UTILIZATION OF DATA MINING METHODS FOR AUTOMATION OF IRON ORE RAW MATERIALS PRODUCTION PROCESSES

Tatiana Ivanova – Violetta Trofimova – Mariia Karelina – Ekaterina Kalinina

Abstract

Digital transformation of production is a today's reality. Automation of a production process implies creation of a system which will provide a real-time forecast of development of the current situation on the basis of data on the process and, thus, will decide optimal operating modes of any equipment. Such forecast should be based on a model. The range of available modeling techniques includes a great variety of methods, from statistical methods to Data Mining techniques. Furthermore, data collected on production processes often requires processing before it will be used in modeling. In this case, modeling is focused on the relationship between output of an active section of an iron ore crushing and concentrating mill and qualitative parameters of the base ore, parameters of production processes (the equipment used and the operating modes of such equipment) and the required quality of iron ore concentrate. This paper provides some results of modeling with Data Mining techniques, and decision trees in particular. It demonstrates advantages such models have in this specific field. It investigates how the proportion of observations in each class and parameters of pruning influence accuracy of decision tree models.

Key words: Data Mining, modeling, optimization, decision tree, crushing and concentrating mill

JEL Code: C1, Q55

Introduction

In our era of industrial automation, almost each and every industrial enterprise is using an automated production process control system. On the basis of a great amount of data on the production process which is transmitted online by sensors installed in such system, it is essential to make optimal decisions on process control to reduce costs and (or) increase the output and (or) improve quality of the products. A lot of papers are being published both in

Russia and abroad, which deal with automation and optimal production process control. The scope of use is rather wide, including mechatronic engineering (Rezchikov et al., 2019), agriculture (Leshchenko, 2020), road construction processes (Prokopyev et al., 2018), metallurgy (Gonzalez-Marcos et al., 2011), machine building (Sakthivel et al., 2012) and many other industries. There are also examples of use of various models for operational control of iron ore concentration processes at crushing and concentrating mills (Biryukov et al., 2013).

1 Statement of the Problem

Within the framework of the project of automation of production processes at an iron ore crushing and concentrating mill, a problem was set to define a model of optimal production process control utilizing methods of mathematical and statistical data analysis and Data Mining techniques. The problem of production process automation is highly topical since the demand for iron ore pellets is high both in the Russian market and all over the world. The optimal control model should provide for real-time forecasting of the influence the controlled parameters of a production process have on the output of the processing section and quality of iron ore concentrate, taking into account qualitative parameters of the base ore and composition of equipment in service. Such modeling is aimed at increase of the output of the process section, i.e. increase in volumes of concentrate produced per time unit, and (or) quality of products, i.e. percentage content of iron in the concentrate produced.

2 Modeling

The production process of a section of a crushing and concentrating mill is determined in a relevant flow chart. The output of a section, which is one of the target modeling parameters, can be expressed as follows:

$$Q = f(u_1, u_2, \dots, u_N, z_1, z_2, \dots, z_K, y_1, y_2, \dots, y_M), \qquad (1)$$

where N, K, M is the number of controlled production process parameters, qualitative parameters of the base ore, and qualitative parameters of iron ore concentrate, respectively, which are used in the model (Ivanova et al., 2019; Shnayder et al., 2020)

Thus, we need to calculate optimal values of controlled factors u_i at pre-set values of qualitative parameters of the base ore z_j and iron ore concentrate y_l, which ensure the maximum output of concentrate of at least the same quality. At the same time, process limitations should be taken into account:

$$u_{i_{min}} < u_i < u_{i_{max}} \qquad (2)$$

$$y_{j_{min}} < y_j < y_{j_{max}} \qquad (3)$$

where i=1…N, j=1…K.

The problem (1)-(3) will be solved taking into account the following process limitations: the operating mode of the process section selected by the type of produced concentrate; adequate quality of the concentrate produced; equipment available; iron loss at the tail end; consumption of grinding media (ball-mill and rod-mill media); and power consumption by process equipment.

The technical limitations listed above should be met taking into account the error of measuring instruments used.

