Big Data Analysis of Manufacturing Processes

Journal of Physics: Conference Series **659** (2015) 012055 doi:10.1088/1742-6596/659/1/012055

Big Data Analysis of Manufacturing Processes

Stefan Windmann[1], Alexander Maier[1], Oliver Niggemann[1], Christian Frey[2], Ansgar Bernardi[3], Ying Gu[3], Holger Pfrommer[4], Thilo Steckel[5], Michael Krüger[6], Robert Kraus[7]

[1]Fraunhofer Application Center Industrial Automation (IOSB-INA), Lemgo
[2]Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB), Karlsruhe
[3]DFKI GmbH, Multimedia Analysis and Data Mining, Kaiserslautern
[4]Hilscher Gesellschaft für Systemautomation mbH, Hattersheim
[5]CLAAS E-Systems KGaA mbH & Co KG, Gütersloh
[6]Karl Tönsmeier Entsorgungswirtschaft GmbH & Co. KG, Porta Westfalica
[7]Bayer Technology Services GmbH, Leverkusen

E-mail: `stefan.windmann@iosb-ina.fraunhofer.de`

Abstract. The high complexity of manufacturing processes and the continuously growing amount of data lead to excessive demands on the users with respect to process monitoring, data analysis and fault detection. For these reasons, problems and faults are often detected too late, maintenance intervals are chosen too short and optimization potential for higher output and increased energy efficiency is not sufficiently used. A possibility to cope with these challenges is the development of self-learning assistance systems, which identify relevant relationships by observation of complex manufacturing processes so that failures, anomalies and need for optimization are automatically detected. The assistance system developed in the present work accomplishes data acquisition, process monitoring and anomaly detection in industrial and agricultural processes. The assistance system is evaluated in three application cases: Large distillation columns, agricultural harvesting processes and large-scale sorting plants. In this paper, the developed infrastructures for data acquisition in these application cases are described as well as the developed algorithms and initial evaluation results.

1. Introduction

Due to increasing demands, manufacturing processes are getting more and more complex [1]. For this reason, users often fail to get an overview of the current system status since data is spread over different computer subsystems or hierarchies. The increasing complexity of manufacturing processes and the permanently growing amount of data lead to an overload of the user with respect to process monitoring, data analysis and fault detection. Therefore, problems and failures are often detected too late, maintenance intervals are not chosen correctly and optimization potential with respect to throughput and energy efficiency is not sufficiently exploited.

The following problems with respect to data acquisition, process monitoring and anomaly detection have to be solved:

- Data acquisition: Data in industrial and agricultural applications is stored in a distributed way, without sufficient time synchronization and without any definition of semantics.
- Process monitoring and anomaly detection: The amount of data and real-time requirements render a manual analysis, which is e.g. based on displayed signals, impossible, even if the

data is already integrated, time-synchronized and augmented with semantics. For these problems, algorithmic solutions exist. However, such approaches are not frequently used in industrial practice because they are not yet suitable for the required processing of big amounts of data.

In the present work, an intelligent assistance system is developed to cope with these challenges (see Fig. 1). The proposed assistance system allows to analyze the available information and to

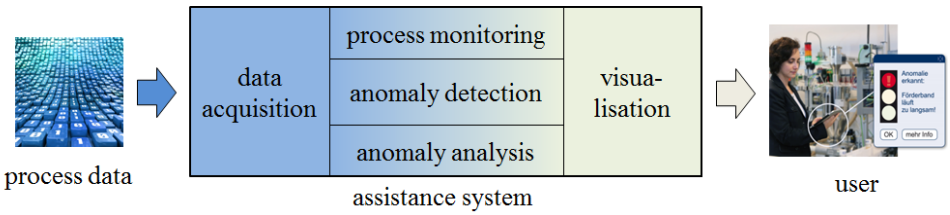

Figure 1. Data analysis of manufacturing processes

detect anomalies. Integration of the assistance system into different manufacturing systems is achieved by suitable data acquisition approaches and flexible methods for process monitoring, which can be adapted with model-learning approaches to changing process behavior. One contribution of the present work is the development of data acquisition approaches for complex distributed manufacturing processes and the integration of big data from several heterogeneous data sources. Furthermore, approaches to process monitoring and anomaly detection in complex manufacturing systems have been developed.

The developed approaches have been evaluated in three application cases, which cover a wide range of agricultural and industrial processes:

- Production processes in distillation columns.
- Anomaly detection in agricultural harvesting processes
- Optimal process control of large-scale sorting plants

The remaining part of this paper is structured as follows: The application cases, which have been investigated in this work, are outlined in section 2. State of the art with respect to data acquisition and anomaly detection in complex industrial and agricultural processes is outlined in section 3. Data acquisition in the described application cases is considered in section 4. The investigated methods for process monitoring and anomaly detection are outlined and evaluated in section 5. Finally, a conclusion and an outlook are given in section 6.

