Application of Ensemble Algorithms to Detect Anode Effects in Aluminum Production

Anton Mikhalev[1], Nina Lugovaya[1], Tatiana Penkova[2], Iliya Puzanov[3] and Andrey Zavadyak[3]

[1] Siberian Federal University, 26, Kirenskogo st., Krasnoyarsk, 660074, Russia
[2]Institute of Computational Modelling of the Siberian Branch of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia
[3]RUSAL Engineering and Technology Center, 37/1 Pogranichnikov st., Krasnoyarsk, 660111, Russia

Abstract

The article considers the approaches for automatically detecting process disruptions known as anode effects in aluminum electrolysis. The suggested method of their identification is based on using ensemble algorithms which are applied to data from immediate and daily average monitoring of reduction cells. The method includes the stage of preprocessing of daily-average inputs, aggregation of immediate and daily average parameters, and construction of a math model. The study determines the most informative parameters, while analyzing how algorithms and approaches featuring ensembles of decision trees stack up against each other. The quality metrics reveal the most effective algorithm for the set task.

Keywords

Ensemble algorithms, detection of disruptions, anode effects, aluminum production

1. Introduction

One of undesirable events arising from aluminum electrolysis is the anode effect. This phenomenon may adversely affect the process of electrolysis leading to excessive power consumption and temporary cell superheating. Moreover, whenever anode effects appear in the middle of electrolysis, this results in the formation of greenhouse gases, as well as in occasional sparks and arc discharges between the anode surface and electrolyte, which is hazardous for an operator addressing the effects [1].

Currently, a number of aluminum producers are developing technological strategies to find ways to make their reduction cells operate without any anode effects. The possibility to detect them is one of the top priority tasks concerning the aluminum production management. Traditionally, anode effects are reduced by maintaining the concentration of alumina within the set range to ensure its most consistent dissolving in the electrolyte, which is made possible by incorporating automatic adjustment units into the automatic alumina handling system [2, 3]. The present-day approach to keeping the number of anode effects to a minimum involves their timely detection and prediction. The state of the process facilities is normally predicted by means of machine learning which allows identifying specific correlations in the data and using them to find process disruptions in the cell operation.

Algorithmic tools to detect the process disruptions are developed in the following stages: 1) analysis and preprocessing of input data; 2) identification of informative features; 3) construction of a math model and validation of results. Taking this into account, the body of the article is arranged accordingly. Chapter 2 sets the task of classification. Chapter 3 presents the description of the input data. Chapter 4 considers the applied methods of the input data preprocessing. Chapter 5 describes the algorithms of classification. Chapter 6 presents the results of the applied classification models.

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

EMAIL: asmikhalev@yandex.ru (A. Mikhalev)
ORCID: 0000-0002-8986-5953 (A. Mikhalev); 0000-0002-2939-0298 (N. Lugovaya); 0000-0002-0057-0535 (T. Penkova)

© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3047/paper11.pdf

2. Research objective

The classification task is set as follows. Let us assume that there is a set of objects $X = \{X^{(1)}, \dots, X^{(n)}\}$, each characterized by the m-dimensional vector of attributes $X^{(i)} = \left(x_1^{(i)}, \dots, x_m^{(i)}\right), i = \overline{1, n}$. Each object under study is attributed to a certain class $C_j \in Y = \{C_1, \dots, C_k\}, j = \overline{1, k}$. In this case, the classification is aimed at the following. It requires a rule (algorithm) to be formulated $a: X \to Y$, so that based on the set point value of attributes, new unknown objects could be attributed to one of the classes.

Being related to the problem of early detection of process disruptions based on the monitoring data, the task of classification is reduced to dividing the states of the process facility into two classes: operative $Y = 0$ and faulty $Y = 1$ (functioning with errors). The input data samples are used as the basis for an algorithm which must be able to use the set operative indicators of the given facility to diagnose its state with sufficiently high accuracy.

Binary classification tasks normally use the following indicators as their metrics:

- *accuracy* is the relation of all the correctly classified objects to the total number of all the classified objects:

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

Here, *TP* stands for the true-positive results (objects classified as “positive” and which are actually positive, i.e. belong to the class $Y = 1$), *TN* stands for the true-negative results (objects classified as “negative” and which are actually negative, i.e. belong to the class $Y = 0$), *FP* stands for the false-positive results (objects classified as “positive” but which are actually negative, i.e. belong to the class $Y = 0$), *FN* stands for the false-negative results (objects classified as “negative” but which are actually positive, i.e. belong to the class $Y = 1$).

