Machine Learning for Enhanced Waste Quantity Reduction: Insights from the MONSOON Industry 4.0 Project

Christian Beecks[1,2], Shreekantha Devasya[2], and Ruben Schlutter[3]

[1] University of Münster, Germany
`christian.beecks@uni-muenster.de`
[2] Fraunhofer Institute for Applied Information Technology FIT, Germany
`{christian.beecks,shreekantha.devasya}@fit.fraunhofer.de`
[3] Kunststoff-Institut Lüdenscheid, Germany
`schlutter@kunststoff-institut.de`

Abstract. The proliferation of cyber-physical systems and the advancement of Internet of Things technologies have led to an explosive digitization of the industrial sector. Driven by the high-tech strategy of the federal government in Germany, many manufacturers across all industry segments are accelerating the adoption of cyber-physical system and Internet of Things technologies to manage and ultimately improve their industrial production processes. In this work, we are focusing on the EU funded project MONSOON, which is a concrete example where production processes from different industrial sectors are to be optimized via data-driven methodology. We show how the particular problem of waste quantity reduction can be enhanced by means of machine learning. The results presented in this paper are useful for researchers and practitioners in the field of machine learning for cyber-physical systems in data-intensive Industry 4.0 domains.

Keywords: Machine Learning · Prediction Models · Cyber-physical Systems · Internet of Things · Industry 4.0

1 Introduction

The proliferation of cyber-physical systems and the advancement of Internet of Things technologies have led to an explosive digitization of the industrial sector. Driven by the high-tech strategy of the federal government in Germany, many manufacturers across all industry segments are accelerating the adoption of cyber-physical system and Internet of Things technologies to manage and ultimately improve their industrial production processes.

The EU funded project MONSOON[4] – *MOdel-based coNtrol framework for Site-wide OptimizatiON of data-intensive processes* – is a concrete example where production processes from different industrial sectors, namely process
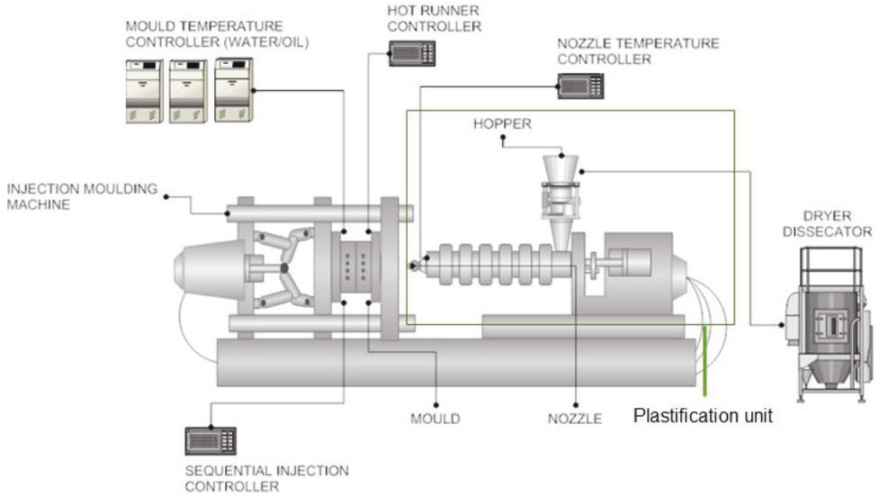
---

[4] http://www.spire2030.eu/monsoon

Fig. 1. Parts and periphery of an injection molding machine (KIMW) [2].

industries from the sectors of aluminum and plastic, are to be optimized via data-driven methodology.

In this work, we are focusing on a specific use case from the plastic industry. We use sensor measurements provided by the cyber-physical systems of a real production line producing coffee capsules and aim to reduce the waste quantity, i.e., the number of low-quality production cycles, in a data-driven way. To this end, we model the problem of waste quantity reduction as a two-class classification problem and investigate different fundamental machine learning approaches for detecting and predicting low-quality production cycles. We evaluate the approaches on a data set from a real production line and compare them in terms of classification accuracy.

The paper is structured as follows. In Section 2, we describe the production process and the collected sensor measurements. In Section 3, we present our classification methodology and discuss the results. In Section 4, we conclude this paper with an outlook on future work.

2 Production Process and Sensor Measurements

One particular research focus in the scope of the project MONSOON lies on the plastic sector, where the manufacturing of polymer materials (coffee capsules) is performed by the injection molding method. Injection molding is a manufacturing process that produces plastic parts by injecting raw material into a mold. The process first heats the raw material, then closes the mold and injects the hot plastic. After the holding pressure phase and the cooling phase the mold is opened again and the plastic parts, i.e., coffee capsules in our scenario, are extracted. In this way, each injection molding cycle produces one or multiple

parts. Ideally, the defect rate of each cycle tends toward zero with a minimum waste of raw material. In fact, only cycles with a defect rate below a certain threshold are acceptable to the manufacturer. In order to elucidate the manufacturing process, we schematically show the parts and periphery of a typical injection molding machine in Figure 1. As can be seen in the figure, the injection molding machine comprises different parts, among which the plastification unit builds the core of the machine, and controllers that allow to steer the production process.

The MONSOON Coffee Capsule and Context data set [2] utilized in this work comprises information about 250 production cycles of coffee capsules from a real injection molding machine. It contains 36 real-valued attributes reflecting the machine's internal sensor measurements for each cycle. These measurements include values about the internal states, e.g. temperature and pressure values, as well as timings about the different phases within each cycle. In addition, we also take into account quality information for each cycle, i.e., the number of non-defect coffee capsules which changes throughout individual production cycles. If the number of produced coffee capsules is larger than a predefined threshold, we label the corresponding cycle with *high.quality*, otherwise we assign the label *low.quality*. The decision about the quality labels was made by domain experts.