2. Application Cases

2.1. Anomaly detection in complex processes of chemical industry

Large scale continuous production processes in chemical industry typically manufacture, produce, or process materials in continuous operation with infrequent maintenance shutdowns. The continuous diagnosis of the functionality, the early detection of potential failures and the monitoring of the underlying physical process itself is essential for the cost-effective operation of such production processes. Chemical production processes are typically dominated by a complex system behavior with a high system order, e.g. industrial column processes in chemical industry can have up to 120 state variables. In this work, a large scale distillation process is considered (see Fig. 2 a)). This process is typical for chemical industry and is characterized by a high complexity and a large number of measured process states (temperatures, pump speed, pressures, etc.). Efficient plant operation requires an early detection of failures such as sensor

failures or blockages of valves, pumps, etc. in order to prevent damages of the plant or production losses.

Figure 2. Application cases: a) Distillation tower b) Agricultural harvester c) Sorting plant

2.2. Process control for mobile agricultural harvesters
The productivity of agricultural harvesters (see Fig. 2 b)) depends on extremely heterogeneous and dynamic conditions such as conditions of field and harvested crop, adjustment of machine parameters and manual skills of the machine operator. For these reasons, many agricultural processes are unstable and low-deterministic. Comparison of several harvesters in the same region and with similar constraints shows performance deviations of more than 20% (tons of crop in a given time unit) between the particular machines. Modern agricultural harvesters provide extensive data streams with a multitude of sensor information with high acquisition rates. Automatic anomaly detection and analysis of performance deviations by evaluation of continuous data streams allow for economic operation of the harvesters in this application case.

2.3. Monitoring of large-scaled sorting plants
The company Tönsmeier operates a lightweight package sorting plant in Porta Westfalica (see Fig. 2 c)) where secondary raw material is extracted from garbage collections. Optimal process control of sorting plants with respect to the quality of the extracted raw materials and the energy consumption of the plant as well as the reduction of downtime of the plant require immediate detection and repair of anomalies. Typical anomalies or causes for anomalies in this application case are sensor failures, decreased pressure in separation units, material jams or too slow operation of spiral conveyors in the sorting plant.

3. State of the art

3.1. Data acquisition
The large amount of data in industrial processes is acquired at several layers of the automation pyramid. At field and control level, IEC 61131-3 function blocks are usually applied for communication (e.g. TCP/IP or SQL) [2]. In this layer, data loggers can be installed in networks to extract sensor signals and signals of actuators (see e.g. [3]). Agricultural harvesters are equipped in most cases with CAN 2.0 networks and GSM-based loggers, which send the acquired data web servers (see [4]). In distributed automation systems, Supervisory Control and Data Acquisition (SCADA) systems employ a multitude of solutions for data acquisition, which are in many cases based on web services (see [5] for an overview). Furthermore, real-time middle-ware such as CORBAe [6] and OSA+ [7] exists. An overview on agent-based control and computer-based manufacturing technology (HMS) is available in [8]. Widespread protocols at the layer of Manufacturing Execution Systems (MES) are Object Linking and Embedding for Process Control (OPC DA, OPC HDA etc.) [9] and the OPC Unified Architecture (OPC UA) [10]. The acquisition of energy data requires either a real-time bus system or technologies such as

ProfiEnergy [11]. Data is described with information models such as CAEX [12] or CIM [13]. The combination of the described methods and protocols leads in industrial practice to heterogeneous networks with industry-sector-specific structures. So far, no unified data acquisition approach exists, which allows for universal application of the considered data mining methods.

3.2. Anomaly detection

Generally speaking, two classes of algorithmic approaches exist for the detection of anomalous situations: model-based approaches and data-driven approaches.

Model-based approaches employ a model to simulate the normal behavior of the manufacturing process. If the actual measurements vary significantly from the simulation results, the behavior is classified as anomalous (see e.g. [14], [15]). Self-learning of process models from data leads to considerable simplification of model creation and configuration. Industrial and agricultural systems are in general hybrid systems, which are composed of both discrete and continuous system parts. Several algorithms for the self-learning of automata for discrete system parts exist, e.g. RTI+ [16], BUTLA [17] and OTALA [18]. Besides, a lot of research is conducted in the related field of model-learning for continuous systems and application to anomaly detection. Clustering-based methods create groups of strongly related objects and find objects which do not strongly belong to any cluster [19]. Self-organising feature maps (SOM) are capable of generating a topology-preserving mapping of a high-dimensional feature space into an output space of a lower dimensionality [20], which can be utilized as an anomaly detector. By applying the so-called UMatrix representation to a self-organizing map, it is possible to perform a classification or a clustering of the feature space [21] where the clusters correspond to typical process phases. Neural networks and regression-based methods have been used to approximate the functional dependency between continuous process variables [22]. Sensor signals are predicted according to this functional dependency and significant deviations of predicted signal values from the observations are reported as anomalies. Statistical approaches to anomaly detection are predominantly based on building a probability distribution model and considering how likely objects are under that model [23]. In most of these approaches, state variables are employed for modeling the temporal transitions of hidden process variables, which are related to the measurements with a measurement model. Common approaches are Kalman filters (e.g. [24], [25], [26]) and particle filters [27].

