Application of Association Rule Mining to Detect "Anode Spike" Disruptions in Aluminum Production

Anna Korobko[1], Anna Metus[2], Dmitry Ogurtsov[3], Tatiana Penkova[2], Iliya Puzanov[4] and Andrey Zavadyak[4]

[1] *Reshetnev Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy ave., Krasnoyarsk, 660037 Russian*
[2]*Institute of Computational Modelling of the Siberian Branch of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia*
[3]*Moscow Institute of Physics and Technology (National Research University), 1 "A" Kerchenskaya st., Moscow, 117303, Russia*
[4] *RUSAL Engineering and Technology Center, 37/1 Pogranichnikov st., Krasnoyarsk, 660111, Russia*

Abstract

The article focuses on applying association rule mining to predict the "Anode spike"-type process disruptions using daily average monitoring data from a series of reduction cells in the experimental area of the Sayanogorsk Aluminum Smelter. The data were binarized according to different criteria for grouping the values of the process parameters into ranges: statistical norms, quartiles, and ranges which are attributed to the occurrence of disruptions. Prediction models were built as a set of association rules. The quality metrics aided in defining the optimum parameters to be used in the model settings, with the results of its validation presented as well. The model selected for the implementa-tion uses binarization based on quartiles. The resulting validation values suggest that the model is effective enough for practical use.

Keywords

Detection of disruptions, data mining, association rule mining, data binarization, aluminum production

1. Introduction

High technical and economic performance in the aluminum industry is largely defined by the operation quality of the process. One of the gravest process disruptions which leads to a significant drop in the metal yield is the anode surface deformation which may be of various types [1]. A "spike" is a buildup of a regular cylindrical or conical shape at the anode bottom; "lagging" is a bulge of a rectangular shape on the anode face, or an irregularity which covers up to 50-60% of the anode area; "overglow" is a buildup of an irregular shape (sphere, mushroom, etc.) at the bottom of the anode which is formed around any side of the anode unit. Currently, such defects are only discovered at a very advanced stage. What causes them is still unclear. There are several hypotheses which are yet to be experimentally validated by means of data mining techniques applied to the monitoring data for reduction cells [1].

A common approach to analyze monitoring data coming from different production processes is to apply association rule mining, a data mining technique which looks for patterns in big data [2-7]. Association rules aid in identifying operation modes leading to higher rates of flawed items [3, 4], finding correlations between various types of defects [5], and predicting the output of the finished product [6]. Applying association rule mining in aluminum production makes it possible to reveal

SibDATA 2021: The 2nd Siberian Scientific Workshop on Data Analysis Technologies with Applications 2021, June 25, 2021, Krasnoyarsk, Russia

EMAIL: gglhroom@gmail.com (A. Korobko); metus@icm.krasn.ru (A. Metus)
ORCID: 0000–0001–5337–3247 (A. Korobko); 0000-0003-0547-5999 (A. Metus); 0000-0002-0057-0535 (T. Penkova)

CEUR Workshop Proceedings (CEUR-WS.org)

patterns across the values of the controlled parameters which are indicative of process disruptions and predict their occurrence in the future.

This paper presents the results of using association rule mining to predict “spike”-type process disruptions based on the daily average data from a series of the reduction cells in the experimental area of the Sayanogorsk Aluminum Smelter as monitored from 01.01.2019 to 30.11.2020. At the stage of preprocessing, based on various criteria the data were binarized according to whatever ranges the process parameters fell into. Prediction models were built as a set of the association rules. Finally, the quality metrics allowed defining the optimum parameters to be used in the model settings, with the demonstration of the validation results.

2. Data preprocessing: binarization

The inputs which make up the association rules are comprised of binarized (or categorized) data with the daily average values of the controlled parameters, to result in a set of elements characterizing the state of the process on a given day.