3 Baseline Data

In the process of defining the control model, we used the following baseline data: statistical data on operation of the process section during several months in 2018–2019. We performed correlation analysis to select the following parameters out of several dozens of baseline parameters:

- Uncontrollable parameters (qualitative parameters of the base ore): Share of free-milling ore; Ore milling characteristics; Iron content in crushed ore; Granulometric composition of ore; Sulfur content in concentrate; Sulfur content in ore;
- Controlled parameters: Density of hydrocyclone overflow 1; Density of hydrocyclone overflow 2; Feed density of a wet magnetic separator 1; Feed density of a wet magnetic separator 2; Solid weight ratio in sands made by a magnetic deslimer 1; Solid weight ratio in sands made by a magnetic deslimer 2; Solid weight ratio in sands made by a magnetic deslimer 3;
- Target parameters (in terms of iron ore concentrate): Output of the process section, tons per h; Iron content in ore, %.

Before the data was used in training, it had been cleared of all peaks, periods in which there were no readings or when readings of some sensors were incorrect, and observations when output of the section was beyond the operating range. Furthermore, after revision of statistics on the parameters we concluded that it was required to distinguish two situations by different types of produced concentrate.

4 Calculations

We considered possible utilization of the following methods of defining control models on the basis of statistical data: Least square method (definition of a regression model); Classification models based on discriminant analysis methods; Multiple choice model based on logistic regression; Neural networks; Decision tree technique.

Earlier (Ivanova et al., 2019) it had been determined that the decision tree model would be the optimal choice to solve this problem. Decision tree models have some advantages, such as high quality of classification; ease of interpretation of obtained results in defining probabilistic recommendations for selection of particular values of parameters of the iron ore milling process; automatic selection of most significant input attributes; a possibility to define non-parametric models and, therefore, to solve Data Mining problems in which there is no a priori information about the type of relationship between the data investigated; high accuracy of models created using decision trees, which is comparable to the accuracy of other methods of defining classification models (e.g., highly accurate neural network models still cannot be interpreted due to everything included to the "black box"); and much less time required to define classification models using decision tree algorithms compared to model training based on other methods.

We performed a series of experiment calculations which showed that accuracy of the decision tree model depended on the following: The number of observations in the training set; The model structure; The proportion of observations in each class; and The pruning parameters.

Since the number of observations within the load range from 295 to 320 tons per hour was small, we decided to model the load value starting from 323 tons per hour, having divided the entire range, in a first approximation, into 4 groups: 1st Group: 323 to 337 tons per hour; 2nd Group: 338 to 352 tons per hour; 3rd Group: 353 to 367 tons per hour; 4th Group: 378 to 382 tons per hour.

Decision tree models were defined on the basis of Classification and Regression Tree (CART) which is one of the state-of-the-art algorithms in the field. (Breiman et al., 1984; Graham, 2011) Calculations are performed in the R programming language. (R-Project.org)

Figures 1 and 2 show results of our experimental calculations, which provide for evaluation of the sample size and the model structure defining accuracy of the forecast. Figure 1 shows values of production load forecast accuracy for the four target groups, which were obtained using 2 binary trees, i.e. at first, we used the entire range of production load and then, accordingly, we divided the training sets in two (less than 252 and more than 252 tons per hour). The two binary trees were defined for those two particular cases.

The white column represents forecast accuracy in the control sample, i.e. in the set of examples which were not used in model training. Low accuracy is observed in marginal minority groups (the first and the fourth one). This is an issue in any classification model, since if the general error of a model is calculated as a sum of errors in all observations, the model is forced to adapt to the majority class to a greater extent.

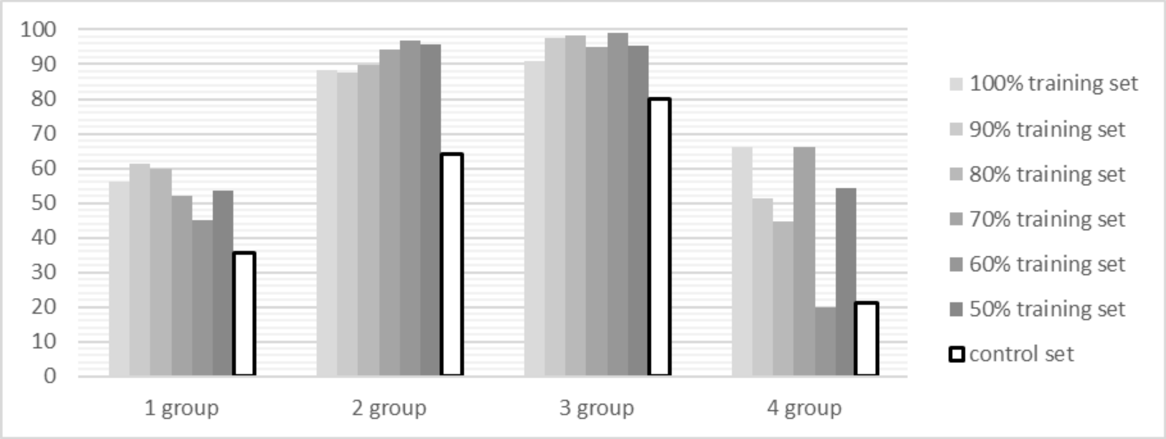
This issue can be solved either by duplication of minority class observations or by exclusion of majority class observations. Unfortunately, each of the said two methods has its weak points. The first one entails a high risk of model re-training on duplicated examples, i.e. the model "fails to learn" to recognize observations beyond the training set; in other words, it will not acquire ability to generalize. The reduced sample of the second method may turn out to be non-representational, and (or) the total number of observations may become inadmissibly small to adequately train the model.