In the case of imbalance between the classes, the regular accuracy is replaced with the balanced accuracy:

$$accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \qquad (2)$$

- *precision* is the relation of all the objects classified as “positive” and which are actually positive to the total number of objects classified as “positive”:

$$precision = \frac{TP}{TP + FP} \qquad (3)$$

Precision characterizes the ability of the given prediction model to correctly classify positive objects in relation to the number of all the objects classified as “positive”.

- *recall* is the relation of all the objects classified as “positive” and which are actually positive to the total number of actually positive objects:

$$recall = \frac{TP}{TP + FN} \qquad (4)$$

Recall characterizes the ability of the given prediction model to correctly classify positive objects from the set of all the positive objects combined.

- *F1* score is the harmonic mean between the values of *precision* and *recall*:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \qquad (5)$$

F1 score demonstrates how many cases are correctly classified by the model, and how many true items can be correctly classified by the model.

3. Description of inputs

The state of the process facilities typically undergoes diagnostic scanning performed as the ongoing monitoring of their current state, and whichever forms, it is manifested over time. The fundamental principle of this diagnostic routine is in taking consistent and systematic measurements of the parameters characterizing the running process at the facility, identifying changes in relation to the standard indications, and their further classification.

The key parameter to control anode effects is the value of voltage in the reduction cell. Anode effects are known to be accompanied by a spike in voltage. However, checking the state of the reduction cells and predicting the process deviations using the voltage parameter do not prove to be highly effective. To ensure better predictive accuracy and to subsequently reduce the number of anode effects, it is necessary to control the range of the parameters.

The daily average data collected through monitoring the operation of cell series include the following parameters: duration of metal tapping (sec), metal level (cm), electrolyte level (cm), electrolyte temperature (°C), alumina dose (kg), bath chemistry parameters, parameters of the point feeding system for alumina and aluminum fluoride, adjustment parameters of the anode-to-cathode distance, amperage (kA), voltage parameters, back EMF (V), state and service life of cells (month). Different time of discretization across the process parameters is one of the obstacles in controlling aluminum electrolysis. Voltage is measured continuously, whereas other parameters are recorded only once a day. Introducing these data into the prediction model entails selecting a suitable scheme for their possible rearrangement when modeling the training set. The underlying strategy of this research is the data aggregation presented earlier [4]. The simulation is carried out as follows: instantaneous values of voltage are averaged out every 10 minutes, the resulting averaged values are then combined with the daily average data retrieved by the staff from sensors at the beginning of their night shift (20:00). The research is based on the monitoring data from the experimental area of the Sayanogorsk Aluminum Smelter, namely a series of reduction cells in potrooms No. 9 and 10. The input data comprise the values of the process parameters for the period of 2020-2021. The output variable represents information about the occurring disruptions (anode effects) in the process.

4. Preprocessing of inputs

The first stage of the daily average data processing involves the search for outliers across the data. The outliers are found using the method of quartiles. The values identified as outliers are substituted for by Nan. The next step is to reconstruct the missing data points. The first features to be removed are those with the gaps exceeding the set threshold (over 50% of entries), and the remaining data are then reiterated with the missing data points reconstructed by means of the EM-algorithm (Expectation-maximization). This is followed by modeling the training set. The final stage is focused on the most informative parameters indicating the occurrence of the anode effects. The datasets contaminated with uninformative features lead to the model overfitting. The features are selected using the method of recursive feature elimination (RFE) combined with the random forest algorithm. The RFE method is based on the consecutive construction of models so that each step results in a model and the feature which proves to be the least informative is eliminated from the set. This eventually forms a set of the most significant features allowing one to detect the anode effects, including the following: electrolyte level, electrolyte temperature, cryolite ratio, alumina dose, Fe concentration, Mg concentration, metal level, duration and number of VIRA, duration and number of MAINA, number of alumina doses in the automatic and manual mode, period of starvation, period of oversaturation, and actual volume of the produced metal. **Ошибка! Источник ссылки не найден.** gives the statistical description of the monitoring data (training and testing) for a reduction cell.

