ICIC Express Letters
Part B: Applications
Volume **11**, Number **5**, May **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 447–452

QUALITY PREDICTION MODELING OF PLASTIC EXTRUSION PROCESS

Eunnuri Cho[1,2], Ji-Hye Jun[1,2], Tai-Woo Chang[1,2,*] and Yerim Choi[1,2]

[1]Department of Industrial and Management Engineering
[2]Intelligence and Manufacturing Research Center
Kyonggi University
154-42 Gwanggyosan-ro, Yeongtong-gu, Suwon, Gyeonggi 16227, Korea
*Corresponding author: keenbee@kgu.ac.kr

Received November 2019; accepted February 2020

Abstract. *Manufacturers are trying to improve their quality through predictive analysis using large amounts of data for smart manufacturing. However, it is difficult to select an appropriate quality prediction model since the characteristics of each process and data of companies are different. Therefore, in this study, the environmental data of plastic extrusion process from a manufacturing company were collected and analyzed by four models of logistic regression, support vector machine, random forest, and bagging method. The best model can be selected through performance evaluation using F1 score of each model. If the measurement data of the product are collected automatically in the future, a better method could be found.*

Keywords: Prediction modeling, Quality prediction, Logistic regression, Support vector machine, Random forest, Bagging, Ensemble learning

1. Introduction.

Because the quality in the shop floor is difficult to measure or evaluate compared to the visible indicator such as production quantity, it is usually maintained after completion of the manufacturing process. However, this has the disadvantage that the quality level of the product during the process cannot be evaluated in real time. Nowadays, manufacturing companies are making efforts to innovate the industrial structure by digitizing the entire manufacturing processes for smart manufacturing. Many factories have processing data stored in time series at each facility, and they are trying quality control and forecasting by utilizing various prediction models using the large amounts of data. However, it is difficult to select a proper prediction model since characteristics of each process and data are different. Therefore, in this study, we conduct an experiment to find a suitable predictive model for plastic extrusion process of a company. The extrusion process produces shape by pushing the heated plastic resin through the dies. The extrusion product is cut to length to complete the whole processes after exposed to air at room temperature or passed through a temperature-adjustable water tank. This study will give an experiment to evaluate quality prediction performance using data collected from the extrusion process, and the experiment will use models of logistic regression, support vector machine (SVM), random forest, and bagging. We selected a model which has best prediction performance from the results in the experimental environment set.

This paper is organized as follows. Section 2 describes the related studies to the extrusion process and the predictive models to be used in this experiment. Section 3 deals with the entire contents of the experiment. Section 3.1 explains the structure and characteristics of the feature dataset and explains how to apply a small amount of target values to the model. After explaining the experimental results in Section 3.2, Section 4 states our conclusions.

DOI: 10.24507/icicelb.11.05.447

2. **Related Works.** The myopic goal of this study is to find the most appropriate model to predict quality based on the collected data in the plastic extrusion process. Most of studies on extrusion process considered predicting factors affecting quality using one methodology. Li and Bridgwater studied a three-layer feed-forward artificial neutral network model to predict the extrusion pressure [1]. The neural network model showed how important parameters affect the extrusion pressure. Lela et al. studied linear regression mathematical models for predicting aluminum extrusion temperatures based on continuously recorded data during manufacturing [2]. To obtain a constant extrusion temperature, an appropriate ram speed curve was calculated based on the proposed model. Painter et al. studied the extrusion forging process for manufacturing automotive exhaust valves using computer modeling method developed by a company for Net Shape Manufacturing [3]. Stoyanov et al. developed a data-driven predictive model using SVM, a machine-learning technique, in device-under-test (DUT), the most common test type in the electronics industry [4]. These researches have been applied to only one model in order to predict the factors of the extrusion process. The prediction results obtained by applying only the specific model can be different from the results obtained through various models.

Raju et al. also discussed various approaches for optimization of plastic extrusion process, including Taguchi technique, artificial neural networks, fuzzy logic, genetic algorithms, non-linear modeling, and response surface methodology from a literature review [5]. Li et al. constructed a quality prediction model for the genetic neural fuzzy system (GNFS)-based injection process. The constructed model was compared with other models based on regression analysis and neural network [6]. Therefore, in this study, various prediction models are applied to the data collected in the plastic extrusion process, and the most appropriate model is derived by evaluating the performance of the models.