Based on this data set, we benchmark different fundamental machine learning approaches and their capability of classifying low-quality production cycles based on the aforementioned sensor measurements. The methodology and results are described in the following section.

3 Application of Machine Learning in Plastic Industry

By applying machine learning to the sensor measurements gathered from a production line of coffee capsules equipped with cyber-physical systems, we aim at detecting and predicting low-quality production cycles. For this purpose, we first preprocess the data by centering and scaling the attributes and additionally excluding attributes with near zero-variance. Preprocessing was implemented in the programming language *R* based on the *CARET* package [7].

Based on the preprocessed data set, we measured the classification performance in terms of *balanced accuracy*, *precision*, *recall*, and *F1* via k-fold cross validation, where we set the number of folds to a value of 5 and the number of repetitions to a value of 100. That is, we used 80% of the data set as training data and the remaining 20% as testing data for predicting the quality of the production cycles. We averaged the performance over 100 randomly generated training sets and test sets.

We investigated the following fundamental predictive models, all implemented via the *CARET* package in *R*:

- *k-Nearest Neighbor* [4]: A simple non-parametric and thus model-free classification approach based on the Euclidean distance.
- *Naive Bayes* [5]: A probabilistic approach that assumes the independence of the attributes.

- *Classification and Regression Trees* [9]: A decision tree classifier that hierarchicaly partitions the data.
- *Random Forests* [3]: A combination of multiple decision trees in order to avoid over-fitting.
- *Support Vector Machines* [11]: An approach that aims to separate the classes by means of a hyperplane. We investigate both linear SVM and SVM with RBF kernel function.

We evaluated the classification performance of the predictive models described above based on the injection molding machine's internal states which are captured by the sensor measurements. The corresponding classification results are summarized in Table 1.

Table 1. Classification results of different predictive models.

	balanced accuracy	precision	recall	F1
k-NN	0.697	0.638	0.686	0.657
Naive Bayes	0.643	0.604	0.563	0.578
CART	0.637	0.595	0.566	0.573
Random Forest	0.653	0.619	0.570	0.589
SVM (linear)	0.632	0.626	0.488	0.540
SVM (RBF)	0.663	0.643	0.563	0.594

As can be seen from the table above, all predictive models reach a classification accuracy of at least 63%, while the highest classification accuracy of approximately 69% is achieved by the k-Nearest Neighbor classifier. For this classifier, we utilized the Euclidean distance and set the number of nearest neighbors k to a value of 7. In fact, the k-Nearest Neighbor classifier is able to predict the correct quality labels for 172 out of 250 cycles on average.

It is worth nothing that this rather low classification accuracy (69%) might have a high impact on the real production process, since in our particular domain hundreds of coffee capsules are produced every minute such that even a small enhancement in waste quantity reduction will lead to a major improvement in production costs reduction. In addition, we have shown that the performance of the k-Nearest Neighbor classifier can be improved to value of 72% when enriching the sensor measurements with additional process parameters [2].

To conclude, the empirical results reported above indicate that even a simple machine learning approach such as the k-Nearest Neighbor classifier is able to predict low-quality production cycles and thus to enhance the waste quantity reduction. Although the provided sensor measurements are of limited extent regarding the number of measurements, we believe that our investigations will be helpful for further data-driven approaches in the scope of the project MONSOON and beyond.

4   Conclusions and Future Work

In this work, we have focused on the EU funded project MONSOON, and have
shown how the particular problem of waste quantity reduction can be enhanced
by means of machine learning. We have applied fundamental machine learning
methods to the sensor measurements from a cyber-physical system of a real
production line in the plastic industry and have shown that predictive models
are able to exploit optimization potentials by predicting low-quality production
cycles. Among the investigated predictive models, we have empirically shown
that the k-Nearest Neighbor classifier yields the highest prediction performance
in terms of accuracy.

As future work, we aim at investigating different preprocessing methods and
ensemble strategies in order to improve the overall classification accuracy. We
also intend to evaluated different distance-based similarity models [1] for improv-
ing the performance of the k-Nearest Neighbor classifier. In addition, we intend
to extend our performance analysis to other industry segments, for instance the
production of surface-mount devices [10], and to investigate metric access meth-
ods [8, 12] as well as ptolemaic access methods [6] for efficient and scalable data
access.

5   Acknowledgements

References

1. Beecks, C.: Distance based similarity models for content based multimedia re-
trieval. Ph.D. thesis, RWTH Aachen University (2013)
2. Beecks, C., Devasya, S., Schlutter, R.: Data mining and industrial internet of
things: An example for sensor-enabled production process optimization from the
plastic industry. In: International Conference on Industrial Internet of Things and
Smart Manufacturing (2018)
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on
information theory **13**(1), 21–27 (1967)
5. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier
under zero-one loss. Machine learning **29**(2), 103–130 (1997)
6. Hetland, M.L., Skopal, T., Lokoč, J., Beecks, C.: Ptolemaic access methods: Chal-
lenging the reign of the metric space model. Information Systems **38**(7), 989–1006
(2013)
7. Kuhn, M.: Building predictive models in r using the caret package. Journal of
Statistical Software, Articles **28**(5), 1–26 (2008)

8. Samet, H.: Foundations of multidimensional and metric data structures. Morgan Kaufmann (2006)
9. Steinberg, D., Colla, P.: Cart: classification and regression trees. The top ten algorithms in data mining **9**, 179 (2009)
10. Tavakolizadeh, F., Soto, J., Gyulai, D., Beecks, C.: Industry 4.0: Mining physical defects in production of surface-mount devices. In: Industrial Conference on Data Mining (2017)
11. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)
12. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity search: the metric space approach, vol. 32. Springer Science & Business Media (2006)