Binarization is performed in two steps. Firstly, the values of each parameter are grouped into non-overlapping ranges. Secondly, a binary matrix is formed by classifying the data according to the ranges they belong to. The resulting matrix is made up of columns which represent the parameter ranges and rows of 1 or 0, depending on whether the parameter value falls into the range in question.

The ranges were defined both generally and individually. Tackled generally, the data from the reduction cells are considered in combination, and the ranges of the parameter values are found for the entire production as a whole. When approached individually, the ranges are determined for each reduction cell separately, taking into account their specific properties. The criteria to sort out the values into ranges are as follows: statistical norms (*STDDEV*), quartiles (*QUARTILES*), and ranges indicative of disruptions (*HISTOGRAMS*) [8].

To perform binarization based on the statistical norm (*STDDEV*), one needs to determine the standard deviation for a given dataset. The standard deviation is found as $\mu \pm \sigma$, where μ stands for the mean average of the dataset, and σ represents the standard deviation. It forms three ranges: $[\mathrm{x}_0;\mathrm{x}_1)$, $[\mathrm{x}_1;\mathrm{x}_2]$, $(\mathrm{x}_2;\mathrm{x}_3]$, where x_0 is the minimum value in the dataset, $\mathrm{x}_1 = \mu - \sigma$ denotes the standard deviation on the left, $\mathrm{x}_2 = \mu + \sigma$ denotes the standard deviation on the right, x_3 stands for the maximum value in the dataset.

Binarization by means of quartiles (*QUARTILES*) is made by analyzing the number of values in the dataset. It forms four ranges: $[x_0;x_1)$, $[x_1;x_2)$, $[x_2;x_3)$, $[x_3;x_4]$, where x_0 is the minimum value in the dataset, x_1 is the first quartile, x_2 is the second quartile, x_3 is the third quartile, and x_4 is the maximum value in the dataset.

To binarize data based on the ranges that are most indicative of the process disruption (*HISTOGRAMS*), the dataset is broken into two subsets: one containing days when the disruptions occurred, another includes the days free of disruptions. The disruption days include the day a disruption was reported on, and five preceding days. The values are distributed across the non-overlapping ranges on the basis of the frequency distribution for each dataset, to identify the areas where one dataset consistently dominates the other ones. The number of ranges in this case varies: $[x_0;x_1)$, $[x_1;x_2)$, …, $[x_{n-1};x_n]$, where x_0 is the minimum value in the dataset; x_n is the maximum value in the dataset; while x_1, x_2, x_{n-1} denote other ranges. This method of binarization proves reasonable when two and more ranges need to be found.

As a result of binarization, each process parameter is associated with a few binary parameters. The resulting ranges make up a binary matrix which is then used to find association rules.

3. Building a prediction model

The association rule represents a statement of the following type: “*If <X event> then <Y event>*», which is interpreted as a cause-and-effect correlation, where <*X* event> is the part of the rule on the left, or an antecedent, and <*Y event*> is the part on the right, or a consequent.

The association rules are calculated in two main steps: 1) searching for data for the patterns (*itemset*) that appear with the pre-set frequency; 2) making up rules from the found datasets.

To identify how frequently *itemsets* appear in the data, the following criteria are used: *support* and *confidence*. Support is the indication of how frequently a given pattern shows up in the full set. Confidence indicates how frequently the antecedent and consequent co-occur. There is another metric to measure significance (*lift*) which is used to check whether the 'if-then' statements are true, or, in other words, how much the event on the right side of the rule results from that on the left. If the rules are drawn from all the possible itemsets, there may be too many of them, and thus, the calculations are constrained by one of the coefficients.

In contrast to the traditional algorithms to generate association rules which identify various atypical patterns, the suggested model identifies the telltale signs of emerging deviations in the operation process considering only the days which preceded the day the process disruption was found (i.e. it considers the time when the disruption only started to form).