Classification analysis is a method of predicting unlabeled data by learning labeled data. We used *Scikit-learn* to run the experiment. *Scikit-learn* is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non-specialists using a general-purpose high-level language, ease of use, performance [7]. There are many models that implement classification problems, including k-nearest neighbors, artificial neural network, logistics regression, SVM, and random forest and provide easy access to numerous different classification algorithms. Among the classification models provided by the library, four models of random forest, logistics regression, SVM, and bagging are used to find the best predictive model for the plastic extrusion process in this paper. The description of each model is as follows. Logistic regression is a statistical technique used by D. R. Cox in 1985 to predict the likelihood of an event using a linear combination of features. The purpose of the logistic regression is to create a predictive model by expressing the relationship between the target and the feature as a concrete function in the same way as the goal of general regression analysis. It is similar to linear regression analysis in terms of describing dependent variables as a linear combination of independent variables. However, unlike linear regression analysis, the logistic regression is a classification technique because the target is targeted to categorical data and when the feature data is given, the result of the data is divided into specific categories.

SVM is a machine learning algorithm and is used for supervised learning. It is mainly used for classification and regression analysis. It is a method of classifying by a reference plane that can maximize the relative distance between different heterogeneous groups. Therefore, given a dataset belonging to one of the two categories, the SVM algorithm creates a non-probabilistic binary linear classification model that determines which categories of new data belong to a given data set. SVM is easy to interpret the results and is effective only with little learning data. In this paper, support vector classifier (SVC), one of SVM, is used to derive the results.

Random forest is a kind of ensemble learning method used for classification and regression analysis. It operates by outputting classification or average predictive value from many decision trees constructed in the training process. The most important feature of the random forest is that the trees have slightly different characteristics due to randomness. This property causes the predictions of each tree to be de-correlated, resulting in improved generalization performance. It also makes the data containing noise even more robust through randomization. Randomization is performed in the training process of each tree.

Bagging which is an ensemble learning method using random learning data extraction method is frequently used. In this study, we use the bagging method as one of the random learning methods of random forest. Bagging is an abbreviation for bootstrap aggregating, which is the process of creating a dataset of the same size as an existing dataset by allowing duplication in the given training data. Because the tree has small deflections and large variances, very deeply grown trees are over-summed against the training data. If all the trees that constitute the forest are trained only in the same data set, the correlation of the trees increases. Bagging involves traversing the different sets of data, thereby causing the tree to become non-correlated.

3. **Data Description and Experiments.**

3.1. **Data description.** In this study, we apply various prediction models to the data collected in a plastic extrusion process for analyzing quality improvement. We discussed with the workers what processes, they thought, would affect quality and attached 26 sensors at the relevant points. Data were collected from the sensors via a data acquisition system. The data have 20,801 numerical values at 1 second intervals. The number of collected feature (X value) data is 26 in total for extrusion temperature, dies temperature, screw speed, extrusion speed, cylinder temperature, external temperature, external humidity and water temperature. The type values of the variables are divided into 'continuous' and 'discrete'. In the case of the dies temperature 5, all the values were 0, so it was excluded from the experiment. Table 1 describes the attributes of feature data.

Table 1. Attributes description

Attribute	Description	Type value	Attribute	Description	Type value
TIEXT1	Extrusion temperature 1	continuous	TI_CYL4	Cylinder temperature 4	discrete
TIEXT2	Extrusion temperature 2	continuous	TI_CYL5	Cylinder temperature 5	discrete
TIEXT3	Extrusion temperature 3	continuous	TI_CYL6	Cylinder temperature 6	discrete
TI_DIES1	Dies temperature 1	discrete	Outtemp1	External temperature 1	continuous
TI_DIES2	Dies temperature 2	discrete	Outtemp2	External temperature 2	continuous
TI_DIES3	Dies temperature 3	discrete	Outtemp3	External temperature 3	continuous
TI_DIES4	Dies temperature 4	discrete	Outwet1	External humidity 1	continuous
TI_DIES5	Dies temperature 5	0	Outwet2	External humidity 2	continuous
RPM1_SCREW	Screw velocity	discrete	Outwet3	External humidity 3	continuous
RPM2_EXT	Extrusion velocity	discrete	Water_1	Water temperature 1	continuous
TI_CYL1	Cylinder temperature 1	discrete	Water_2	Water temperature 2	continuous
TI_CYL2	Cylinder temperature 2	discrete	Water_3	Water temperature 3	continuous
TI_CYL3	Cylinder temperature 3	discrete	Water_4	Water temperature 4	continuous

The target (Y value) data are classified into two types, good or defective, and they are collected at irregular intervals because the workers record them manually at the shop floor. Since a small amount of target values were collected compared to the number of feature data, we generated target values in consideration of the relationship between feature values. First, we calculated the correlation between each feature data. And we selected the feature variables with the highest correlation. Then, the minimum and maximum values were considered for the selected feature data. If it is within the selected feature range, target value is generated with good product. The details are as follows.