Table 1
The statistical description of the inputs

Parameter	Records	Mean	Min	Max	Standard deviation
Electrolyte temperature	49,900/ 5,790	956.908/ 953.946	931/ 937	977/ 971	8.9056/ 6.7454
Electrolyte level	49,900/ 5,790	17.788/ 19.265	11/ 13	22/ 25	1.7363/ 1.7479
Alumina dose	49,900/ 5,790	7.383/ 6.955	7.09/ 7.12	7.43/ 7.43	0.1174/ 0.2984
Fe concentration	49,900/ 5,790	0.056/ 0.054	0.0274/ 0.0435	0.0842/ 0.0775	0.013/ 0.0068
Cryolite ratio	49,900/ 5,790	2.309/ 2.323	2.0279/ 2.1304	2.5572/ 2.51	0.0968/ 0.0874
MgF_2 concentration	49,900/ 5,790	0.549/ 0.546	0.4206/ 0.42	0.77/ 0.69	0.053/ 0.053
Metal level	49,900/ 5,790	16.585/ 15.963	12/ 12	22/ 22	1.8536/ 2.1254
Anode-cathode distance: number of VIRA	49,900/ 5,790	12.955/ 15.659	4/ 7	31/ 31	4.6043/ 4.9156
Anode-cathode distance: number of MAINA	49,900/ 5,790	6.786/ 6.983	0/ 2	24/ 20	3.7558/ 3.3994
Anode-cathode distance: duration of VIRA	49,900/ 5,790	20.632/ 25.182	6/ 11	58/ 53	8.2172/ 9.8158
Anode-cathode distance: duration of MAINA	49900/ 5790	15.285/ 16.327	2/ 3	46/ 44	9.0693/ 10.025
Number of alumina doses in the automatic mode	49,900/ 5,790	6791.489/ 6922.808	5723/ 6185	7533/ 7409	317.8467/ 254.9121
Period in the manual mode	49,900/ 5,790	0.167/ 0.238	0/ 0	0.8/ 0.6	0.1227/ 0.1432
Period in starvation	49,900/ 5,790	6.402/ 7.107	2.6/ 3.1	10.2/ 9.2	1.2498/ 1.2347
Period in oversaturation	49,900/ 5,790	6.709/ 5.702	1.1/ 1.3	9.4/ 8	1.28/ 1.0982
Volume of the produced metal	49,900 / 5,790	4165.372 / 4096.7568	1760 / 2050	5040 / 4880	303.5068 / 398.0336
Cell voltage	49,900/ 5,790	3.7448/ 3.75	0.2853/ 3.0466	5.86/ 4.252	0.07/ 0.0554

5. Description of algorithms

Diagnostic models are built using a number of machine learning techniques. Most commonly, they include decision trees [5], ensembles of algorithms [6], artificial neural networks [7], etc. In this study, the prediction model features the ensembles of algorithms based on decision trees.

Gradient boosted trees is an algorithm which uses an ensemble of decision trees where each consecutive tree fits on the data on errors in the preceding decision tree. Gradient boosting involves the serial construction of algorithms, where each successive algorithm tries to compensate the errors in the composition of the previous ones. The resulting classifier is obtained as a linear combination of the classifiers. The optimum linear combination coefficients are found using a greedy algorithm, which implies the gradual addition of classifiers similar to the gradient descent. This study uses *XGBoost Classifier* (*XGBC*) [8] and *Catboost Classifier* (*Catboost*) [9].

Bagging on decision trees is an algorithm which applies the bootstrap technique to an ensemble of decision trees, with each of them built from the dataset generated from inputs. The classification result is defined by voting. This study applies *Balance Bagging Classifier* (*BBC*) [10] and *Balance Random Forest Classifier* (*BRFC*) [11].

The models are trained and tested using the generated dataset. The hyperparameters are tuned through random search with cross validation. The parameters for the model are selected by means of maximum precision. Tables 2-5 demonstrate the optimum values of the hyperparameters in each model.

Table 2
The hyperparameters of the *XGBClassifier* model

Parameter name	Description	Value
n_estimators	Number of gradient boosted trees	150
max_depth	Maximum tree depth for base learners	13
learning_rate	Learning rate	0.3

Table 3
The hyperparameters of the *CatBoostClassifier* model

Parameter name	Description	Value
iterations	Number of iterations	25
depth	Tree depth	13
learning_rate	Learning rate	1

Table 4
The hyperparameters of the *BalanceBaggingClassifier* model

Parameter name	Description	Value
n_estimators	The number of base estimators in the ensemble	290
max_samples	The number of samples to draw from X to train each base estimator	1
max_features	The number of features to draw from X to train each base estimator	1

Table 5
The hyperparameters of the *BalanceRandomForestClassifier* model

Parameter name	Description	Value
n_estimators	The number of trees in the forest	50
max_depth	The maximum depth of the tree	14
min_samples_leaf	The minimum number of samples required to be at a leaf node	1
min_samples_split	The minimum number of samples required to split an internal node	2
criterion	The function to measure the quality of a split	‘gini’

6. Analysis and comparison of the results

The model was fitted on the data divided into two sets, training and testing ones, in the following proportion: the 2020 data were used for training, while the 2021 data were used for testing. The results of the model training are presented in Table 6. The best results across the chosen metrics were shown by the *XGBoost Classifier* model.