The rule-generating algorithm is based on the application of the concept lattice theory (Formal concept analysis) [9, 10]. Formal concepts are derived from the resulting binary matrix. The formal concept is a pair of sets (A, B), where B is a combination of the binary parameters which partially or fully describes the day before the disruption was discovered. The key element to B is the target parameter which stands on the right side of the rule, and the rest of the binary parameters stands for what is on the left. The element A represents a set of events which are covered by the created rule.

The size of the set |A| allows calculating the rule support coefficient juxtaposing it with the size of the input set of the events (both with favorable and undesired outcomes). The confidence coefficient is determined by associating the rule support coefficient with the frequency of the consequence, separately from the antecedent. Both coefficients help evaluate the strength of the rule (lift). In this regard, the rules with high confidence and low support indicate a rare combination of the parameters which are seen only in one or two events. Conversely, the rules with low confidence and high support describe a frequent pattern which is found both on the disruption days and disruption-free days.

Figure 1 shows a scatter plot for the rules worked out for reduction cell No. 6 in the experimental area PA-550. The training set included the data from 2019 when there were 10 "spike"-type disruptions detected in the reduction cell. The algorithm with STDDEV binarization generated 577 association rules. In the graph, the set of rules with the highest confidence is highlighted in yellow.

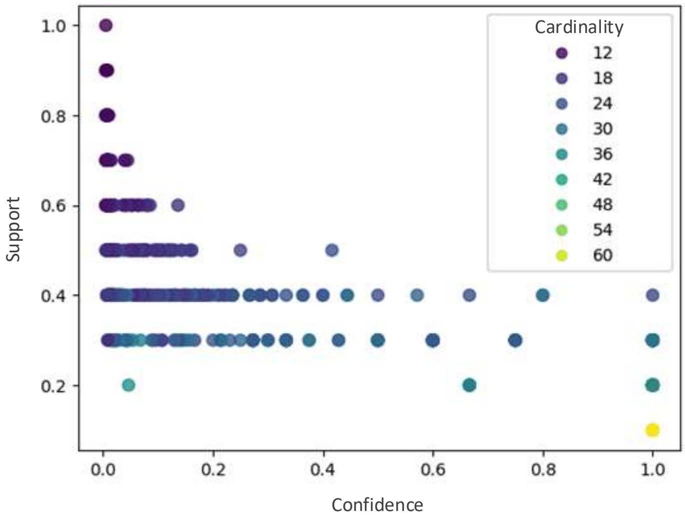

Figure 1: The scatter plot for the rules in Reduction Cell No. 6 in the Experimental Area PA-550, 2019

The association rules based on the analysis of the preceding events allows determining the ranges of values and combinations of the controlled parameters which describe the state of the process before the disruption was detected, considering the specific characteristics of the reduction cells. The resulting set of rules can be used to predict process disruptions and evaluate the probability of an unfavorable event.

4. Validating and adjusting the prediction model

The predictive accuracy of the model is determined using the following indicator:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (1)$$

where TP is the true positive outcome; FP is the false positive outcome; while FN stands for the false negative result, and TN, for the true negative one.

At the same time, the percentage of True Positives is found as a ratio of the TP outcomes to the number of entries which correspond to the days prior to disruptions, while the percentage of False Positives is defined as a ratio of the FN results to the number of entries which describe the disruption-free days as per the current day and the one that follows.

The main parameters used in the model tuning are the following:

- type of the rule formation (individual or general approach): $\text{ind\_type} = (\text{true}, \text{false})$;
- type of binarization: $\text{bin\_type} = (\text{STDDEV}, \text{QUARTILES}, \ \text{HISTOGRAMS})$;
- data completeness (with or without the consideration of empty values): $\text{keep\_nan} = (\text{true}, \text{false})$;
- range of values (with or without the consideration of the "routine" parameter values): $\text{anomalies\_only} = (\text{true}, \text{false})$.