Table 2 shows the correlation between extrusion temperature and dies temperature, which has a meaningful correlation among the overall feature data. In this study, we confirmed that there is high correlation between the extrusion temperature (TIEXT) 1, 2, and 3. Also, extrusion temperature (TIEXT) 1, 2, and 3 and dies temperature (TI_DIES) 4 have high correlation. However, since each sensor for measuring the extrusion temperature is installed at a close position, we excluded correlation between the extrusion temperature (TIEXT) 1, 2, and 3. Therefore, extrusion temperature 3 and dies temperature 4 were selected. The ranges of the good values of the extrusion temperature 3 and the dies temperature 4 are (202.88 to 210.48) and (202.00 to 209.00), respectively. In this study, the data with range of the feature value was selected as good product, and then the experiment was performed by generating 10% of outlier to prevent overfitting of predictive analysis.

Table 2. Correlation analysis between extrusion temperature and dies temperature

	TIEXT2	**TIEXT3**	**TI_DIES1**	**TI_DIES2**	**TI_DIES3**	**TI_DIES4**
TIEXT1	0.901	0.868	−0.001	−0.239	−0.151	0.802
TIEXT2		0.859	−0.019	−0.233	−0.135	0.762
TIEXT3			0.001	−0.045	0.000	0.884
TI_DIES1				0.000	−0.016	−0.006
TI_DIES2					0.868	−0.179
TI_DIES3						−0.091

3.2. **Experiment result.** Experiments were conducted on four models, and the performance of the classification results was confirmed using four evaluation indexes. Four evaluation indexes were calculated using the module in *Scikit-learn*. A confusion matrix is obtained to calculate precision, recall, accuracy and F1 score. Confusion matrix is a matrix representation for the classification. It is shown in Table 3.

Table 3. Confusion matrix

	Classified as good	Classified as bad
Actual good	TP (True positive)	FN (False negative)
Actual not good	FP (False positive)	TN (True negative)

The formulae shown below are used to calculate precision, recall, accuracy and F1 score. The explanation of each index is as follows.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1\ score = \frac{2 \times precision \times recall}{(precision + recall)}$$

Precision means the true percentage of true that is true. In this experiment, it is an index to judge whether the model is a good product when it is predicted as a good product. Recall is the percentage of what the model is true to be true. It is an indicator that the model is a good product among actual goods. Precision and recall were discussed in the case of predicting true as true. Accuracy refers to the degree to which the actual value and the predicted value agree with each other among all observations. Accuracy not only predicts true but also predicts false. Accuracy is the most intuitive evaluation index for model performance. However, since the bias of data is not considered, the performance of predicting false when the data is false is inevitably low. F1 score means the harmonic mean of precision and recall. If most of the data in the classification model have a specific value, the value of the accuracy is not significant because it is not considered as biased data. This data is called the imbalance dataset and the F1 score is an appropriate indicator to handle it. In this study, we used the F1 score, which considers data bias, as a performance evaluation index reflecting the precision and recall.

Table 4. Classification performance evaluation

Models	Precision	Recall	Accuracy	F1
Logistic regression	0.837933	0.84883	0.813497	0.843346
SVC	0.891058	0.920026	0.886176	0.90531
Random forest	0.900565	0.933355	0.899635	0.916667
Bagging	0.89582	0.93368	0.896558	0.914358

