Machine learning in the development of regulators: the specifics of data preparation and the choice of the ML model algorithm

N A Staroverova¹, M L Shustrova and M Z Fatykhov

Kazan National Research Technological University, 72 Karl Marx St, Kazan, 420015, Russian Federation

¹E-mail: nata-staroverova@yandex.ru

Abstract. The article considers the stages of regulators development based on machine learning. Machine learning, like a fuzzy logic, is used in PID regulators in two ways: it is able to build the regulator or a unit for adjusting its coefficients. A feature of machine algorithms is the "learning" ability which allows you to transfer the experience of an expert to the system. The presence of a learning procedure distinguishes the configuration of regulators based on machine algorithms from the classical approach. The publication analyzed the possibility of using regression machine algorithms (models) of the Scikit-learn library (Python) to solve control problems using the example of a level regulator. The search for an effective data processing model led to the construction of ensemble algorithms based on stacking. The high degree of compliance between predicted and test data confirms the accuracy and stability of the results. The results presented in the article can be used in a simulating or in developing a controller based on a machine algorithm.

1. Introduction

The economic development of modern society is impossible without the development of industrial production. It is important to ensure an increase in the volume of manufactured products, an increase in their quality characteristics, as well as ensuring the safety of the operation of enterprises.

A significant contribution to the achievement of these indicators is made by the development of production technologies, hardware and process control technologies.

An important role in ensuring the quality of control of technological processes is played by the laws of regulation, which are the basis for the functioning of control systems. Historically, proportional controllers were the first to be created. Soon it was supplemented with integral and differential components and PID controllers were appeared. Despite the fact that operation principles of PID controllers were developed at the beginning of the twentieth century, this type of regulator is still the most widely used in automated control systems. The popularity of PID controllers is due to their reliability, functionality, and simplicity. Other advantages are the small amount of overshoot and the relatively high speed of reaching the set point. However, there are also unresolved problems. It is necessary to eliminate the integral saturation. Controlling objects with hysteresis and nonlinearities is not always effective. The issues of automatic tuning and adaptation are worthy of special attention [1-3].
The growing level of development of modern technologies forms the trend of introducing intelligent technologies into production management processes. With an increasing in the volume of information flows, the degree of complexity of their processing are also increasing. The dynamics of modern control systems are often unknown. Processes requiring regulation cannot be called independent. The measurement results of key technological parameters are often very noisy, and the load on the system is not constant, which additionally causes certain difficulties. Often there are tasks that require non-standard solutions. These problems cannot always be solved by conventional algorithmic methods, and it is often impossible to find a clear solution algorithm. It is advisable to solve problems such as monitoring, control and prediction of process parameters, the direct measurement of which turns out to be expensive, difficult to implement or impossible. These factors determine the relevance of the development of adaptive regulators. Adaptive control can provide monitoring of regulator settings of continuously changing properties of control object, which allows improving quality of control. For example, in highly inertial processes, it is advisable to carry out operational correction of control signals long before changing process parameters.

Artificial intelligence technologies (machine algorithms, artificial neural networks, deep learning) are one of the key areas of development in this context. These technologies have become widely used both for the implementation of predictive analysis functionality and for ensuring optimal regulation of process parameters [4-5].

The relevance of a regulator based on artificial intelligence lies in its ability to provide maximum peak performance, high readiness for operation, the ability to predict events, as well as convenience in operation and maintenance, which is a significant advantage. One of the difficulties in developing adaptive regulators is the high degree of individualization of each regulator. The creation of an adaptive regulator includes the stages of determining a data set and parameters for conducting training, selecting a data processing model, conducting direct training, testing and implementing the regulator in a functioning system.

The purpose of this work is to determine an effective machine learning algorithm, as well as a data set and parameters for investigating the possibilities of machine learning in regulating technological processes.

