Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 81 (2019) 809–814

www.elsevier.com/locate/procedia

52nd CIRP Conference on Manufacturing Systems

Data mining for fault diagnostics: A case for plastic injection molding

Dominik Kozjek[a,*], Rok Vrabič[a], David Kralj[b], Peter Butala[a], Nada Lavrač[c]

[a]*University of Ljubljana, Faculty of Mechanical Engineering, Aškerčeva 6, 1000 Ljubljana, Slovenia*
[b]*ETI Proplast d.o.o., Obrezija 5, 1411 Izlake, Slovenia*
[c]*Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*

* Corresponding author. Tel.: +386-1-4771-742; fax: +386-1-2518-567. *E-mail address:* dominik.kozjek@fs.uni-lj.si

Abstract

In manufacturing processes the automated identification of faulty operating conditions that might lead to insufficient product quality and reduced availability of the equipment is an important and challenging task. This paper proposes a data mining approach to the identification of complex faults, i.e. unplanned machine stops in plastic injection molding. Several data mining methods are considered, with a focus on the abilities to reveal patterns of faulty operating conditions and on the interpretation of the induced models with the objective to find the data mining method that best corresponds to the nature of the plastic-injection-molding process and the related data. Well-known data mining methods, i.e. J48, random forests, JRip rules, naïve Bayes, and k-nearest neighbors are applied to real industrial data. The results show that tested data mining methods can be effectively used to reveal patterns related to faulty operating conditions. The interpretation capacity of the tested methods, their ability to describe the operating conditions, and to reveal patterns related to faulty operating conditions, are demonstrated and discussed.

Peer-review under responsibility of the scientific committee of the 52nd CIRP Conference on Manufacturing Systems.

Keywords: fault diagnostics; plastic injection molding; data analytics; data mining; industrial data

1. Introduction

Plastic injection molding (PIM) is one of the most widely applied processes for the production of plastic parts. PIM is a cyclic process in which, under high pressure and with high speed, the mold (also called the tool) is filled with hot melt. This is followed by the cooling of the resulting form (product) in the tool, and, in the final stage, the ejection of the product from the tool [1]. One cycle typically lasts about 20 seconds. The cycle time in general depends on the part geometry, the tool construction and the plastic material used.

The settings of the process parameters on the programmable logic controller of a PIM system have to be fine-tuned to achieve process stability and adequate product quality. Monitoring the process parameters can be used for feedback and, in turn, to devise actions for improvement. The parameter setting is usually determined empirically by a process engineer or by using CAE (computer-aided engineering) simulations. The process-parameter settings have an influence on the productivity, the quality, and the cost of production. When a new batch of product is started after a tool change, the initial setting of the control parameters and the ramp-up of the operation involve time-consuming and tedious work performed by an operator, which results in numerous defected products and scrap. Typically, it takes several cycles until the process is stabilized and until products of prescribed quality are generated. This is the case not only when a new tool is introduced but also when a batch is repeated using an existing tool, and previously optimized and stored parameters are used as a starting point during the setup.

PIM process quality can deteriorate due to deviations in the condition of the input material, drifting or shifting of the processing conditions caused by machine or tool wear, environmental change or operator fatigue, or due to planned or unplanned PIM system stops, a change of material batch, etc. The interplay of many phenomena, nonlinearity, and the manipulation of many process parameters during real-time production makes the PIM process difficult to handle.

Peer-review under responsibility of the scientific committee of the 52nd CIRP Conference on Manufacturing Systems.
10.1016/j.procir.2019.03.204

Unplanned machine stops (UMSs) are highly undesirable due to the decrease in quality, the increase in production costs and the loss of productivity and availability. The objective is to have a high-quality, stable process without UMSs.

Modern controllers usually have the option to temporarily store the process-parameter values and other data that describe the working process and the cycle, and the option to export these data. The operations of the production system are usually supported by a manufacturing execution system (MES) that stores the information about work orders, operations, products, tools, operators, etc. With data integration, transformation and using advanced data-analysis approaches, new insights and knowledge can be obtained.