Data-driven anomaly detection methods have been receiving considerable attention in recent years as they do not require an analytical model, depending only on the measurable process data [28]. Examples are principal component analysis (PCA), partial least squares (PLS) regression, subspace-aided approach (SAP), etc. [29]. Furthermore, a variety of unsupervised statistical and proximity based methods for the direct detection of anomalies in raw data has been developed in recent years, e.g. k-NN anomaly detectors [30], LOF [31], LOCI and aLOCI [32]. By means of these methods, additional anomalies that are currently not obvious from domain knowledge or that cannot be discovered using model-based approaches, are detected. Fast algorithms have been developed for anomaly detection in large scale data sets, such as HBOS, LDCOF and FastLOF (see [33] for an overview).

4. Data Acquisition

Data analysis in the application cases is based on centralized and synchronized data acquisition. Big amounts of data from several heterogeneous data sources such as energy data, process data, MES data or ERP data have to be analyzed in a central infrastructure. Therefore, data acquisition in distributed manufacturing has to cope with the following challenges:

- Heterogenous manufacturing processes: In the manufacturing processes, products of different vendors, proprietary network protocols and different MES are used.

- Time synchronization: Data, which is used for process analysis must be related to the same time instance. Therefore time synchronization of data, which is acquired at different positions in the manufacturing processes, is required.
- Data integration and standard interfaces: Data shall be integrated in a system-wide infrastructure with standard interfaces. Furthermore, the interfaces should contain meta data about the acquired data.

It is outlined in the following sections, how these challenges are met in the particular application cases.

4.1. Distillation columns
In chemical distillation columns, data is gathered on different levels and for different purposes. For safety and control of the product quality, the process data is directly used on the control level for automation. Additionally, the data is used in management execution systems (MES) and enterprise resource planning (ERP) systems for the control of the general production process. Different software tools are used for the different purposes with a set of relevant and well established interfaces, namely OPC-DA, OPC-HDA, OPC-UA and SQL. Due to the sensitive nature of the production plants, the access on the process data is secured and strictly limited. Depending on the level of IT security, any form of OPC can be blocked and SQL remains as only generally allowed interface. The process data, which is available in the considered distillation columns, consists of temperatures, pressures, flow rates and control values. The measuring points are not equidistant with respect to the time and can differ, depending on the measured value.

4.2. Agricultural harvesters
The infrastructure for the agricultural application case is depicted in Fig. 3. Core of the environment are harvesters equipped with embedded data units for acquisition, pre-processing, time-synchronization and transmission of process data (the embedded system in Fig. 3) and a powerful back-end, which is composed of a Hadoop cluster and a data mining tool) for finding patterns in data being delivered by machines and other parts of the infrastructure. The embedded system is configured in a way that data for subsequent analysis is defined as well

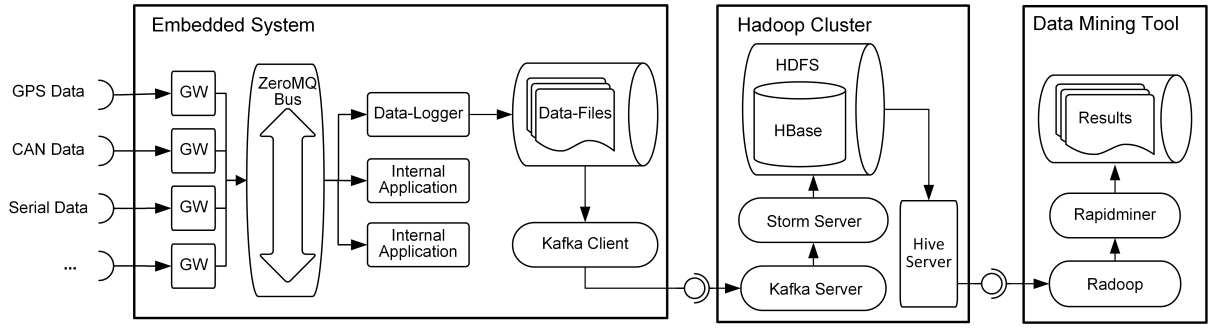

Figure 3. Infrastructure for machine analytics

as specified rules for acquisition (e.g. "record position if the driving direction deviates more than five degrees"). For 2015 harvest season, 10 combine harvesters are equipped with such technology, which provides approximately 3000 different attributes. Data is captured from each available interface (Global Positioning System (GPS) data, CAN bus, serial data, etc.) and sent via a Kafka client to the back-end. In addition, a subset of combine harvesters had been equipped with cameras to provide imagery information of the machine's surrounding, since no

sensors or information sources are available to characterize the crop respectively the harvesting conditions in detail. On the back-end side, the Hortonworks distribution is installed to provide a Hadoop based cluster. Important components are the Kafka server for secure and reliable reception of machine data, the Storm server for fast processing of data and deposition, HDFS as distributed file system, and REST and Hive for data access by analytic tools like R and RapidMiner. The physical infrastructure consists of six servers (five data nodes and a name node). Environment data like field boundaries and digital terrain models were integrated into the cluster to serve as filters in analytic processes.