There is another approach to this issue – in addition to the widely used CART decision tree algorithm, we could use modified machine training techniques which take into account both the maximum information gain in the process of branching and the difference between the "cost" of classification errors in different groups of observations. For instance, we could use the EG2, CS-IS3 and IDX algorithms. In our problem, however, it is impossible to estimate the cost of classification errors.

Having evaluated changes in forecast accuracy in the training sample depending on the sample size, he concluded that accuracy depends on the proportion of examples of different groups of the training sample rather than on the sample size. Columns of different shades of gray in Figure 1 represent situations with different numbers of observations in the training set within the range from 2208 observations (100%) to 1104 observations (50%). The white column represents accuracy in the control set.

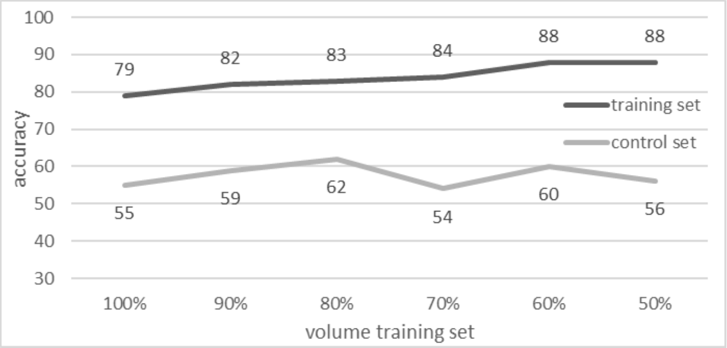
Overall accuracy of the decision tree model in the training set and the control set is shown in Figure 2. The overall accuracy grows as the training sample reduces, which is, in the first instance, preconditioned by growth in accuracy in the majority groups, while the decrease in accuracy in the marginal (minority) groups is drastic. Accuracy in the control set remains almost the same suggesting that the half-sample is sufficiently representational. The overall low accuracy in the control is attributable to insufficiency of parameters defining ore fed to the milling section or to poor quality of such parameters.

Fig. 1: Relationship between accuracy of the target parameter forecast based on two binary decision trees and the training sample size in the four target groups



Source: author’s own work

Fig. 2: Accuracy of decision tree-based forecast in the training set and the control set at different sizes of the training sample



Source: author’s own work

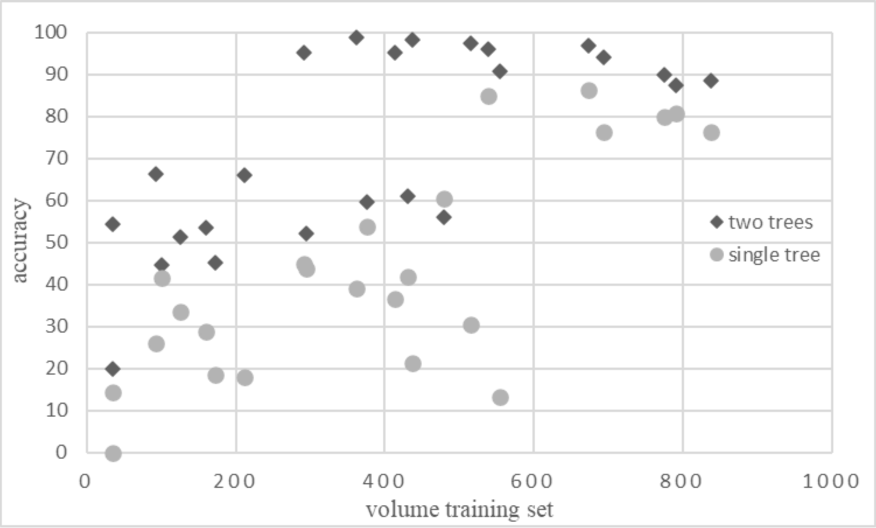
In addition to the training sample size, we investigated influence of the model structure on accuracy of the result. We analyzed a decision tree model in which the output variable has four possible values instead of two, matching the four groups distinguished by production load. These results are much worse than the previous ones as we can see that the 3rd and 4th groups of the control set were not recognized at all.