The aim of this work is to demonstrate how the quality and productivity of a production process can be improved by using data mining (DM) approach to analyze the available data of the PIM process. In this study the focus is on the ability of DM to reveal the patterns of faulty operating conditions and on the interpretation of the induced models with the objective to find the DM method that best corresponds to the nature of the PIM process and the related data.

The general analysis workflow that includes the application of DM methods is called the knowledge discovery process [2, 3]. This process starts with selecting the data, possibly from several sources. The selected data need to be pre-processed and transformed in a way that is suitable for the DM process. The outputs of the DM process are predictive models and/or descriptive patterns. The interpretation of these models or patterns can lead to the discovery of new and potentially useful knowledge.

The application of DM techniques represents an alternative to using traditional techniques for dealing with quality problems in manufacturing, such as statistical process control (SPC) and design of experiments (DOE). DM offers tools for the discovery of patterns, associations and statistically significant structures and events in data, including techniques for clustering, association rule mining, classification, and regression [4–6]. DM has been widely applied for quality improvement (QI) in manufacturing [7]. DM methods, either used as stand-alone or as combined (hybrid), and are frequently used for prediction and optimization. Applications of DM approaches in manufacturing mainly address quality analysis (correlating the output quality with the system parameters), failure analyses of production resources, maintenance analysis to enhance the availability of production resources, and production planning and scheduling analysis to improve planning [8].

This paper presents real industrial PIM data used for experimental comparison of the well-known DM methods J48 decision trees, random forests, JRip rules, naïve Bayes and k-nearest neighbors. The results are discussed in terms of the ability to reveal patterns that are related to faulty operating conditions, prediction ability as well as interpretation ability.

2. Data description

The data used in this research are obtained from a real manufacturing environment. The data coming from the PIM process were generated by five PIM machines of a well-known European producer and by the MES. The data were acquired over a period of six months.

There are three different types of data collected and used in the experiments: (1) process parameters records - the tables holding the values of dozens of input and output process parameters for each process cycle, (2) alarms records - the tables holding information about alarm occurrences, and (3) tool exchange records - the data about the tools and tool exchanges. In Table 1, the content of the input data types, are indicated.

Table 1. Experimental data.

Data type	Attributes
process parameters records	timestamp, cycle no., cycle time [s], injection time [s], max. inj. pressure [bar], etc.
alarms records	timestamp, cycle no., alarm code, etc.
tool exchange records	machine ID, tool ID, product code, timestamp of mounting, timestamp of dismounting

Due to the limited storage capacity of machine controllers, the data about process parameters and alarms are daily exported from the machine controllers and stored in hundreds of files.

In process parameters tables, the timestamp, and the cycle number are recorded in addition to the cycle's process-parameter values.

In alarm tables, the timestamp, the corresponding cycle number, and the type of alarm that occurred during a certain cycle are recorded. During the production process, over one hundred different types of alarms occur. Generally, these alarms are divided into two groups: 1st-degree alarms, causing an immediate shutdown of the working process, (some of them can be recognized as an UMS), and 2nd-degree alarms or warnings that report, e.g., about some overruns of the process parameters or minor problems and they are not causing an immediate shutdown of the working process.

The data about the used tools are manual entries of the machine operators into the MES at times when the tools were replaced. These tables hold the information about the machine ID, the tool ID, the date, the approximate time when tool exchange was performed, etc.

Approximately 2.2 million cycles were performed on five observed PIM machines and over ten different tools were changed on each machine during the observed period. The ratio of cycles with to cycles without alarms is generally lower than 1%.

3. Data pre-processing and transformation

Raw data was transformed into an appropriate form for DM. Fig 1. shows an IDEF0 diagram of the implemented procedure of data pre-processing and transformation which consists of the following steps: (1) merging all the files and different data tables into one big common data table, (2) selection of instances referring to selected injection tools, (3) identification of UMS instances, (4) determining the target attribute, (5) choosing the parameters that are in the domain of interest, (6) determining the features that describe the

operating conditions, (7) eliminating the instances of transition areas, and (8) under-sampling instances representing normal operating conditions with a tendency to take into account multiple operating regimes during the observed period.

