Computers in Industry 94 (2018) 16–25

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Towards a data science toolbox for industrial analytics applications

Christoph M. Flath*, Nikolai Stein

Julius-Maximilians-University, Würzburg, Germany

ARTICLE INFO

Article history:
Received 1 June 2017
Accepted 8 September 2017
Available online xxx

Keywords:
Predictive analytics
Manufacturing
Process mining

ABSTRACT

Manufacturing companies today have access to a vast number of data sources providing gigantic amounts of process and status data. Consequently, the need for analytical information systems is ever-growing to guide corporate decision-making. However, decision-makers in production environments are still very much focused on static, explanatory modeling provided by business intelligence suites instead of embracing the opportunities offered by predictive analytics. We develop a data science toolbox for manufacturing prediction tasks to bridge the gap between machine learning research and concrete practical needs. We provide guidelines and best practices for modeling, feature engineering and interpretation. To this end, we leverage tools from business information systems as well as machine learning. We illustrate the usage of this toolbox by means of a real-world manufacturing defect prediction case study. Thereby, we seek to enhance the understanding of predictive modeling. In particular, we want to emphasize that simply dumping data into "smart" algorithms is not the silver bullet. Instead, constant refinement and consolidation are required to improve the predictive power of a business analytics solution.

1. Introduction

In the last decade, the manufacturing sector has seen a tremendous digital transformation. Wireless connectivity as well as cost decreases for sensors and data storage have paved the way towards a next-generation industrial infrastructure. In particular there has been a considerable convergence of industrial IT systems and shopfloor automation (see Fig. 1). Going forward, ubiquitous IT on the shop-floor will be instantiated by self-monitoring production equipment and networked production systems [1]. Unsurprisingly, manufacturing companies today have access to a vast number of data sources providing gigantic amounts of process and status data. Manyika et al. [2] estimate that the manufacturing sector generated more than two exabytes of data in 2010. This data ranges from production status and utilization data to continuous tool and machinery condition monitoring. Yet, creating ever-growing data dumps will not contribute to business value generation. However, if appropriately managed data can be a highly valuable resource that is becoming more and more critical to worldwide business operations. This has led to widespread agreement that *data is the new oil* in future IT-augmented systems [3].

In turn, companies are hard-pressed to establish novel analytics tools and use cases to benefit from their data treasures. Leveraging this data by means of new analytics tools offers opportunities to foster data-driven decision-making and increase both efficiency and effectiveness of existing business processes. Such approaches have been discussed in both academic and practitioner literature [5,6]. Revisiting the new oil analogy, an analytics solution resembles an oil refinery which turns basic resource into a valuable products .[1]

While a plethora of IT consultants has been courting companies to buy into the "Big Data Revolution", companies are often disappointed by the outcomes and overwhelmed by the amount and variety of data [7,8]. For now, the promise of industrial analytics mostly remains a mixture of promises, visions and pilot projects instead of large-scale implementations. To become a indispensable part of the manufacturing engineer's toolbox, it still has a long way to go. The recent influx of machine learning research has brought forward a host of capable algorithms and tools but has not equipped operators and decision makers with the necessary work-flows and tools. Consequently, there is an urgent need for tool-kits and templates which assist manufacturing decision-makers navigate through a world of new opportunities.

* Corresponding author.
E-mail addresses: christoph.flath@uni-wuerzburg.de (C.M. Flath), nikolai.stein@uni-wuerzburg.de (N. Stein).

[1] In a recent IoT Analytics study 15% of respondents consider industrial data analytics as a crucial success factor today. Additionally, 69% think it will be crucial in 5 year's time.

https://doi.org/10.1016/j.compind.2017.09.003
0166-3615/

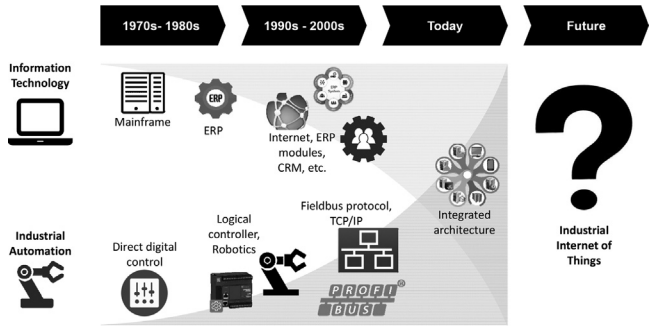

Fig. 1. Convergence of industrial IT systems and shopfloor automation. Adapted from [4].