4.3. Sorting plant

Data acquisition in the sorting plant is accomplished according to Figure 4. Data is collected from three sources:

- TiTech Analysis Kits
- Syperion Flow Meters
- Programmable logic controllers (PLCs).

The TiTech Analysis Kits provide measurements with respect to the material and the composition of the volume flow. The shape and extension of the volume flow is measured by the Syperion Flow Meters. The TiTech Analysis Kits and the Syperion Flow Meters are connected to a central control unit (CCU) and an industry PC (IPC), respectively. Further data is collected from the PLCs of the sorting plant. APIs have been developed in the present work to connnect TiTech analysis kits and Syperion flow meters to the central infrastructure, respectively. The PLC provides an OPC DA interface and is connected via an OPC DA/UA wrapper in the MES API. Due to different data sources, time synchronization has been applied to obtain suitable data sets.

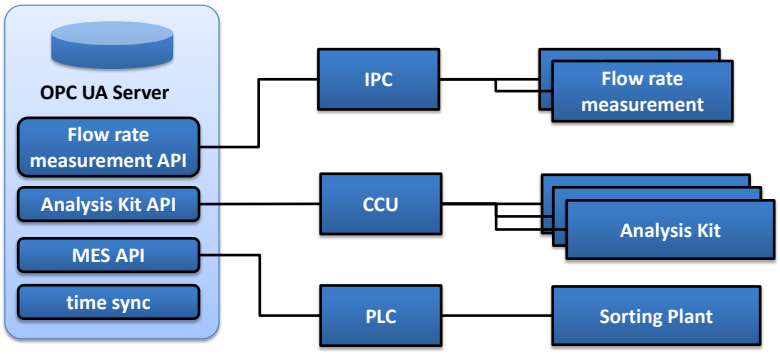

Figure 4. Data acquisition in the sorting plant.

5. Algorithms

In the present work, several model-based and data-driven anomaly detection algorithms have been investigated. Distance based methods compute the distances of potential anomalies to the nearest neighbors in the set of fault-free data. Regression models and self-organizing maps are used to predict observations and to compare the predicted values with the actual observations. Furthermore, PCA-based methods have been investigated.

5.1. Distance based approaches

Initial evaluations have been conducted with distance based methods, which are based on the k-Nearest-Neighbors (k-NN) algorithm (see e.g. [34]). In these approaches, each instance $x = (x_1, \ldots, x_n)$ in a set X of n data points is assessed with respect to the subset $N_k(x) \subset X$ of the k nearest neighbors of x. In doing so, a chosen distance function d, e.g. the Euclidean distance, is used. A score based on the average distance of nearest neighbors is the k-NN Global Anomaly Score (GAS) [30]:

$$GAS_{kNN}(x) = \frac{\sum_{o \in N_k(x)} d(x, o)}{|N_k(x)|}. \tag{1}$$

Instances $x \in X$ with the highest GAS are considered as potential anomalies. The Local Outlier Factor (LOF) [35] of the instance is defined as the ratio of the average inverse GAS of its neighbors and the inverse GAS of x:

$$LOF_k(x) = \frac{\frac{1}{|N_k(x)|} \sum_{o \in N_k(x)} [GAS_{kNN}(o)]^{-1}}{[GAS_{kNN}(x)]^{-1}}. \tag{2}$$

If the densities of an instance are comparable to those of its neighbors the LOF score is approximately equal to 1. An instance with a low density yields a high LOF score and can be considered an outlier. A minimal LOF value of 1.2 is chosen to label an instance as an outlier.

Fig. 5 a) shows the result of the k-NN Global Anomaly Score algorithm applied to the sensor data of three combine harvesters. The x-, y- and z-axis represent motor rotation (x), driving speed (y) and throughput (z) respectively. The colorbar displays the outlier score. The blue instances are considered normal whereas the red and yellow points indicate outliers. The outlier classification highly depends on the user-chosen parameters k and p.