3.1. Steps

Merging the data. The first major step was to merge all the files and different data tables into one big common data table in which for each cycle, the machine ID, the product ID, the date, the time, the numerical process-parameter values and the alarm occurrences are given. The merging of heterogeneous data is performed using the information about the machine ID, the cycle numbers, and the times of records. In the following steps a transformation of this big data table into a form suitable for data mining was performed.

Selection of instances referring to selected injection tools. In the experiments the data from one machine and one selected tool are used. There were 16 different tools used on the selected machine. Only the data belonging to a tool with the highest number of cycles performed, which is approximately 250,000 cycles, were selected and analyzed.

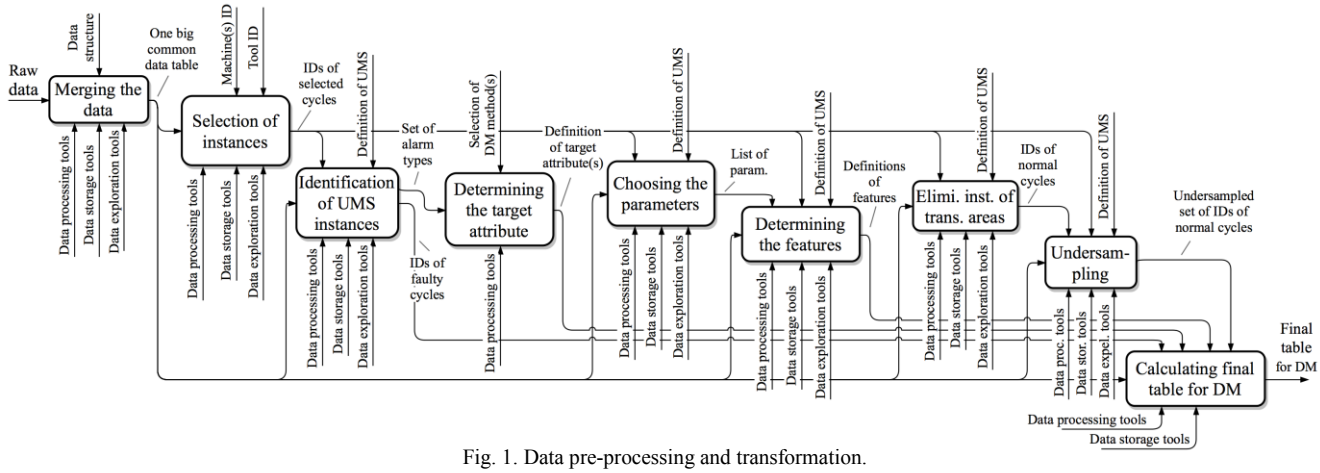

Fig. 1. Data pre-processing and transformation.

Identification of UMS instances. UMSs are unplanned machine stops that are not a consequence of process start-ups, e.g., machine stops that are not in the vicinity (not closer than 50 consecutive cycles) of another machine stop, or machine stops close to cycles when a new batch of product is started after a tool change. Seven 1st-degree alarms were selected. These alarms were used to identify cycles when the UMSs occurred. The examples of chosen alarm types are *too high tool-closing force* and *too high hydraulic-oil temperature*. The UMS cycles are identified by knowing the times when the 1st-degree alarms occurred and the times when the tools were replaced. During the recording period for the selected machine and tool, 87 UMSs were identified.

Determining the target attribute. The chosen alarms were merged into one common attribute, which holds information about whether during some cycle the UMS did or did not occur. This merged attribute is the target attribute of the final table for data mining and it has two possible values: (1) NORMAL (normal operating conditions), or (2) FAULTY (when a UMS occurred).

Choosing parameters in the domain of interest. The attributes of the final table for data mining need to describe the dynamics of the input and output process parameters and the patterns related to UMSs. With respect to these requirements, 31 process parameters (*tool-closing force*, *cycle time*, *time of injection*, *maximum injection pressure in the cylinder*, *temperature 1 of the tool*, *temperature 2 of the tool*, etc.) were selected.

