Analysis of machine learning methods to improve efficiency of big data processing in Industry 4.0

A A Prudius\(^1\), A A Karpunin\(^1\), A I Vlasov\(^1\)

\(^1\)Bauman Moscow State Technical University, 5/1, 2-nd Baumanskaya Str., Moscow, 105005, Russian Federation

Abstract. The article considers the basic methods of machine learning applied by individual entrepreneurs within the framework of transition to digital production to improve the efficiency of data processing, classification of existing and would-be customers and their subsequent work with them. The main attention is paid to the problem of increasing the effectiveness of methods of machine learning applied for solving the current questions. Areas of application of technology are shown. The peculiarities of machine learning are briefly analyzed. The main features and prospects of the development of Machine Learning services are shown on the basis of the concept of a step-by-step combination of the methods under consideration. One of the main algorithms for working with data is analyzed; its main features, scope and procedure are described. Recommendations are given for the further use of machine learning algorithms. The role of machine learning in the development of modern science and industry is analyzed, the main tendencies of the industry development are determined, and the practical application of big data is shown. As part of the transition to Industry 4.0, the main areas of application of machine learning, big data, Artificial Intelligence and their relations with the corresponding fields of science and production are described. The article also offers a review of the application of Artificial Intelligence and machine learning in particular in the context of the transition to digitalization and the issues of individual entrepreneurship.

1. Introduction

The issue of analyzing the processing of large amounts of data has to be addressed starting from the earliest stages in the development of any applied task. Transitioning the development of modern industry to Industry 4.0 [1], which primarily implies the implementation of cybernetics and AI (Artificial Intelligence) and its components in different processes, which includes raising radically new questions in development. So today implementing machine learning and using big data are an integral part of efficient work in every organization.

Machine Learning (ML) is one of the most relevant directions in the development of modern technology, which represents methods of artificial intelligence aimed at training systems in the course of practical resolution of a number of applied tasks. When considering machine learning as a task, the task can be formulated as a generalization of classical approximation tasks, since in real-life of the applied projects the data can be non-numeric, incomplete and heterogeneous. Meanwhile, many machine learning methods are closely related to retrieving information and data mining [2].

The relevance of this research stems from the fact that the extensive growth of information volumes in recent times results in the emergence of Big Data as a standalone class of information processed,
which requires using networked computing capacity to process, as it is beyond the human mind to handle these amounts of data.

Machine learning currently has a broad range of applications [3]. The applications of machine learning are expected to continue growing due to ubiquitous digitization and accumulation of huge amounts of data in science, industry, transportation and business sectors of the society [4].

The work offers a methodology to analyze customer base in financial and insurance organizations, in order to classify existing and prospective clients using several machine learning algorithms. Its scientific novelty comes from using the sequence of information processing and applying the most efficient combinations of different methods.

The results of the study are practically relevant, as they increase the precision of the data obtained during the analysis at each level, which can further be applied in building corporate behavior strategies.

2. Literature review
This work reviews the issue of analyzing the efficiency of data analysis and applying radically new methods of dealing with data. Statistical analysis methods are used worldwide, the only thing that changes with time is the technology, which is currently based on machine learning. By using advanced analytics, financial organizations obtain practical applications of universal statistical methods and as a result gain valuable insights on their customer base, peculiarities of customer behavior and causal relationships in their interaction with the business.

Machine learning generally is one of the most common methods of research used by artificial intelligence. In the process of ubiquitous implementation of cybernetic systems in various areas of human life and developing Industry 4.0, AI plays one of the most important roles in the development of modern science and industry, thanks to the broad range of applications [5].

Some of the key practical applications of artificial intelligence include such trendy technologies as BigData [6], Internet of Things (IoT) [7], Virtual Reality/Augmented Reality (VR, AR) [8], and distributed ledgers (Blockchain) [9-12]. These and many others help to form a new stage in the development of public conscience and scientific development, directed to a broad range of theoretical and practical tasks addressed on a daily basis.

3. Methods
The key methods of machine learning [10] examined in this work for addressing the tasks at hand are as follows:

Clustering, the goal of which is to break down a multitude of various objectives into groups (called “clusters”) that contain “similar” objects with the maximum possible differences between the clusters. Clustering is the most general method of machine learning of those considered in this article; its main difference from classification is that the list of groups resulting from the breakdown is not fixed and is defined in the process of work (unsupervised learning) [11];

Classification, which is required to break down a finite number of objects into predefined classes by one or more parameters (supervised training);

Regression, which involves studying the impact of one or more independent variables on the dependent variable. Regression is the most accurate data analysis method.