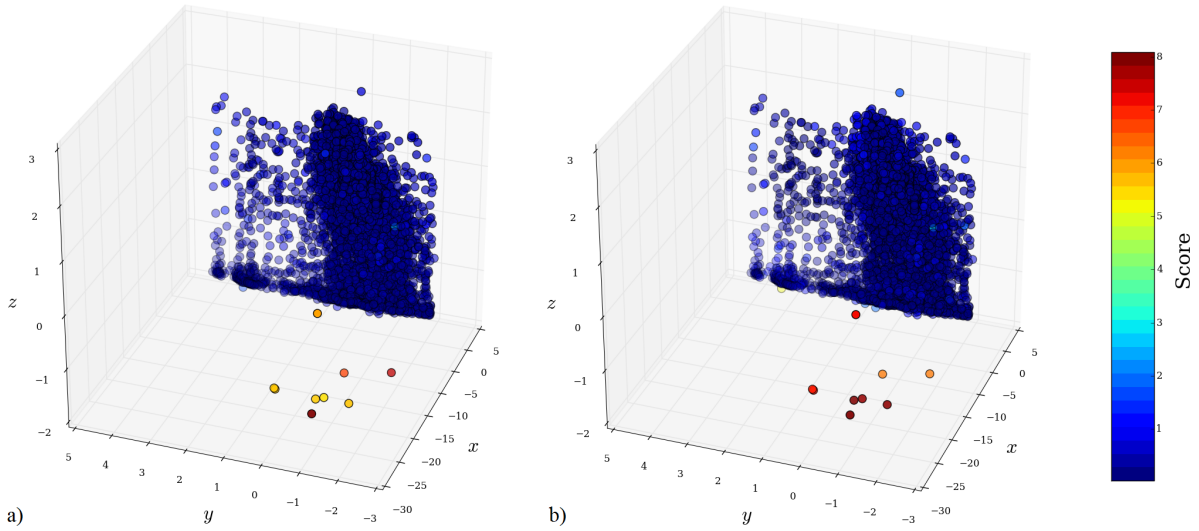

Figure 5. Outlier classification based on $k = 10$: a) GAS score b) LOF score. Outliers are indicated by red and orange circles.

Applying the LOF score on the same dataset as above, we obtain the results shown in Fig. 5 b). In this setup, LOF detects one anomaly more than GAS and allows for more reliable discrimination between anomalies and fault-free data points.

5.2. Regression models

In a regression analysis, the relationship between a specific dependent sensor signal and a set of independent sensor signals has been exploited. Application of linear regression for outlier detection is based on strongly correlated sensor signals. In this case, different attributes, which are represented by the sensor data, are usually generated by the same underlying process in closely related ways [36].

Table 1 shows the correlation between the throughput y of an agricultural harvester and some highly dependent sensor signals $x_1, \ldots, x_m$ in terms of the magnitude of the correlation coefficient.

Table 1. Absolute correlation coefficient $|\rho_{y,x}|$ between throughput y of a harvester and the correlated sensor signals x (for signals with $|\rho_{y,x}| > 0$)

Sensor signal x	$\rho_{y,x}$	Sensor signal x	$\rho_{y,x}$
peripheral reel speed	0.74	motor speed	0.47
fan speed	0.56	shredder	0.44
working position	0.55	water content in grain	0.42
return speed	0.55	average water content	0.36
threshing drum speed	0.53	crop type	0.27
motor load	0.52	lower sieve position	0.25
shaker speed	0.52	upper sieve position	0.24
revolving drum on/off	0.50		

The magnitude of the correlation coefficient has a value of 1 for completely correlated signals and a value of 0 for uncorrelated signals. Based on correlation analysis, a linear model

$$y(x_1, ..., x_m) = a_0 + \sum_{i=1}^{m} a_i x_i \tag{3}$$

has been established for m highly correlated signals with respect to y. Outliers are selected as those instances that exhibit a large deviation from the model.

Fig. 6 displays the outliers (red triangles) found from the successive linear regression method [37], which is based on this model. The x axis is general position system (GPS) latitude, the

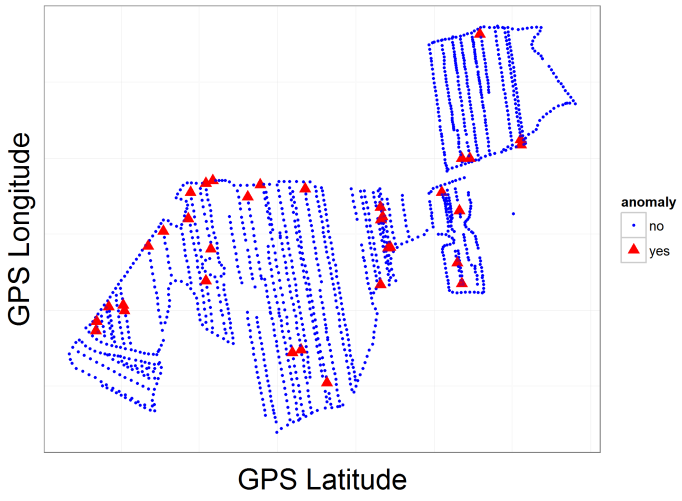

Figure 6. Outliers found from the linear successive regression method

y axis is the GPS longitude. It can be observed that the outliers often lie near corners, field boundaries or crowded regions. In the present work, the linear regression method has further been extended to quadratic regression models. Furthermore, system states with individual regression models assigned to each state have been identified. The system states have been obtained by the OTALA algorithm [18]. For each system state, a particular quadratic regression model has been determined by least squares estimation. Both model learning algorithms - OTALA and the learning of QRMs - have been parallelized by application of MapReduce technology. Initial results have been published in [38]. The MapReduce version of OTALA allows to distribute the workload on $|E|$ edges, by processing the edges E of the created automaton in parallel. For the MapReduce version of QRM, distribution of workload on $|S|$ nodes is possible by processing the states S of the automaton in parallel. Furthermore, online algorithms have been proposed, which efficiently handle novel observations to update the models, which have been created from large historical data sets. In the next step, more complex nonlinear regression models with dependent variables, such as neural networks, have to be evaluated in order to improve the prediction accuracy and to fully exploit the large amounts of data.