Determining the features. Parameter values in the cycles when the UMSs occurred can be meaningless for further analysis. For example, the value of the parameter that holds information about the time of the cycle. Consider that the process is running and that at a certain point the inappropriate tool-closing force causes a UMS. When that UMS occurs, the process is stopped in the middle of the cycle, the measured time of the cycle differs significantly from the measured cycle times of previous cycles on the observed machine, but note that this is not the reason for UMS, it is a consequence of the UMS. If the overall goal is to forecast UMSs, just observing certain process-parameter values when the UMS occurs makes it too late to act. For these reasons, it is more convenient to observe and find out what is happening with the parameter values at a time before the UMSs' occurrence. The data-transformation mechanism makes it possible to determine the features describing the dynamics of the process parameters in the time interval before the observed cycle. Each of the chosen process parameters is described with two features that hold information about the values and the dynamic of this parameter before the observed cycle. A feature named *AV+parameter name* (e.g., *AV temperature 2 of the tool*) holds a value that is the average of the values of the 1st, 2nd and 3rd cycles performed on the same machine before the observed cycle, and the other feature, called *SL+parameter name*, holds information about the average increasing or decreasing of the parameter value of the 1st and 2nd cycles performed on the same machine before the observed cycle.

Eliminating instances of transition areas. Instances (cycles) that are, in terms of time, in the vicinity (50 cycles or less) of UMSs, near other alarm occurrences or near a tool change, are the instances of the transition areas. These instances are eliminated from the set of instances representing normal operating conditions.

Undersampling instances representing normal operating conditions. The set of instances that represent normal operating conditions was undersampled to a final number of 87 instances. This was done in the following way. In the step *identification of UMS instances*, cycles when the UMSs occurred are identified, then in the step *eliminating instances of transition areas*, from the group of remaining instances, the instances of the transition areas are eliminated. After execution of the above-mentioned steps, some normal instance areas of different lengths (measured in the number of consecutive cycles) are left. Next, from the centres of these normal instance areas, 87 instances are selected in proportion to the lengths of the normal instance areas.

Final table for DM. The final table for data mining has 62 numeric attributes (features) and 1 binary target attribute. The table contains 174 instances. The ratio of normal (NORMAL) to faulty (FAULTY) instances is 87:87. The classification accuracy of the ZeroR classifier is 50%.

3.2. Implementation

The system that generates data mining tables through the steps of reading, encoding, integrating, filtering, querying, transformation, etc., was developed in the Python programming language.

3.3. Possible sources of uncertainty

The following sources of uncertainty and disturbance may affect the considered process: (1) The part of the input data, i.e. tool exchange records, may contain errors (e.g., inaccurate timestamps) because they are manual entries by the machine operators. (2) Sensor data may be incorrect (e.g., due to sensor failure). (3) The determined features may not be able to distinguish faulty from normal examples although the patterns in data exist. (4) The patterns of faulty examples in the final table for data mining may only differentiate operating conditions of faulty and normal groups of examples, and they may not be generally valid for UMSs, but such patterns may have appeared in faulty examples just by coincidence. The most susceptible steps for this type of uncertainty in the considered process are *the elimination of instances of transition areas* and *undersampling of instances representing the normal operating conditions*.

4. Data mining algorithms and experimental setup

This section gives brief descriptions of basic data mining algorithms used in the experiment and their experimental setup. For inducing and evaluating the predictive models the Weka data mining toolkit [9] was used. The accuracy of the models was evaluated with a 10-fold cross validation procedure.

J48 decision tree. The input to the J48 algorithm, also called C4.5, is a set of instances belonging to known classes. The result of using the algorithm is a predictive model in the form of a decision tree or a set of *if-then* rules, which can be used to classify new instances, with the goal of inducing a model that is understandable as well as accurate [10]. In this experiment, the confidence factor is set to 0.25 and pre-pruning is performed by defining the minimum number of instances in a leaf.