2. Selecting and preparation data for learning
Regulator is designed for automatic level control in the separator condensate collection compartment (figure 1), which is a part of the complex gas treatment unit. Hydrocarbon liquid is fed to separator R-2 where it is separated into weathering gas, light phase flow and water-methanol solution. The adjustable parameter is the level in the separator controlled by the level meter LT-1 [5].

![Figure 1. Regulation object scheme.](image-url)

The analysis is carried out on data accumulated in the specified area during the year. The data for analysis is divided into two sets - for training and for testing.
According to classical control theory, level control is calculated based on the mismatch signal between the regulator set point and the current level value. However, in order to implement the machine learning strategy, it is advisable to consider at the initial stage all the technological parameters available for measurement in this area.

Training the model on a primary experimental data often does not produce the expected result. This fact can be explained by several reasons. The data for learning may contain inconsistent experimental values, i.e. the same values of independent parameter vectors have different values of dependent parameter vectors. Measurement errors and subjective errors can also lead to inconsistencies [6-8].

Another cause is a model parameter selection error if the model does not include meaningful input parameters. Such data strings affect learning speed, the magnitude of learning errors and generalizations, and may ultimately lead to inoperability of the constructed machine learning model. Therefore, this data should be excluded from the training sample. The values in the source database can be divided into three components: noise, non-informative values, and datum objects. The main task in data pre-processing is to eliminate noise and non-informative values and bring the reference parameters into normal form.

The process of excluding parameters is divided into two stages: the elimination of noise emissions and the elimination of uninformative peripheral objects. The process stops when only datum objects remain, the exclusion of which significantly increases the functionality. The correlation method allows you to clearly determine the dependence of the input, transition, and output values. It can be used to estimate the influence of parameters. The primary data analysis made it possible to determine the effect of each of the analyzed parameters on the valve opening degree. Diagrams of some parameters are shown in figure 2.

![Diagram](image.png)
Several dataframes with different sets were formed to analyze the joint effect of parameters on the output. In the next step this data frames were tested on regression models suitable for the forecasting problem. The influence of the analyzed technological parameters on the resulting variable (the state of the control valve) was assessed by analyzing their correlation (table 1). A correlation coefficient below 0.5 indicates almost no influence of the parameters on each other [9-10].

The process state of the valve depends directly on the level reading: the maximum level value corresponds to the maximum degree of valve opening. However, the heat map shows that this relationship is not so obvious between the valve and the level readings. However, there is a high correlation with hydrocarbon condensate flow from the separator. This is natural: this flow rate determines the increase in the level.

### Table 1. Correlation between valve condition and sensor readings.

|       | LT-1  | TT-2  | FT-2  | FT-3  | FT-1  | TT-1  | V-1  |
|-------|-------|-------|-------|-------|-------|-------|------|
| LT-1  | 1.0000| -0.1519| 0.6333| 0.3282| 0.5994| -0.2509| 0.6150|
| TT-2  | -0.1519| 1.0000| -0.4830| -0.5075| -0.2395| 0.8295| -0.5683|
| FT-2  | 0.6333| -0.4830| 1.0000| 0.6293| 0.4743| -0.5370| 0.9663|
| FT-3  | 0.3282| -0.5075| 0.6293| 1.0000| 0.0613| -0.4642| 0.6548|
| FT-1  | 0.5994| -0.2395| 0.4743| 0.0613| 1.0000| -0.3645| 0.4978|
| TT-1  | -0.2509| 0.8295| -0.5370| -0.4642| -0.3645| 1.0000| -0.5947|
| V-1   | 0.6150| -0.5683| 0.9663| 0.6548| 0.4978| -0.5947| 1.0000|

### 3. Instruments and methods

Before starting work, it is advisable to normalize the data and bring them to a single dimensionless scale. There are several methods of normalization: decimal scaling, minimax normalization, and normalization using standard deviation.
Decimal scaling means dividing the original numbers by the maximum number of digits from the sample. The disadvantage of this method is the absence of any clear range of numbers. It will always spread up or down.

