*, 15, 395-408.

ISO/IEC JTC 1, (2014) Information technology, Big Data, Preliminary Report.

Laney, D. (2001). 3-D data management:controlling data volume, velocity and variety, *META Group Research Note* , February, pp. 1-4

Lee, J., Lapira, E., Bagheri, B., Kao, H.-a. (2013). Recent advances and trends in predictive manufacturing systems in big data environment, *Manuf. Lett*. 1, 38–41.

Liu, T.-I., Jolley B. (2015). Tool condition monitoring (TCM) using neural networks. *Int. J. Adv. Manuf. Technol*., 78, 9-12.

Liukkonen, M., Laasko, I., Hiltunen, Y. (2013). Advanced monitoring platform for industrial wastewater treatment: Multivariate approach using the self-organizing map. *Environ. Model Softw*. 38, 193-201.

Loboda, I, Robles, M.A.O. (2015). Gas turbine fault diagnosis using probabilistic neural networks. *Int. J. Turbo Jet-Engines*, 32(2), 175-191.

Manyika, J., Sinclair, J., Dobbs, R., Strube, G., Rassey, L., Mischke, J., Remes, J., Roxburgh, C., George, K., O'Halloran, D., Ramaswamy, S. (2012). *Manufacturing the future: The next era of global growth and innovation.* Report, McKinsey Global Institute.

Munirathinam, S., Ramadoss, B. (2014). Big Data Predictive Analtyics for Proactive Semiconductor Equipment Maintenance: A Review. *ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference*, Stanford University.

Nagpal, T., Brar, Y.S. (2014). Artificial neural network approaches for fault classification: comparison and performance. *Neural Comput. Applic*. 25, 1863-1870.

Negri, E., Fumagalli, L. & Macchi, M., 2017. A review of the roles of Digital Twin in CPS-based production systems. *Procedia Manufacturing*, 11(June), pp.939–948.

NIST Special Publications 1500-1 (2017), DRAFT NIST Big Data Interoperability Framework: Volume 1, Definitions. US Department of Commerce.

Noorossana, R., Eyvazian, M., Amiri, A,, Mahmoud, M. (2010). Statistical monitoring of multivariate multiple linear regression profiles in phase I with calibration application. *Quality Rel. Eng. Int.*, 26(3), 291-303.

Perera, A., Papamichail, N., Barsan, N., Weimar, U., Marco, S. (2006). On-line novelty detection by recursive dynamic principal components analysis and gas sensro arrays under drift condition. IEEE Sensors J. 6(3), 770-783.

Pham, D.T., Afify, A.A. (2005). Machine learnign techniques and their application in manufacturing. *Proceedings of the Institution of Mechanical Engineers. Journal of Engineering Manufacture: Part B*. 219, 395-412.

Pian, J., Zhu, Y. (2015). A hybrid soft sensor for measuring hot-rolled strip temperature in the laminar cooling process. *Neurocomputing*, 169, 457-465.

Pimental, M.A.F., Clifton, D.A., Clifton, L., Tarassenko, L. (2014). A review of novelty detection. Signal Process, 99, 215-249.

Rawlings, J. O., Puntula, S. G., Dickey, D., (1998). *Applied Regression Analysis (Springer Texts in Statistics),* NY, USA, Springer-Verlag.

Rivera, D., Sandoval, M., Godoy, A. (2014). Exploring soil databases: A self-organizing map approach. Soil Use Manage, 31(1), 121-131.

Robertson, G., Thomas, M.C., Romagnoli, J.A. (2015). Topological preservation techniques for nonlinear process monitoring. *Comput. Chem. Eng*. 76, 1-16.

Schroeck, M. , Shockley, R. , Smart, J. , Romero-Morales, D. and Tufano, P. (2012), *Analytics: The Real-World Use of Big Data* , IBM Institute for Business Value, Said Business School, New York, NY.

Shrouf, F., Ordieres, J. & Miragliotta, G., 2014. Smart Factories in Industry 4.0: A Review of the Concept and of Energy Management Approached in Production Based on the Internet of Things Paradigm. In *Industrial Engineering and Engineering Management (IEEM), 2014 IEEE International Conference on*. pp. 697–701.

Simon, P. (2013). *Too big to ignore: the business case for big data*. Hoboken, NJ, USA. Wiley.

Vinodh, S., Prakash, N.H., Selvan K.E., (2011). Evaluation of agility in supply chains using fuzzy association rules mining. *International Journal of Product Research*, 49 (22), 6651-6661.

White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective, *Neural Computation,* 1(4), 425-464.

Zaslavsky, A., Perera, C., Georgakopoulos, D. (2013). Sensing as a service and big data. *arXiv preprint.*

Zhang, Y., Ren, S., Yang L., Si, S., (2017). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *Journal of Cleaner Production*, 142, 626-641.

Zhao, X., Li, w., Zhou, L. Song, G.-B., Ba, Q.m Ou, J. (2013). Active thermometry based DS18B20 temperature sensor network for offsgre pipeline scour monitoring using k-means clustering algorithm. *Int. J. Distrib. Sensor Netw*.

Zhou, J, Guo, A., Celler, B., Su, S. (2014). Fault detection and identification spanning multiple process by integrating PCA with neural network. *Appl. Soft Comput,* 14, 4-11.

Zikopoulos, P. and Eaton, C. (2011), *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* , McGraw-Hill Osborne Media, NY.