Random forests. Random forests (RF) are a combination of tree predictors [11]. Each tree depends on a random sample of cases from the original data, and on a random selection of variables at each node. The class of a new instance is predicted by means of voting by all the constructed trees. In the experiment, different settings of the algorithm parameter, depending on the number of trees, are tested.

JRip rules. The JRip algorithm is a propositional rule learner, implemented incremental pruning to produce error reduction (RIPPER) [9, 12]. In this experiment, the option of pruning and checking for error rate is enabled, the number of folds was set to 3, the minimum total weight of the instances in a rule was set to 2.0 and the seed used for randomizing the data was set to 1. Different settings of the number of optimization runs are tested in the experiment.

Naïve Bayes. Naïve Bayes (NB) is a simple and efficient probabilistic classifier. Supervised discretization is used in the experiment.

k-Nearest Neighbors. The principle of the k-nearest neighbors (kNN) method is to find a predefined number of training instances closest in distance to the new instance and to predict its classification from these instances [2]. Normalization and weighting are important when using distance-based algorithms. There are many possible ways of measuring the distance between two instances. In the experiments, the Manhattan distance measure is used, and different settings of the number of nearest neighbors are tested.

5. Results

This section gives the results of the classification analysis using five well-known data mining algorithms on the input dataset for data mining analysis. Several different algorithm-parameter configurations were tested for J48, random forests, JRip and k-nearest neighbors algorithms.

5.1. Performance measures

Fig. 2 shows classification accuracies of (a) J48, (b) random forests, (c) JRip and (d) k-nearest neighbors for several algorithm-parameter configurations (naïve Bayes is not shown in Fig. 2 because only one parameter configuration was tested for this algorithm). Table 2 gives the estimated classification accuracy, precision and recall measures of all the tested algorithms for the best algorithm-parameter configuration of the individual algorithm with regard to the accuracy measure.

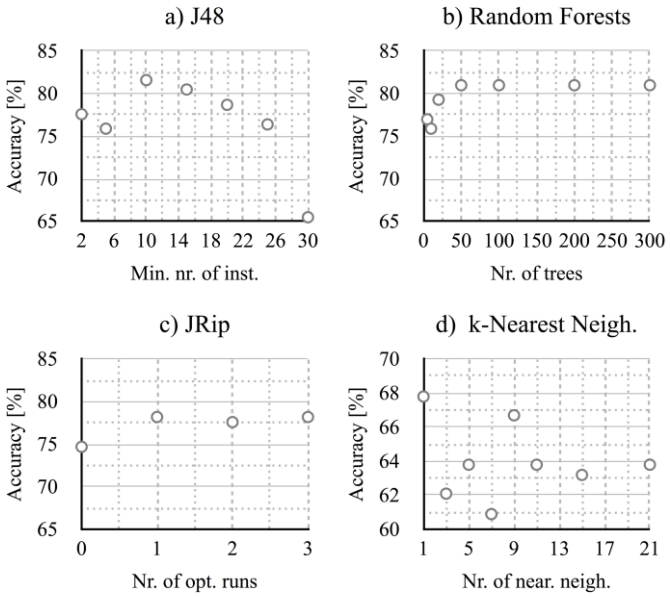

Fig. 2. Classification accuracies of tested algorithms.

Table 2. Performance results of the tested algorithms, using 10-fold cross-validation.

	J48	RF	JRip	NB	kNN
Accuracy [%]	**81.6**	81.0	78.2	74.1	67.8
Precision FAULTY [1]	0.831	**0.864**	0.827	0.862	0.718
Precision NORMAL [1]	**0.802**	0.781	0.747	0.681	0.650
Recall FAULTY [1]	**0.793**	0.759	0.713	0.575	0.586
Recall NORMAL [1]	0.839	0.862	0.851	**0.908**	0.770

5.2. Interpretation

For the two algorithms that allow for results interpretation, the interpretations are given below.