One of the most common algorithms in machine learning, as used in addressing applied tasks, is the logistic regression of input parameters, which is characterized by lack of predicted value of numeric variables from the sampling of input data [13]. Instead of returning a certain numeric value, the function determines the probability of the value being analyzed matching a certain class of system parameters. For clarity, let us suppose only two classes are being analyzed, and the probability to be determined, \( P_+ \), is the probability of a certain value belonging to the class “+”. Respectively, \( P_- = 1 - P_+ \) for all other cases. Thus, the output of logistic regression is always in the interval of \([0, 1]\).

The idea of logistic regression is that a multitude of input parameters can be separated by a plane (in case of two dimensions – with a straight line with no inflection points) into two areas corresponding to the classes of output data. The resulting (see Fig. 1) plane is called ‘linear discriminant’, as from the
viewpoint of its function it is linear and allows the input parameters to be discriminated into different classes [14].

**Figure 1.** Visual representation of discrimination by logistic regression

In the simplest case, when considering a two-dimensional space, for a point with coordinates \((a, b)\) the logistic regression algorithm will include the following steps:

1. Calculate the value \(\beta_0 + \beta_1 a + \beta_2 b\) of the boundary function (or odds ratio function) identified as \(t\);
2. Calculate the odds ratio: \(OR_+ = e^t\) (as \(t\) is a logarithm);
3. Once the value \(OR_+\) is obtained, calculate \(P_+\) using a simple dependency: \(P_+ = \frac{OR_+}{1 + OR_+}\).

After getting the value \(t\) in p.1 and combining the remaining steps, we get \(P_+ = \frac{e^t}{1 + e^t}\). The formula to the right of the equation is the logistic function, which lends its name to the method of analysis.

By combining this algorithm with such ML methods as gradient boosting and random forest, we can improve the accuracy of input value classification, increasing precision of data analysis.

The implementation of algorithms of machine learning is possible in virtually any programming language. The most popular are the following ones: Python; R; Scala; Java.

Each of the listed programming languages implies the availability of its own tools for working with machine learning algorithms. In the framework of this paper, it is proposed to consider some libraries of the Python language as the most common ones.

One of the main libraries is NumPy; it is a Python library that adds support for large multidimensional arrays and matrices together with a large library of high-level mathematical functions for operations with these arrays.

Mathematical algorithms implemented in Python often work much slower than the same algorithms implemented in compiled languages (for example, Fortran, C, Java). The NumPy library provides implementations of computational algorithms (in the form of functions and operators), optimized for working with multidimensional arrays. As a result, any algorithm that can be expressed as a sequence of operations on arrays of data and implemented using NumPy is equivalent in speed to the code processed in the MATLAB environment.

To visualize the results of work, libraries such as Matplotlib, Bokeh, Plotly, Scikit are used. For example, Matplotlib is a Python library for building qualitative two-dimensional graphs. Matplotlib is a flexible, easily configurable package that along with the main language libraries provides capabilities similar to the MATLAB environment. At present, the package works with several graphics libraries, including wxWindows and PyGTK. Simple three-dimensional graphics can be built using the toolkit.
mplot3d. There are other sets of tools: for cartography, for working with Excel, utilities for GTK and others.

There is also a class of libraries specially designed to work with machine learning, including Theano, TensorFlow, Keras, etc. Theano is an extension of the Python language that allows you to efficiently compute mathematical expressions containing multidimensional arrays. The library provides a basic set of tools for the configuration of neural networks and their learning. Theano has received the greatest recognition in the tasks of machine learning when solving optimization problems. The library allows you to use the GPU features without changing the program code, which makes it irreplaceable when performing resource-intensive tasks.

4. Results and Discussion

In practice, solving applied tasks by using machine learning methods as part of this work is reduced to the joint usage of the algorithms described above.

The first issue to address is data clustering. To do this, all characteristic parameters are put in an N-dimensional space, where N is the number of dimensions for, and the distance between anchor points (which represent customer characteristics) is calculated. This way the entire volume of data is segmented by the criteria needed in each specific case, for further targeted work with each group. As part of the first stage of data processing, we also segment input data based on a custom quantity of parameters being analyzed and the most detailed segmentation by the maximum number of factors, depending on the goals set. The peculiarity of this stage lies in obtaining the results of statistical analysis without indicating priority parameters for further data work.

When the resulting data are classified based on the correlation between data points and a zero variable, the factors of the information analyzed are weighed (e.g., paying capacity has the maximum correlation and is the target variable when analyzing paying capacity), based on which each segment is fit within one class or another. Multi-collinearity (see Fig. 2) also plays a role – when using multiple factors, their individual contribution to classification can change substantially, in presence of a strong correlation between the factors – even though an individual factor can have substantial weight, it can be reduced in combination with additional factors with high correlation.