Minimax normalization eliminates the drawback of the previous example by bringing it to a certain interval \([a, b]\). Basically, intervals \([0, 1], [-1, 1]\) are taken for its further use, for example, in activation functions in neural networks. This method is effective with relatively uniform value distributions. The presence of abnormal emissions in the sample can affect the entire distribution, even if these emissions do not significantly affect the dataset.

Normalization by standard deviation solves this problem by introducing an estimate of mean and variance. In this case, the average value and standard deviation of the sample are calculated, and then the data is converted for each individual sample value.

In the work, all three methods considered were analyzed to determine the method that is most effective within the given task.

The tools for solving the prediction problem are regression algorithms of the Scikit-learn library of the Python development environment. Linear algorithms were used in the task are Linear Regression ('LR'), Ridge Regression ('R'), Lasso Regression ('L') (LASSO — Least Absolute Shrinkage and Selection Operator), Elastic Net Regression ('ELN'), Least Angle Regression (LARS) ('LARS'), Bayesian ridge regression ('BR').

Non-linear algorithms were also used. There are k-nearest neighbors regressor ('KNR'), Decision Tree Regressor ('DTR'), Linear Support Vector Machine – Regression / ('LSVR'), Epsilon-Support Vector Regression ('SVR') [9-13]

Ensemble algorithms considered in the research are AdaBoost Regressor ('ABR') (AdaBoost = Adaptive Boosting), Bagging Regressor ('BR') (Bagging = Bootstrap aggregating), Extra Trees Regressor ('ETR'), Gradient Boosting Regressor ('GBR'), Random Forest Classifier ('RFR').

Each model was tested on four datasets. Each dataset was used both without normalization and with three methods of normalization. Testing was carried out to determine an algorithm that would show consistently accurate results and be resistant to data noise.

Model testing results were performed by estimating the determination coefficient. For acceptable models, it is assumed that the determination coefficient should be not less than 50%. In this case, the absolute value of the multiple correlation coefficient exceeds 70%.

Models with a determination coefficient above 80% can be considered statistically valid. The variable is exactly described by the model if the determination coefficient is close to one.

4. Results

The efficiency of each algorithm is estimated by cross-validation. During testing, it was revealed that almost all models show accuracy above 0.8. The exception was Least Absolute Shrinkage and Selection Operator/Lasso Regression ("L") and Elastic Net Regression ("ELN"). The Normalizer method, in turn, is not effective for Lasso Regression ("L") and Elastic Net Regression ("ELN"). It shows satisfactory results for other models.

The algorithms selected for further work meet a number of requirements: a stable and high result greater than 0.9, both on raw data and with all methods of normalization, and with any set of parameters. These features indicate a good generalizing ability of algorithms and their resistance to noise and outlier, which is an important property.

Such algorithms include linear algorithms - Least Angles Method ("LARSCV"), Bayesian ridge regression ("BR"), nonlinear algorithms - k-closest neighbour method (regression) ("KNR"), Regression trees ("DTR"), Reference vector method (regression) ("SVR") and ensemble algorithms - ("BR") (Bootstrap aggregating), Extra Trees Regressor, Gradient Boosting Regressor ("GBR"), Random Forest Classifier ("RFR").

The accuracy in 0.9 is considered quite sufficient. Therefore, an ensemble algorithm Random Forest Classifier ("RFR"), which showed an average determination coefficient of 0.97, could be recommended to predict the degree of valve opening in this process scheme. However, in addition to accuracy, it is
advisable to achieve greater stability of the prediction process in critical situations, for example, at
outlier points in parameter readings. Stacking allows to combine several weaker algorithms and train a
final estimator on their results. Any algorithm, or neural network, can be used as a final estimator. Linear
and nonlinear regression algorithms ('LARSCV', 'BR', 'KNR', 'DTR', 'SVR') were used as an "ensemble
of weak students. As a final estimator, ensemble algorithms (table 2) with the largest determination
coefficients are adopted. Standard error (mse) is used to estimate the response to outliers, mae is an
estimate of absolute prediction error. The correspondence of the predicted data and the original
information is estimated by the determination coefficient.