J48 decision tree. The rules obtained from the induced decision tree model (where min. number of instances in a leaf is set to 10) are:

```
if AV temperature 2 of the tool [°C] > 270.33, then
FAULTY; (37/1)
else
    if AV tool opening time [s] > 1.56, then
        if SL cycle time [s cycle^-1] > 0.01, then FAULTY;
        (21/5)
        else NORMAL; (95/18)
    else FAULTY; (21/4)
```

Explanation: Based on the decision tree model shown above, we find that patterns related to faulty operating conditions (conditions which might lead to unplanned machine stops) in the cycles before the observed cycle are: (1) the average value of the second heater in the tool exceeds 270.33 °C, (2) if the average time of opening the tool exceeds the value of 1.56 seconds and at the same time the value of the cycle time is increasing (increasing more than 0.01 seconds per cycle), and (3) the average time of opening the tool is less than or equal to 1.56 seconds.

JRip rules. The induced rules are as follows (number of optimisation runs is set to 3):

```
if AV temperature 2 of the tool [°C] ≥ 272.0, then
  FAULTY; (36/1)
if AV tool opening time [s] ≤ 1.56, then FAULTY; (21/4)
if SL cycle time [s cycle^-1] ≥ 0.01, then FAULTY; (21/5)
if SL time of closing force increasing [s cycle^-1] ≥
  0.005, then FAULTY; (7/2)
else NORMAL; (89/14)
```

6. Discussion

The results show that the tested data mining algorithms J48, random forests, JRip rules, naïve Bayes and k-nearest neighbors are capable of revealing the patterns related to faulty operating conditions. The default accuracy of 50.0% is exceeded for all the tested algorithms.

Algorithms J48, random forests, JRip and naïve Bayes resulted in a higher classification accuracy in comparison to the k-nearest neighbors method. They use the information about the target attribute during the model's induction, unlike the k-nearest neighbors method.

The benefits of using random forests algorithm are the ease of tuning its parameters and a relatively good predictive performance.

The interpretation ability is often an important factor that determines the usefulness of the models. The random forests algorithm gives relatively good results, but the induced model does not have a good interpretation ability. Similarly, the k-nearest neighbors model does not have good interpretation ability. The naive Bayes model could be interpreted using the information contributions of the attributes. The J48 and JRip models can be interpreted in the form of rules. The interpretations suggest that important process parameters with respect to the faulty operating conditions related to UMSs, for a selected tool and machine, are *temperature 2 of the tool*, *tool-opening time*, and *cycle time*.

Reliability factors of the suggested approach can be divided into two groups: (1) data-related, and (2) analysis-related. Data-related factors concern data quality. As already indicated in Section 3 (in *Possible sources of uncertainty*), data may be incorrect, e.g., due to manual input, or due to failure or inaccuracy of the sensors. This can result in an incorrect identification of patterns and incorrect estimation of algorithm performance for identification of investigated faults. Analysis-related factors, i.e. the way how pre-processing parameters are determined, which algorithm implementation and algorithm-parameter configurations are used, selection of algorithm-performance estimation method, etc., mostly affect how well we are able to reveal patterns in the input data. Analysis-related factors also affect the quality of algorithm-performance estimation and the interpretation correctness, e.g., a large number of identified patterns may be just a consequence of model overfitting.

Nevertheless, data mining and data-analysis techniques offer great potential for manufacturing systems, where the

incompleteness of the information and the lack of advanced, system-specific knowledge are the key issues affecting their performance. In manufacturing systems, large amounts of heterogeneous data are being generated and continuously captured, and it is expected that the volumes and heterogeneity of the data will even increase in the future. However, poor data quality, lack of holistic databases, and lack of models for analysis are some of the recent, most pressing challenges in manufacturing systems [13]. The paper proposes an approach to the smart usage of manufacturing data for a widely applied manufacturing process, i.e. plastic injection molding, for the identification of complex faults using several data mining methods, and shows how these methods can be useful in manufacturing practice. To indicate the potential and opportunities of the presented approach, some related work that is associated with advanced analysis in PIM and is based on the same PIM data, is briefly described next.