5.3. Self-organizing maps

In this approach, a data-driven model of the system in the form of a self-organizing map is generated. With the aid of the UMatrix transformation [21], the various process phases or operating states of the system (e.g. emptying containers or filling containers) were identified using the Watershed transformation. An an example, the experimental demo plant, illustrated in Fig. 7 a) is used, which consists essentially of two containers between which liquid is pumped around in cycles at varying pumping powers and valve positions. The obtained Watershed

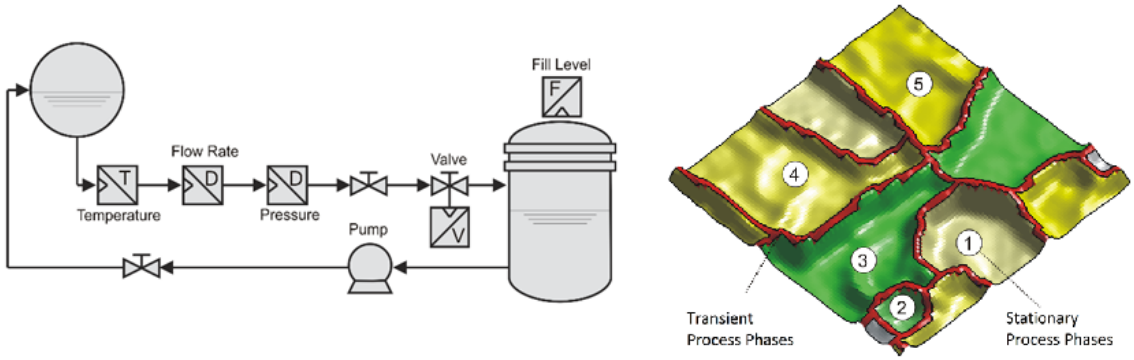

Figure 7. a) Demo plant b) self-organizing map

transformation for this process is depicted in Fig. 7 b). Based on the trained self-organizing map an online-diagnosis of the process behavior is performed. Interventions in the process behavior of the system are clearly revealed in the curve of the quantization error (Q_{error}) of the map by applying a threshold evaluation. The concept is currently evaluated to very large-scale datasets from industrial distillation columns - first results show, that the proposed concept can successfully be applied to real-world applications.

5.4. PCA-based anomaly detection

The PCA-based anomaly detection method that is used in this project [39] works on a data base. Each point in this data base represents the complete sensor data at a given time instance. This interpretation leads to a high-dimensional problem, which can be reduced by dimensionality reduction. Dimensionality reduction leads to a compact representation of system behavior. In

the next step, patterns are extracted, which represent the normal behavior of the system. Fig. 8 shows the data of PCA transformed data of the sorting plant on the left hand side. The area of normal operations is depicted with green points, while known errors are shown as red points. The interference is small so that the green cluster can be used as model of the normal plant operation. Deviations of the learnt model of normal operation can be related to failures. The distance

$$d(y_{pca}(k); x_{pca}(k)) = \|y_{pca}(k) - x_{pca}(k)\|^2 \tag{4}$$

of PCA-transformed measurements $y_{pca}(k)$ with respect to PCA-reduced data points x_{pca} of the normal behavior has been assessed by application of Marr-Wavelets

$$\Psi(y_{pca}(k), x_{pca}(k)) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1 - \frac{2}{\sigma^2}\right) e^{-\frac{d(y_{pca}(k);x_{pca}(k))}{2\sigma^2}}. \tag{5}$$

The standard deviation σ of the Marr-Wavelets is obtained from training data. For $\Psi(y_{pca}(k); x_{pca}(k)) > 0$, a normal process state is assumed at time instance k. $\Psi(y_{pca}(k); x_{pca}(k)) \leq 0$ indicates potential faults, which are displayed to the user.

The PCA-based anomaly detection method has been used to detect and predict a process anomaly in the sorting plant. Based on a test dataset, an initiating process anomaly could be detected approximately 1-2 minutes beforehand.

Figure 8 shows the visualized PCA on the left hand side and the raw data on the right hand side. Further it can be seen that the data points pointing at a process anomaly (yellow/red) are clearly separated from the data points representing the normal behavior (green). The status of the plant (for each sorting module) is visualized using a traffic light.

As a benefit, the error cause can be determined at an early stage by analyzing which signals from the raw data lead to a certain warning.