The model was validated based on the monitoring data from Reduction Cell No. 7 in the Experimental Area PA-550. The training dataset includes entries for the period of 019.01.01-2020.01.01, namely 730 entries, and 13 "spike"-type disruptions. The test dataset is comprised of entries for the period of 2020.01.01-2020.05.01, including 121 entries, and 18 "spike"-type disruptions. The results of the model validation with different tuning parameters are presented below.

Model 1: $ind\_type = true$; $bin\_type = STDDEV$; $keep\_nan = true$; $anomalies\_only = true$. Number of rules – 101.

How the model performs is demonstrated in Figure 2. On the left is a graph which shows the changes in the percentage of the true and false positive predictions, depending on the confidence threshold value, and on the right is a diagram which juxtaposes the prediction results with the actual events of the process disruptions. The maximum accuracy of the model amounts to 0.85 (i.e. 85% of true predictions) with the confidence threshold value of 0.61, $TP/FN = 8/10$; $TN/FP = 95/8$.

Refining the model tuning parameters – $ind\_type = false$, $anomalies\_only = false$ – decreases the overall accuracy and results in a higher number of false positive predictions. The maximum accuracy is achieved by increasing the number of true positive outcomes: $TN = 99$, $TP = 0$.

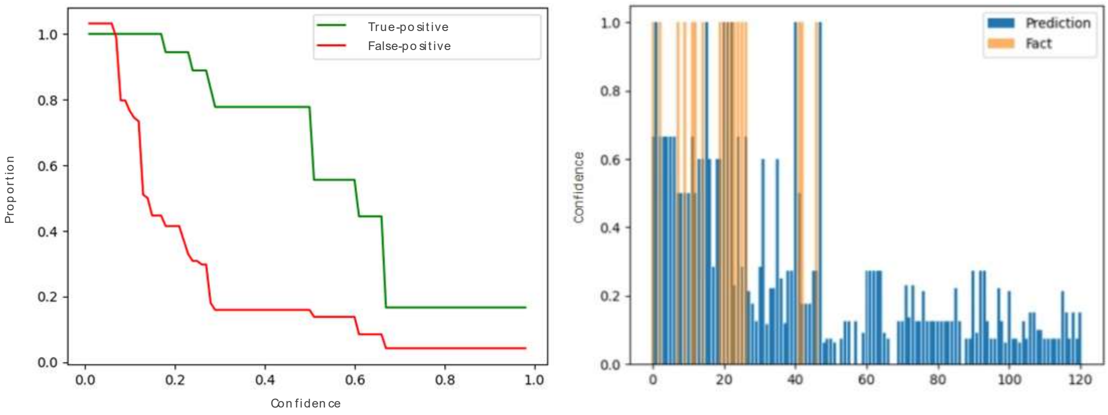

Figure 2: The performance of Model 1

Model 2: $ind\_type = true$; $bin\_type = QUARTILES$; $keep\_nan = true$; $anomalies\_only = true$. Number of rules – 557.

Figure 3 shows the model performance. Its maximum accuracy is 0.88 with the threshold value of 0.41, $TP/FN = 13/5$; $TN/FP = 93/10$. At the confidence threshold values within the range of

0.41-0.42, the percentage of true positive predictions amounts to 72%, while that of False Positives is equal to 11%.

When the model tuning parameters are changed for $ind\_type = false$, $anomalies\_only = false$, there is a drop in the overall accuracy, with a higher number of false positive predictions.