![Figure 2. Results of correlation analysis of the variables](image)

Additionally, more efficient algorithms can be used to optimize performance during the data classification phase – e.g., gradient boosting, which is based on multiple decision trees and can learn from its own mistakes – when building each subsequent decision tree (see Fig. 3), it considers and corrects errors made in earlier iterations.
Finally, regression analysis methods are applied, which, unlike classification, help identify specific values of different factors, such as choosing an interest rate that is the most profitable for the company that still meets the customer requirements, considering all the factors and aspects of the given situation. Even without having specific values but being aware of all the parameters, one can give a highly accurate estimation of the figure that can be adapted for each specific case.

Thus, by using all methods of machine learning in sequence, an organization gets a highly detailed analysis of its customer base, which enables highly targeted work with different customer groups and each individual customer. This method is highly optimized for the business, by determining specific values on a case-by-case basis, following the overall targets formulated by the company.

5. Conclusion
The work studies key methods of machine learning, their peculiarities and application areas. The authors review algorithms for analyzing big data using ML, identify their optimum combinations for personalizing the approach to individual customers, improve the efficiency of a financial organization (e.g., reducing costs, improving conversion and margins) and optimizing and automating business processes within companies. It can be concluded that analyzing ordinary individual events in the context of big data improves the performance of an individual company operation and the industry in general, capturing the benefits that could not be accessed previously.

6. Acknowledgement
Some results of the work were obtained with the support of project No. 14.579.21.0142 UID RFMEFI57917X0142 within the framework of the federal target program “Research and development in priority areas of development of the scientific and technological complex of Russia for 2014-2020”.

References
[1] Akberdina V V, Tretyakova O V and Vlasov A I 2017 A methodological approach to forecasting spatial distribution of workplaces in an industrial metropolis Problems and Perspectives in Management 4(87) 50–61
[2] Hastie T, Tibshirani R and Friedman J. 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag) p 746

[3] Vlasov A I and Yuldashev M N 2018 Analysis of the methods for processing information from a cluster of sensor *Sensors and Systems* **1(221)** 24–30

[4] Vlasov A I, Yudin A V, Shakhnov V A, Usok K A and Salmina M A 2017 Design methods of teaching the development of internet of things components with considering predictive maintenance on the basis of mechatronic devices *International Journal of Applied Engineering Research* **12(20)** 9390–9396

[5] Witten I H and Frank E 2016 Data Mining: Practical machine learning tools and techniques *Morgan Kaufmann* 7–9

[6] Vlasov A I, Novikov P V, and Rivkin A M 2015 Peculiarities of airtraffic planning using weather maps built with BIG DATA technologies *Bulletin of the Bauman Moscow State Technical University. Series: Machine Building* **6(105)** 46–62

[7] Vasconcelos D R 2017 Mathematical model for a collaborative indoor position system (IPS) and movement detection of devices within IoT environment *Proceedings of the Symposium on Applied Computing* 602–608.

[8] Muravyev K A, Averyanikhin A E and Kotelnitsky A V 2016 A methodology of calculating the optimal number of nodes in a virtualization cluster for a private virtual desktop cloud for maximum efficiency *International Research Journal* **5-3(47)** 6–13

[9] Vlasov A I, Karpunin A A and Novikov I P 2017 Systems analysis of blockchain data exchange and storage technology *Modern technology. Systems Analysis. Modeling* **3(55)** 75–83

[10] Berduygina O N, Vlasov A I and Kuzmin E A 2017 Investment capacity of the economy during the implementation of projects of public-private partnership *Investment Management and Financial Innovations* **14(3)** 189–198

[11] Yuldashev M N, Vlasov A I and Novikov A N 2018 Energy-efficient algorithm for classification of states of wireless sensor network using machine learning methods *Journal of Physics: Conference Series* **1015** 032153

[12] Karpunin A A and Kozlov A A 2017 Analysis of the methods for implementing decentralized applications in design and engineering information technology *Information technology in design and manufacturing* **4(168)** 39–44

[13] Vlasov A I and Yuldashev M N 2017 Gaussian processes in regression analysis of the states of wireless sensor network, including electromagnetic interference *Electromagnetic compatibility technologies* **3(62)** 35–43

[14] Andreev K A, Vlasov A I and Shakhnov V A 2016 Silicon pressure transmitters with overload protection *Automation and Remote Control* **77(7)** 1281–1285