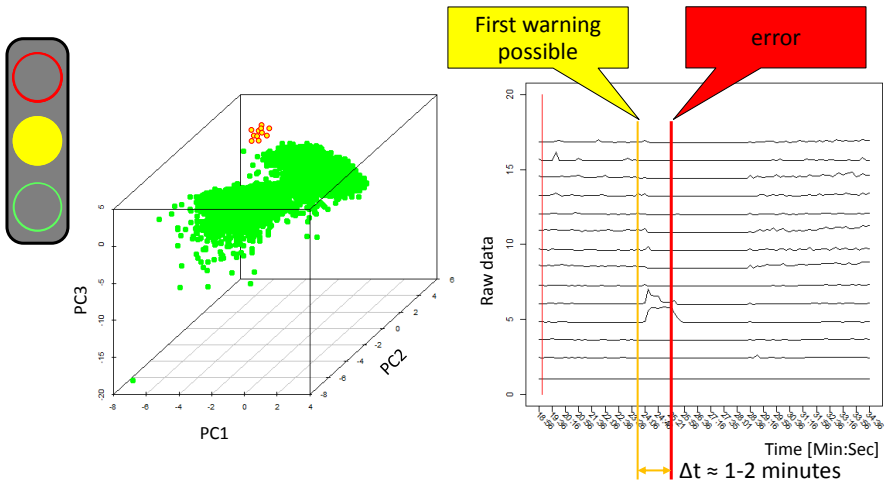

Figure 8. Using the PCA in the sorting plant, the error could be predicted approximately 1-2 minutes beforehand.

6. Conclusion and outlook

In the present work, three industrial and agricultural application cases for big data analysis have been investigated. In these application cases, heterogeneous networks and protocols required the development of convenient data connectors between the central infrastructures and the manufacturing systems. Data analysis in the central infrastructure was accomplished

independently of the particular application cases. For this purpose, several application-independent anomaly detection algorithms have been investigated in the present work. The distance-based anomaly detection method LOF outperformed the GAS score in the investigated agricultural application case. Furthermore, regression-based anomaly detection methods were observed to find anomalies at reasonable positions in this application case. MapReduce algorithms have been developed for efficient learning of hybrid system models, which are based on quadratic regression models. A further step on the research agenda is the development of non-linear regression methods, which are based on neural nets and support vector machines. Self-organizing maps have been shown in the present work to be suitable for anomaly detection in a demo plant of the chemical industry and have to be extended to large-scale plants. A PCA-based method showed promising results for anomaly detection in a large-scale production plant, where process anomalies could be detected 1-2 minutes beforehand.

So far, initial evaluations of the developed algorithms have been conducted in particular application cases, respectively. However, the proposed algorithms are designed to work in each of the three application cases. Systematic evaluation of the proposed algorithms is a further step on the research agenda.

Acknowledgments

The present work was funded by the German Federal Ministry of Education and Research (BMBF) within the cooperative project AGATA (Analyse großer Datenmengen in Verarbeitungsprozessen, engl.: analysis of large amount of data in manufacturing processes) and managed by the Project Management Agency DLR. The author is responsible for the contents of this publication.

References

[1] MANUFUTURE-EU, "Factories of the Future PPP Strategic Multi-annual Roadmap," 2010.
[2] *IEC 61131-3: Speicherprogrammierbare Steuerungen, Teil 3: Programmiersprachen*, International Electrotechnical Commission Std.
[3] F. Pethig, , B. Kroll, O. Niggemann, A. Maier, T. Tack, and M. Maag, "A generic synchronized data acquisition solution for distributed automation systems," in *17th IEEE International Conference on Emerging Technologies and Factory Automaion*, 2012.
[4] *ISO 11783 Tractors and machinery for agriculture and forestry - Serial control and communications data networks*, International Organization for Standardization Std.
[5] W. Emmerich, M. Aoyama, and J. Sventek, "The Impact of Research on the Development of Middleware Technology," *ACM Transactions on Software Engineering and Methodology*, vol. 17, no. 4, pp. 19:1–19:48, 2008.
[6] Object Management Group (OMG), "Common Object Request Broker Architecture (CORBA) Specification Version 3.1," http://www.omg.org/spec/CORBA/, 2008.
[7] F. Picioroaga, A. Bechina, U. Brinkschulte, and E. Schneider, "OSA+ Real-Time Middleware, Results and Perspectives," in *7th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing*, 2004.
[8] P. Leitao, "Agent-based distributed manufacturing control: A state-of-the-art survey," *Engineering Applications of Artificial Intelligence*, vol. 22, pp. 979–991, 2009.
[9] *OPC Data Access Custom Interface Standard Version 3.00*, OPC Foundation Std.
[10] M. Damm, S. Leitner, and W. Mahnke, *OPC Unified Architecture*. Springer-Verlag Berlin Heidelberg, 2009.
[11] P. . P. International, "The profienergy profile," Tech. Rep., 2010.
[12] M. Schleipen and M. Okon, "The caex tool suite - user assistance for the use of standardized plant engineering data exchange," in *IEEE Conference on Emerging Technologies and Factory Automation (ETFA)*, 2010.
[13] S. Rohjans, M. Uslar, and H. Appelrath, "Opc ua and cim: Semantics for the smart grid," in *IEEE Transmission and Distribution Conference and Exposition*, 2010.
[14] L. Christiansen, A. Fay, B. Opgenoorth, and J. Neidig, "Improved diagnosis by combining structural and process knowledge," in *Emerging Technologies Factory Automation (ETFA), 2011 IEEE 16th Conference on*, 2011, pp. 1–8.
[15] S. Faltinski, H. Flatt, F. Pethig, B. Kroll, A. Vodenčarević, A. Maier, and O. Niggemann, "Detecting