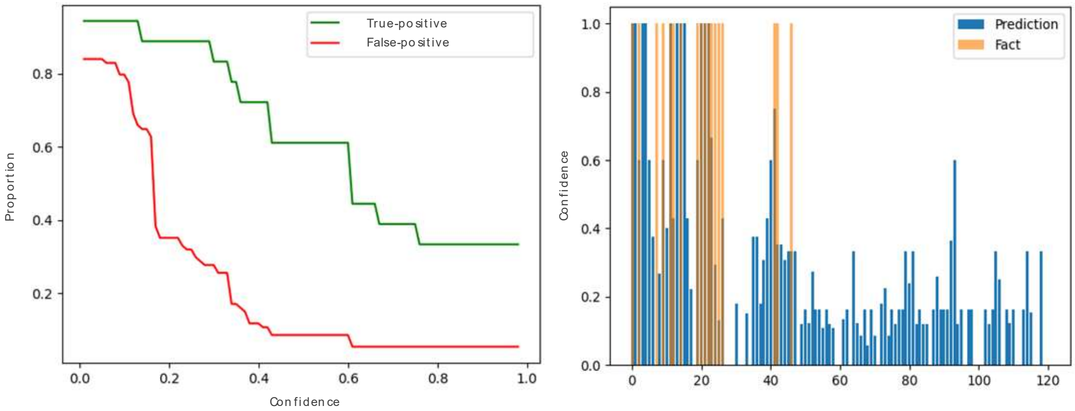

Figure 3: The performance of Model 2

Model 3: $ind\_type = true$; $bin\_type = HISTOGRAMS$; $keep\_nan = true$; $anomalies\_only = true$. Number of rules – 270.

The performance of the model is demonstrated in Figure 4. Its accuracy reaches 0.86 at the confidence threshold of 0.51, $TP/FN = 1/17$; $TN/FP = 103/0$. The performance results suggest that the prediction error is rather high, while the evaluations of the true positive and false positive predictions proves to be similar. The high confidence of the rule may indicate a disruption which is currently developing and thus, has not been detected yet.

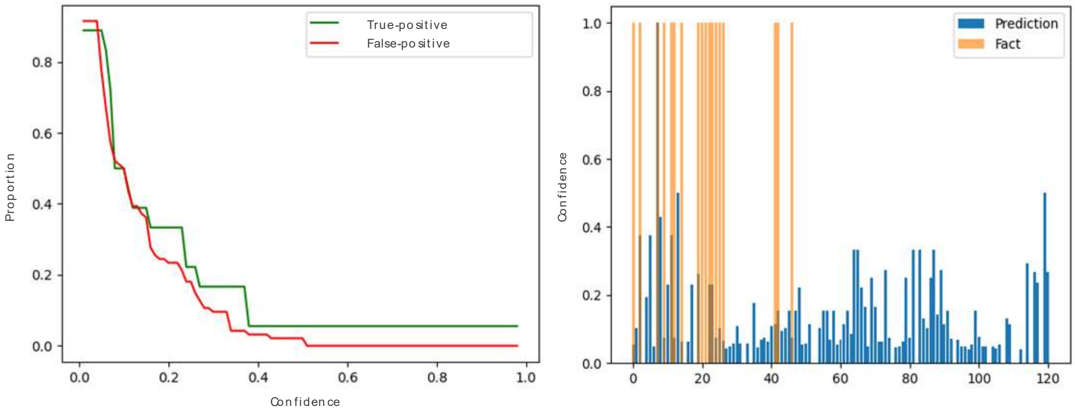

Figure 4: The performance of Model 3

Table 1 shows the results of testing the model with different types of data binarization. It presents the dates preceding the days the “spike”-type disruptions were reported on, and the corresponding confidence of the operating rule which is interpreted as the probability of a disruption to occur if the controlled parameters show the values as observed.

The validation results show that the predictive accuracy of the model is higher with the individual approach. When selecting the parameters at the model tuning stage, it is important to consider the data completeness and to look at the entire range of values of the process parameters. To a great extent, the accuracy of the given model depends on the period covered by the training set. The conditions of the technology process do change over time, so do the ranges of values and combinations of the parameters characterizing the occurring disruptions, which inevitably affects the results of their prediction. Among the models with different types of binarization, the best results were demonstrated by the one with quartile-based binarization. Therefore, the model with the following parameters was chosen to be implemented: $ind\_type = true$; $keep\_nan = true$; $anomalies\_only = true$; $bin\_type = QUARTILES$.