On the whole, accuracy of forecasts based on two binary trees and on one tree can be compared using Figure 3.

In average, accuracy with two trees is 73%, while accuracy with one tree is 44% with dispersions of 508 and 628, respectively. Validation of significance of a hypothesis of equality of selected means using non-parametric methods showed that this hypothesis is rejected at the significance level less than 1%, i.e. at this level a significant difference is detectable in accuracy of the decision trees. This applies particularly to small samples.

We then determined relationship between accuracy of the decision tree model in the training set and the control set and the tree pruning method.

Fig. 3: Accuracy of forecasts in the training set based on two binary trees and on one tree at different sizes of the training set



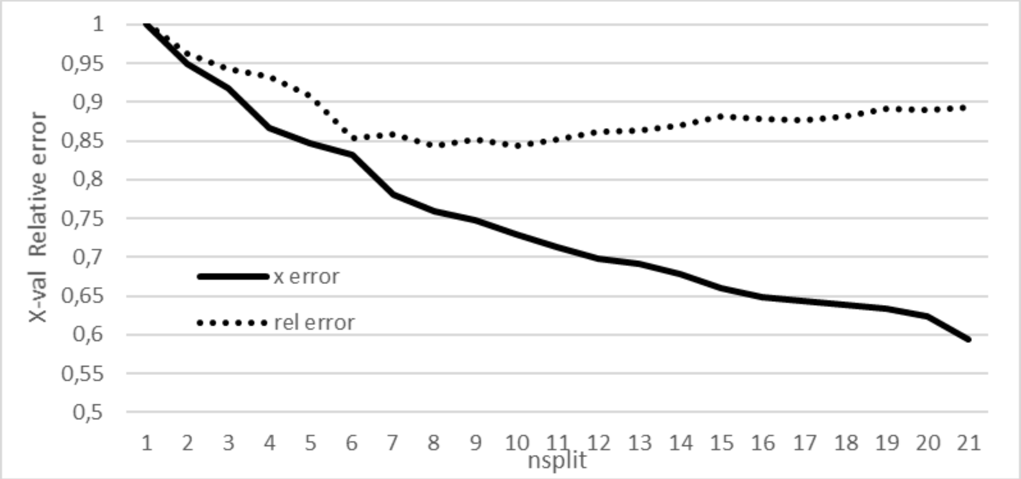
Source: author's own work

It is well-known that accuracy of a decision tree model and its value are in inverse relation. The maximum possible value of a tree has one example on each leaf and is absolutely out of further use, since the rules defined are of low significance. It is essential to find the golden mean in the process of pruning, i.e. to prune branches with the lowest classification capacity first. The CART algorithm distinguishes different subtrees and selects those which will provide the smallest classification error in a test set. Candidates for pruning are selected on the basis of the adjusted error rate, i.e. a "fine" for complexity of the tree is added to the error of a particular branch in the training set.

In our opinion, an error in a test set is not an objective index of model accuracy, in spite of the fact that the tree definition procedure implies validation of model accuracy in a test set using the cross-validation procedure. For instance, in our calculations we used tenfold cross validation. In addition to the training error (the error in the training examples) and the generalization error (the error in the test set), model accuracy shall be evaluated in the control set elements of which were used neither in training nor in testing of the model.

Figure 4 demonstrates relationship between the training error (x error) and the generalization error (rel error) of a binary decision tree and its size (nsplit). We can see that branching should be stopped in a range of about 6 to 15 leaves since, further decrease in the training error with concurrent increase in the generalization error suggests re-training of the model.

Fig. 4: Relationship between the decision tree training (x error) and the generalization error (rel error) and its size (nsplit)



Source: author’s own work

In order to make a decision on where exactly the tree should be pruned, we used 2 approaches: 1) the level matching the minimum error in the test set. However, the problem is, this error in the test set tends to change with occasional separation into the training set and the test set; 2) the level at which the error in the test set becomes less than the minimum error plus root mean square error deviation. This method takes into account variability of the error in the test set which emerges in the process of cross-validation.