This paper seeks to address this gap by compiling and explicating a data science toolbox for prediction task in manufacturing systems. We highlight key data preparation and analysis steps. In particular, we combine methods from machine learning and business information systems to guide the development of predictive analytics solutions. We subsequently apply the toolbox to a case study from a major manufacturing company. Thereby, we illustrate how a predictive analytics solution can be set up, refined and evaluated. Prediction tasks in other manufacturing settings will face very similar challenges. Therefore, we are confident that these research questions and our results can be generalized and applied beyond the specific case at hand.

2. Related work and preliminaries

Data ubiquity due to the integration of networked machines as well as the rise of machine learning algorithms lead to a transformational change throughout all major industries. Recent research conducted by General Electric, Accenture [9] estimate that the Industrial Internet offers a $15 trillion opportunity due to reduced costs, productivity gains and new products. They show that the need to leverage the potential of the available data-sets is of high urgency in the manufacturing sector. However, modern manufacturing environments are characterized by large amounts of sensors leading to data-sets that are complex in terms of volume and variety. In the upcoming section, we show the different techniques that are available to tackle these problems.

2.1. Data science in manufacturing

With the rise of data ubiquity, the desire to generate insights and business value from this data is ever-growing. Hence, the idea of “business analytics” describing “data science” in a business context [6] has experienced a rapid growth over the last years.

Shmueli and Koppius [10] carve out the difference between explanatory statistical modeling and predictive modeling. They emphasize that explanatory power derived from traditional models does not imply predictive power. Consequently, predictive analytics is needed not only to create models for practical applications but also for theory building and theory testing. Manufacturing companies need to embrace business analytics in order to remain competitive in the global marketplace [11]. Historically, manufacturing firms have relied on observable process outcomes through shop-floor initiatives like standardized work or continuous improvement. By incorporating advanced analytics they can also address unobservable problems like machine degradation or hidden defects.

Recent research regarding machine learning applications for manufacturing mainly focuses on technical solutions that are used to identify relevant information from large data-sets with many variables. To predict the level of machine degradation, Mosallam et al. [12] apply unsupervised learning to select meaningful variables from a set of monitoring data. The authors report good results in a turbofan engine as well as a battery health setting. Sipos et al. [13] design an information system using multiple linear classifiers to predict failures of medical equipment based on log data. Bleakie and Djurdjanovic [14] propose a method that is capable of predicting system condition by comparing the similarity of recent sensor readings with known degradation patterns. They successfully apply this method in a semiconductor manufacturing setting.

To this end, the existing research mainly provides solutions for specific problems in case study settings. Hence, the goal of this paper is to provide a toolbox for the implementation of data driven approaches in various manufacturing settings.

2.2. Machine learning

The algorithms behind predictive manufacturing applications can be assigned to the field of data mining. Unlike “normal” algorithms it is the data that tells these data-driven algorithms what the good answer is. In a manufacturing setting, a traditional approach would try to define a set of variables (e.g., weight and form) that identifies defective parts. In contrast, a machine learning algorithm does not need such coded rules but would learn them by examples. These learning techniques can be either unsupervised or supervised. In unsupervised machine learning, the observations have no “labels.” Hence, an algorithms is used to identify hidden patterns in the input variables. In contrast, supervised learning is the task of inferring a function from labeled training data. In supervised learning, each example is a pair consisting of an input object (in most cases a vector) and a output value. Problems with a continuous output space are summarized under the term regression problems while classification describes problems with a discrete output space.

2.2.1. Unsupervised learning

Unsupervised learning summarizes machine learning algorithms that find hidden structures in unlabeled data [15]. While there are many possible applications in different fields (e.g., association rule mining for recommender systems, generative adversarial networks for image generation [16]) we focus particular on the algorithms used for dimensionality reduction as they are of special interest in manufacturing settings with increasing amounts of high dimensional monitoring data.

Principle components analysis (PCA) is a popular and well studied method to transform high-dimensional data-sets into low-dimensional data-sets. 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 the variance between the points is the highest. By selecting the axes that explain most of the variance, the number of variables is reduced from m to n. Thereby, the bulk of information is preserved as the new variables are combinations of the old variables [17]. However, PCA reaches its limitations if the relationships between the variables are non-linear. This shortcoming is tackled by the recently developed method t-distributed stochastic neighbor embedding (t-SNE). This technique takes a set of points in a high-dimensional space and embeds them in a lower-dimensional space by solving a problem known as the crowding problem [18]. Due to its flexibility, t-SNE is often able to find structures in data-sets where other dimensionality-reduction algorithms fail. However, this advantage comes at the costs of a decreased interpretability as well as the need for a complex hyper-parameter tuning [19]. The selection of the best dimensional reduction algorithms is typically a trial and error