Table 2. Staking test results with different final estimators.

| final estimator | determination coefficient | mae; mse |
|-----------------|---------------------------|---------|
| RandomForestRegressor (max_depth = None, n_estimators = 700), figure 3 - a | 0.988 | 0.2677; 0.1675 |
| BaggingRegressor (n_estimators=700), figure 3 –b | 0.987 | 0.2694; 0.1753 |
| ExtraTreesRegressor (n_estimators=700), figure 3 –c | 0.988 | 0.2704; 0.1686 |
| GradientBoostingRegressor (n_estimators=700) figure 3 –d | 0.988 | 0.2695; 0.1648 |

All ensemble sets were analyzed with different levels of noise, i.e. with 3 sets of parameters tied to
one valve and showed satisfactory results (figure 3), not much different from the results in table 1. The
influence of the “set” of parameters on the accuracy of the results will be discussed in more detail in a
separate article.

Figure 3. Staking test results with different final estimators: agreement of the predicted (red) and test
(green) predicate.
The coefficient of determination of the obtained ensemble sets is 0.98, which shows the sufficient predictive ability of the model. In addition, the absolute and root-mean-square errors are small, which proves that most of the model's predictions correspond to the actual values of the test sample. This effect is also noticeable in figure 3: small deviations are present only at the peak values of the predicted and testing data, in other positions there is a coincidence of parameters.

5. Conclusion
The article provides analysis of annual cycle data of gas condensate treatment shop separators unit. The data were cleared and prepared for machine learning. The results obtained by testing 12 regression algorithms conclude that machine regression algorithms show good accuracy and reproducibility in predicting the degree of valve opening within the set task.

In this example, ensemble stacking with linear and nonlinear regression algorithms (‘LARSCV’, ‘BR’, ‘KNR’, ‘DTR’, ‘SVR’) turned out to be the most effective. Among final estimators, the best results were demonstrated by a final estimator including ‘Random Forest Regressor’, ‘Bagging Regressor’, ‘Extra Trees Regressor’, ‘Gradient Boosting Regressor’. In each case, we have a determination coefficient of 0.98. These results will be used to analyze the effect of various parameters on the stability of the model forecast, as well as to build regulators based on them.

References
[1] Jimmi W Key 2016 Control engineering Russia 63(3) 62-6
[2] Jimmi W Key 2016 Control engineering Russia 64(4) 106-10
[3] Gaynullin R, Zatsarinnaya Y and Staroverova N 2019 IOP Conference Series: Materials Science and Engineering 537(3) 032019
[4] Mukhamadiev R, Staroverova N and Shustrova M 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 1-8
[5] Shustrova M, Fafurin A and Baytimirov A 2019 International Conference on Industrial Engineering Applications and Manufacturing ICIEAM 8743033
[6] Trawiński B, et al. 2012 International Journal of Applied Mathematics and Computer Science 22 867-81
[7] Gibadullin R, Baimukhametova G and Perukhin M 2019 International Conference on Industrial Engineering Applications and Manufacturing ICIEAM 8743064
[8] Botchkarev A 2019 Interdisciplinary Journal of Information, Knowledge & Management 14 35
[9] Järvelä Sanna, et al. 2020 British Journal of Educational Technology 51 2391-406
[10] Kosarevskaia S A, et al. 2021 IOP Conference Series: Materials Science and Engineering 1047(1)
[11] Yang Qijiu 2020 IOP Conference Series: Materials Science and Engineering 782(3)
[12] Oh sugi Hideharu, et al. 2017 Scientific reports 7(1) 1-4
[13] Kangash A, Maryandyshiev P, Zatsarinnaya Y and Volkova M 2019 IOP Conference Series: Materials Science and Engineering 643(1) 01215