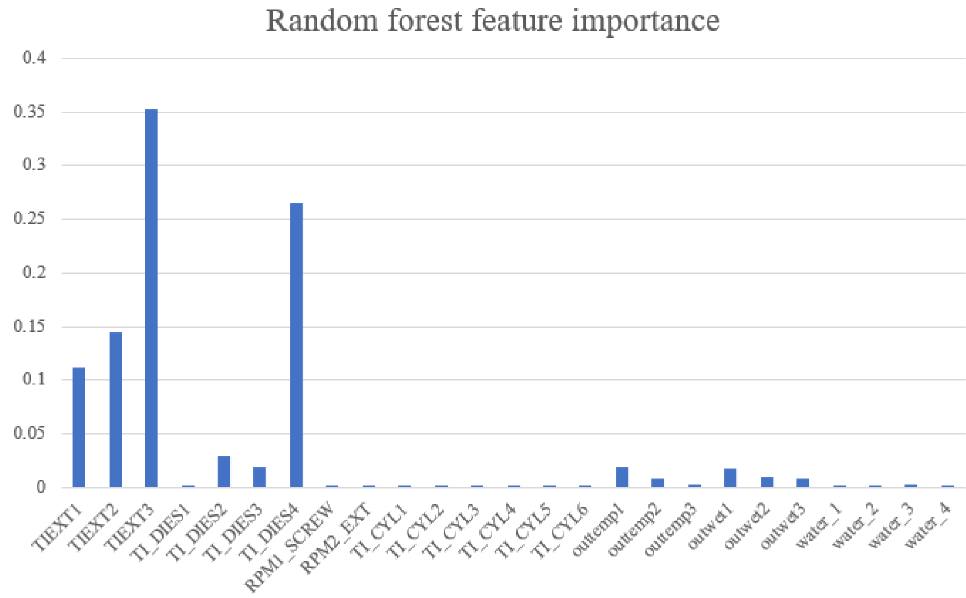

Figure 1. Feature importance in case of random forest model

If the scores of precision and recall are similar, the F1 score will be higher. Therefore, when checking the F1 score, we should also check the scores of precision and recall. When the classification model was applied based on the data collected in the plastic extrusion process, the random forest model among the four models showed the highest F1 score. When the scores of precision and recall were checked, it was found that random forest model had the highest performance with precision 0.900, recall 0.933, and F1 score 0.916. Figure 1 shows the importance of variables in the random forest model, which is the optimal model in this study. The most important in classifying the quality of plastic pipes in this study are the variables of extrusion temperature 3.

4. **Conclusion.** Because there was a requirement from a manufacturer for the use of an advanced technique, we analyzed the manufacturer's real field data. The purpose of this study is to predict the quality of extruded plastic products based on process data collected

by attaching 26 sensors to mechanical equipment. Since the workers measured the quality of products manually, the target data of quality were not collected in real time. In this case, quality prediction is difficult due to lack of target values. Therefore, in order to solve this problem, the correlation between feature data is grasped and target values are generated based on the range of extrusion temperature 3 (TIEXT3) and dies temperature 4 (TI_DIES4), which have the highest correlation. Experiments using virtual data show that the random forest model has the highest performance with F1 score 0.916. However, this may be different from the actual result because it is virtual data. The collection of target values is important for accurate quality prediction. To ensure accurate target values, it is necessary to introduce automation systems and equipment. Therefore, we will build a system that automatically judges the quality of the outer diameter and inner diameter in real time through the image. Therefore, if the model for quality prediction is applied based on the data collected in real time, it will contribute to the improvement of product quality.

Acknowledgement. This work was supported by Kyonggi University's Graduate Research Assistantship 2019.

REFERENCES

[1] Y. Y. Li and J. Bridgwater, Prediction of extrusion pressure using an artificial neural network, *Powder Technology*, vol.108, no.1, pp.65-73, 2000.
[2] B. Lela, A. Musa and O. Zovko, Model-based controlling of extrusion process, *The International Journal of Advanced Manufacturing Technology*, vol.74, nos.9-12, pp.1267-1273, 2014.
[3] B. Painter, R. Shivpuri and T. Altan, Prediction of die wear during hot-extrusion of engine valves, *Journal of Materials Processing Technology*, vol.59, nos.1-2, pp.132-143, 1996.
[4] S. Stoyanov, M. Ahsan, C. Bailey, T. Wotherspoon and C. Hunt, Predictive analytics methodology for smart qualification testing of electronic components, *Journal of Intelligent Manufacturing*, vol.30, no.3, pp.1497-1514, 2019.
[5] G. Raju, M. L. Sharma and M. L. Meena, Recent methods for optimization of plastic extrusion process: A literature review, *International Journal of Advanced Mechanical Engineering*, vol.4, no.6, pp.583-588, 2014.
[6] E. Li, L. Jia and J. Yu, A genetic neural fuzzy system-based quality prediction model for injection process, *Computers & Chemical Engineering*, vol.26, no.9, pp.1253-1263, 2002.
[7] F. Pedregosa, G. Varoquaux, A. Gramfort et al., Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, vol.12, pp.2825-2830, 2011.