Table 6
The classification results in the test dataset

	XGBC	*Catboost*	*BBC*	*BRFC*
accuracy (2)	0.99	0.96	0.97	0.96
precision	0.99	0.22	0.43	0.33
recall	0.85	0.54	0.93	0.99
F1 score	0.86	0.32	0.58	0.49

Figure 1 shows the confusion matrix for the *XGB Classifier* model. The rows show the actual classes, while the columns represent the predicted instances. The diagonally positioned elements show where the actual and predicted classes coincide. The total sum of all the values in the matrix cells represents the number of the test items.

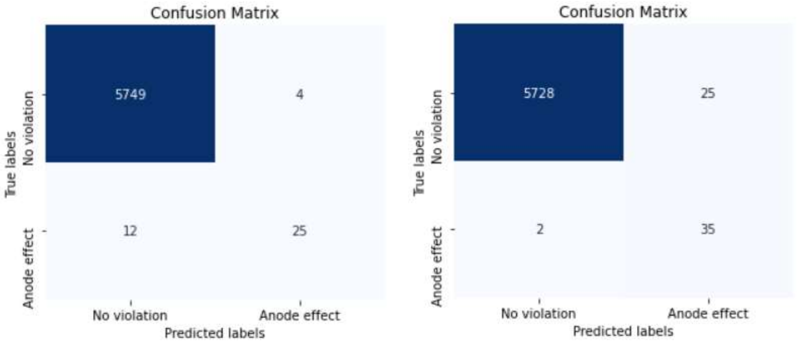

Figure 1: The confusion matrix for the XGB Classifier:
1a – threshold value 0.5, 1b – threshold value 0.3

As is seen in the matrix, in most cases the trained classifier detects the process disruptions correctly, with 12 type I errors (when the anode effects are not detected by error) and 4 type II errors (when the anode effects are falsely reported). When objects are classified according to one or another class, the commonly used threshold value amounts to 0.5. However, this value is not always optimum, for instance, as concerns the imbalanced data distribution in the inputs. The classifier threshold controls the ratio between the False positive and the False negative instances. To reduce type I errors (to increase the number of False positive instances and to reduce the False negative ones), the threshold value was decreased to 0.3 (Figure 1b). Therefore, the *XGBoost Classifier* model was selected for the purpose of predicting the anode effects, since the evaluation results suggest that the model quality proves suitable for practical use.

7. Conclusion

The article presents the results of the study aimed at developing a detection toolkit for process disruptions classified as anode effects based on ensembles of decision trees. The suggested models predict deviations in the process of aluminum production using the combined data of immediate and daily-average monitoring data. The method includes preprocessing the daily average inputs, aggregating the immediate and daily-average data, and building a math model. The study reveals the most informative parameters characterizing the current state of the facility, as well as how the deviations develop, which allows predicting the process disruptions. *XGBoost Classifier* stands out among other tested algorithms. The validated results suggest that the quality of this model is rather high for practical use. Additional research on the inputs is required to ensure higher accuracy of the prediction

8. References

[1] A. Tabereaux, Anode Effect and PFC Emission Rates, in: Proceedings of the Eighth Australasian Aluminium Smelter Techn, New Zealand, 2004, pp. 532–540.

[2] V. Yu. Bazhin, A. A. Vlasov, A. V. Lupenkov, Anode effect control on an aluminum electrolyzer, Metallurgist 5 (2011) 89–93.

[3] V. Yu. Bazhin, D. V. Makushin, A.V. Saitov, Conception of operating an aluminum electrolyzer without anode effect, Notes of the Mining Institute, volume 202, 2013.

[4] K. Zhou, G. Xu, S. Guo, Anode effect prediction based on support vector machine and K nearest neighbor, in: Chinese Automation Congress, 2017, pp. 341-345. doi: 10.1109/CAC.2017.8242789.

[5] L. Rokach, O. Maimon, Data Miningwith Decision Trees. Theory and Applications, London, World Scientific Publishing Co, 2008.

[6] C. Zhang, Y. Ma, Ensemble machine learning: methods and applications, USA, Springer, 2012.

[7] I. Goodfellow, Y. Bengio, A. Courville, Deep learning,. Cambridge, MIT press, 2016.

[8] XGBoost Classifier, URL: https://xgboost.readthedocs.io/en/latest/.

[9] Catboost Classifier, URL: https://catboost.ai/docs/concepts/pythonreference_catboostclassifier.html.

[10] Balance Bagging Classifier, URL: https://imbalancedlearn.org/stable/references/generated/imblearn.ensemble.BalancedBaggingClassifier.html.

[11] Balance Random Forest Classifier, URL: https://imbalancedlearn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html.