Table 1
The results of testing the association rule

Date	STDDEV	HISTOGRAMS	QUARTILES
14.08.2020	0.545	0.5	1
18.08.2020	0.375	0.666	0.75
23.08.2020	0	0.417	0.625
04.09.2020	1	0	1
07.09.2020	0.375	1	1
11.09.2020	1	0.429	1
09.10.2020	0	1	0.6
11.10.2020	0	0.615	1
13.10.2020	0	0.412	0.385
31.10.2020	0.429	0	1
04.11.2020	0	0	0.5
10.11.2020	0.375	0	1
14.11.2020	0.5	0.429	1
15.11.2020	0	0,5	1
18.11.2020	1	1	1
19.11.2020	0.625	0.833	1
22.11.2020	0.4	0.5	1
23.11.2020	0.714	0.5	1

5. Conclusion

The article presents the results of the association rule technology applied to predicting “Anode spike”-type process disruptions on the basis of daily average monitoring data from a series of reduction cells in the experimental area of the Sayanogorsk Alumi-num Smelter. At the preprocessing stage, the data were binarized using various criteria to divide the values of the process parameters into ranges: statistical norms, quartiles, and ranges indicative of disruptions. The predictive models were built as a set of association rules. The testing results suggest that the predictive accuracy is higher with the individual approach. Moreover, it is crucial to take into account the completeness of the data and consider the entire range of the values of the process parameters. The accuracy of the model proved to largely depend on the period covered by the training set because the very conditions in which the given process runs change considerably over time. The model with the quartile-based binarization was selected for the implementation. The validation results indicate that the model is of rather high quality for practical use. Further research into monitoring inputs is required to obtain a higher predictive accuracy.

6. References

[1] Yu. G. Mikhalev, P. V. Polyakov, A. S. Yasinsky, S. G. Shakhrai, A. I. Bezrukikh, A. V. Zavadyak, Causes of Process Disruptions Involving Anodes. Review of Russian and Overseas Experimental Data, SFU Journal. Engineering and Technologies 10(5) (2017) 593–606.

[2] J. Treinen, T. Ramakrishna, A Framework for the Application of Association Rule Mining in Large Intrusion Detection Infrastructures, in: Proceedings 9th International Symposium on Recent Advances in Intrusion Detection, LNCS, volume.4219, 2006, pp. 1–18.

[3] J. Jeon, S. Y. Sohn, Product failure pattern analysis from warranty data using association rule and Weibull regression analysis: A case study, Reliability Engineering and System Safety 133 (2015) 176–183.

[4] J. Kim, H. Hwangbo, Real-Time Early Warning System for Sustainable and Intelligent Plastic Film Manufacturing, Sustainability 11(5) (2019) 1490.

[5] K. Wongwan, W. Laosiritaworn, Application of Association Rules in Woven Wire Mesh Defects Analysis, in: 7th International Conference on Industrial Technology and Management, 2018, pp. 325–329.

[6] H. T. Hu, R. Z. Zhang, X. Guan, Application on Crude Oil Output Forecasting Based on TB-SCM Algorithm, in: Proceedings 5th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2015, pp. 398–401.

[7] J. Kim, Y. Lee, Progress of Technological Innovation of the United States' Shale Petroleum Industry Based on Patent Data Association Rules, Sustainability 12(16) (2020) 6628.

[8] A. Metus, T. Penkova, Analysis of Aluminium Electrolysis Data in the Context of Extreme Values of Technological Parameters, in: CEUR Workshop Proceedings, volume. 2727, 2020, pp. 92–98.

[9] B. Ganter, R.Wille, Formal Concept Analysis: mathematical Foundations, Springer-Verlag, Berlin Heidelberg, New York, 1999.

[10] R. Wille, Restructuring Lattice Theory: an approach based on hierarchies of concept, Reidel, Dordrecht-Boston, 1982.