anomalous energy consumptions in distributed manufacturing systems," in *Industrial Informatics (INDIN), 2012 9th IEEE International Conference on*, 2012, pp. 358 – 363.
[16] S. Verwer, "Efficient identification of timed automata: Theory and practice," Ph.D. dissertation, Delft University of Technology, 2010.
[17] O. Niggemann, B. Stein, A. Vodenčarević, A. Maier, and H. Kleine Buning, "Learning behavior models for hybrid timed systems," in *Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*, Jul 2012.
[18] A. Maier, "Online passive learning of timed automata for cyber-physical production systems," in *The 12th IEEE International Conference on Industrial Informatics (INDIN 2014)*. Porto Alegre, Brazil, Jul 2014.
[19] S. Haykin, *Neural Networks and Learning Machines (3rd Edition)*, 3rd ed. Upper Saddle River, New Jersey: Prentice-Hall, 2008.
[20] T. Kohonen, *Self Organization and Assoziative Memory*. Springer, 1990.
[21] C. Frey, "Monitoring of complex Industrial processes based on self-organizing maps and watershed transformations," in *IEEE International Conference on Industrial Technology (ICIT)*, 2012.
[22] A. Vodenčarević, H. Kleine Buning, O. Niggemann, and A. Maier, "Identifying behavior models for process plants," in *Emerging Technologies & Factory Automation (ETFA), 2011 IEEE 16th Conference on*, 2011, pp. 1–8.
[23] A. Shui, W. Chen, P. Zhang, S. Hu, and X. Huang, "Review of fault diagnosis in control systems," in *Control and Decision Conference, 2009. CCDC '09. Chinese*, june 2009, pp. 5324 –5329.
[24] B. C. Williams and M. M. Henry, "Model-based estimation of probabilistic hybrid automata," Tech. Rep., 2002.
[25] F. Zhao, X. D. Koutsoukos, H. W. Haussecker, J. Reich, and P. Cheung, "Monitoring and fault diagnosis of hybrid systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 6, pp. 1225–1240, 2005.
[26] S. Windmann, S. Jiao, O. Niggemann, and H. Borcherding, "A Stochastic Method for the Detection of Anomalous Energy Consumption in Hybrid Industrial Systems," in *INDIN*, 2013.
[27] M. Wang and R. Dearden, "Detecting and Learning Unknown Fault States in Hybrid Diagnosis," in *Proceedings of the 20th International Workshop on Principles of Diagnosis, DX09*, Stockholm, Sweden, 2009, pp. 19–26.
[28] G. Wang, S. Yin, and O. Kaynak, "An LWPR-Based Data-Driven Fault Detection Approach for Nonlinear Process Monitoring," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2016–2023, 2014.
[29] S. Yin, S. Ding, A. Haghani, H. Hao, and P. Zhang, "A Comparison Study of Basic Data-driven Fault Diagnosis and Process Monitoring Methods on the Benchmark Tennessee Eastman Process," *J. Process Control*, vol. 22, no. 9, pp. 1567–1581, 2012.
[30] F. Angiulli and C. Pizzuti, *Lecture Notes in Computer Science: Principles of Data Mining and Knowledge Discovery*. Springer, 2002, ch. Fast Outlier Detection in High Dimensional Spaces, pp. 43–78.
[31] M. Breuning, H.-P. Krigel, R. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000.
[32] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast Outlier Detection Using the Local Correlation Integral," in *International Conference on Data Engineering*, 2003.
[33] M. Amer and M. Goldstein, "Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner," in *3rd RapidMinder Community Meeting and Conference*, 2012.
[34] R. Duda, P. Hart, and D. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.
[35] C. Phua, D. Alahakoon, and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, 2004.
[36] K. Chakrabarti, E. Keogh, M. Pazzani, and S. Mehrotra, "Local adaptive dimensionality reduction for indexing large time series databases," *ACM Transactions on Database Systems*, vol. 27, no. 2, pp. 188–228, 2002.
[37] C. Aggarwal, *Outlier Analysis*. Springer, 2013.
[38] S. Windmann and O. Niggemann, "MapReduce Algorithms for Efficient Generation of CPS Models from Large Historical Data Sets," in *IEEE International Conference on Emerging Technologies and Factory Automation (ETFA 2015)*, Luxembourg, 2015.
[39] J. Eickmeyer, P. Li, O. Givehchi, F. Pethig, and O. Niggemann, "Data driven modeling for system-level condition monitoring on wind power plants," in *26th International Workshop on Principles of Diagnosis (DX 2015)*, 2015.