A holistic approach that includes data generation, acquisition, storage, pre-processing, and prognostics is presented in [14], and it is shown how the approach is able to tackle the high dimensionality and the large size of the PIM data to create and evaluate prediction models for the prognostics of the UMSs. A data-analysis method for cyclic manufacturing processes, such as PIM and die casting, is presented in [15]. The proposed data-analysis method integrates well-known methods, i.e. decision trees and clustering, with the aim of identifying types of faulty operating conditions, and it supports the use of the benefits that are derived from high-performance, data-analytics technologies for storing and retrieving the data. The research described in [16] explores whether in PIM a machine-to-machine approach can be used, in which several work systems share the process data to improve the fault-detection model. The results show that the model can be improved by sharing the data among similar work systems and that it is possible to generalize the process knowledge and apply it to a different work system without prior knowledge.

7. Conclusions

The presented data analysis of real industrial data for a plastic-injection-molding process showed that the tested data mining algorithms J48, random forests, JRip rules, naïve Bayes and k-nearest neighbors can be used to support decision-making during the plastic-injection-molding process. The tested algorithms can be used to reveal patterns related to faulty operating conditions. Regarding the predictive performance and the diagnostics needs, using the presented approach, if we have to choose among the algorithms tested in this study, we suggest the use of J48 and JRip algorithms. The main reasons for this conclusion are relatively good predictive power and at the same time good interpretation ability of models induced by these algorithms.

The presented examples clearly indicate the potentials of data mining as data-analysis techniques in manufacturing systems. However, many challenges and opportunities remain for the development and introduction of data analysis for the smart usage of data in manufacturing systems.

Acknowledgements

This work was partially supported by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia, grant no. 1000-15-0510 and by the Slovenian Research Agency, grant no. P2-0270.

References

[1] Rosato, D.V., Rosato, M.G.. Injection molding handbook. Springer Science & Business Media, 2012.
[2] Bramer, M.. Principles of data mining. Springer, London, 2007.
[3] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.. From data mining to knowledge discovery in databases. AI magazine 1996;17(3): 37–54.
[4] He, S., He, Z., Wang, G.A., Li, L.. Quality Improvement using Data Mining in Manufacturing Processes. Ponce, J., Karahoca, A. (eds.), Data Mining and Knowledge Discovery in Real Life Applications. InTech; 2009, p. 357–372.
[5] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[6] Gardner, R.M., Bieker, J., Elwell, S.. Solving tough semiconductor manufacturing problems using data mining. Advanced Semiconductor Manufacturing Conference and Workshop, 2000 IEEE/SEMI; 2000, p. 46–55.
[7] Köksal, G., Batmaz, İ., Testik, M.C.. A review of data mining applications for quality improvement in manufacturing industry. Expert Systems with Applications 2011;38(10):13448–13467.
[8] Gröger, C., Niedermann, F., Mitschang, B.. Data mining-driven manufacturing process optimization. Proceedings of the World Congress on Engineering; 2012, vol. 3.
[9] Frank, E., Hall, M.A., Witten, I.H.. The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Fourth Edition, 2016.
[10] Quinlan, J. R.. C4.5: programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[11] Breiman, L.. Random forests. Machine learning 2001;45(1):5–32.
[12] Cohen, W.W.. Fast Effective Rule Induction. Twelfth International Conference on Machine Learning; 1995, p. 115–123.
[13] Esmaeilian, B., Behdad, S., Wang, B.. The evolution and future of manufacturing: A review. Journal of Manufacturing Systems 2016;39:79–100.
[14] Kozjek, D., Vrabič, R., Kralj, D., Butala, P.. A Data-Driven Holistic Approach to Fault Prognostics in a Cyclic Manufacturing Process. Procedia CIRP 2017;63:664–669.
[15] Kozjek, D., Vrabič, R., Kralj, D., Butala, P.. Interpretative identification of the faulty conditions in a cyclic manufacturing process. Journal of Manufacturing Systems 2017;43:214–224.
[16] Vrabič, R., Kozjek, D., Butala, P.. Knowledge elicitation for fault diagnostics in plastic injection moulding: A case for machine-to-machine communication. CIRP Annals - Manufacturing Technology 2017;66(1):433–436.