Tab. 1: Mean model accuracy in the test set and the control set

	Mean model accuracy in the test set	Mean model accuracy in the control set
Tree with excessive number of branches	84.7	55.4
Tree pruned using the 1st method	80.3	59.3
Tree pruned using the 2nd method	78.0	60.5

Source: author’s own work

We performed a series of calculations and defined 828 binary decision trees in three variants: a tree with a deliberately excessive number of branches, a tree pruned using the 1st method, and a tree pruned using the 2nd method. Results of mean accuracy of the models in the test set and the control set are provided in the table below.

Conclusion

The decision tree method having some obvious advantages is the optimal method for solving problems related to modeling of performance of the process section of a crushing and concentrating mill. This study showed that the following factors influenced quality of the models we defined: the number of observations in the training set and the proportion of examples of various groups in this set; the model structure, since we found that several separate binary trees defined in different ranges of the target parameter are more efficient than a single tree with multiple values of the target parameter; and the pruning method. With regard to the latter, we concluded that pruning using the second of the proposed methods contributes to increase in accuracy of trees in the control set; and, thus, we recommend using it in spite of concurrent decrease in accuracy of trees in the test set.

References

Biryukov, V.V., Oleynik, A.G., Opalev, A.S., Shcherbakov, A.V. (2013) Modernization of iron ore miling techniques at OJSC “OLKON” utilizing imitation modeling. *Proceedings of Kola Science Center of the Russian Academy of Sciences*. 5 (18). 183-188.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A. (1984) Classification and regression trees. Boca Raton: CRC Press, 366 p.

Gonzalez-Marcos, A., Alba-Elias, F., Castejon-Limas, M., Ordieres-Mere, J. (2011) Development of neural network-based models to predict mechanical properties of hot dip galvanised steel coils. *International Journal of Data Mining, Modelling and Management (IJDMMM),* 3(4), 389-405. DOI: 10.1504/IJDMMM.2011.042936

Graham, W. (2011) Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, 395 p.

Ivanova, T.A., Trofimova, V.Sh., Shnayder, D.S., Kalinina, E.A. (2019) Optimization of parameters of iron ore raw materials production processes: comparative analysis of combined mathematical and statistical models. *Applying Mathematics in Economic and Technical Studies: Collection of proceedings of international scientific and practical conferences under general editorship of V.S. Mkhitaryan*. 90-101.

Leshchenko, N. (2020) Digital development of agro-industrial organizations in Russia. *The 14th International Days of Statistics and Economics. Conference Proceedings,* 653-662. DOI 10.18267/pr.2020.los.223.0

Prokopyev, A.P. et al. (2018) Implementing the concept of automation and intellectualization of road construction process control. *Bulletin of Moscow State Construction University.* 13(1 (112)). 61-70. DOI: 10.22227

Rezchikov, A.F., Kushnikov, V.A., Ivashchenko, V.A. et al. (2019) Welding process control in robot-aided process systems by the criterion of product quality. *Mechatronics, Automation and Control*. 20(1). 29-33. DOI: 10.17587/mau.

Sakthivel, N. R., Nair, Binoy B., Sugumaran, V., Roy, R.S. (2012) Application of standalone system and hybrid system for fault diagnosis of centrifugal pump using time domain signals and statistical features. *International Journal of Data Mining, Modelling and Management (IJDMMM)*, 4(1), 74-104. DOI: 10.1504/IJDMMM.2012.045137

Shnayder, D.A., Kalinina, E.A. (2020) Automated System-Adviser Based on a Model for Control of the Technological Process of Concentrate Production. *Global Smart Industry Conference (GloSIC)*. 335-341. DOI: 10.1109/GloSIC50886.2020.9267859.

Contact

Tatiana Ivanova
Nosov Magnitogorsk State Technical University
38, Lenina st., Magnitogorsk, Russian Federation, 455000
jun275@mail.ru

Violetta Trofimova
Limited Liability Company "Belka Digital"
89, Lenin prospekt, Chelyabinsk, Russia, 454080
violat@mail.ru

Mariia Karelina
Nosov Magnitogorsk State Technical University
38, Lenina st., Magnitogorsk, Russian Federation, 455000
marjyshka@mail.ru

Ekaterina Kalinina
South Ural State University (National Research University)
76, Lenin prospekt, Chelyabinsk, Russia, 454080
kalininaea@susu.ru