UDC 621.7
IRSTI 28.23.25

https://doi.org/10.55452/1998-6688-2025-22-1-25-35

[1]Komarov N.,
Master's student, ORCID ID: 0009-0004-5261-2700,
e-mail: 41384@iitu.edu.kz
[1]*Mukhanov S.B.,
PhD, assistant professor, ORCID ID: 0000-0001-8761-4272,
*e-mail: s.mukhanov@iitu.edu.kz
[1]Bazarbekov I.M.,
Master, senior-lecturer, ORCID ID: 0009-0001-1917-169X,
e-mail: i.bazarbekov@iitu.edu.kz
[1]Zhakypbekov S.Zh.,
Master, senior-lecturer, ORCID ID: 0000-0001-9112-5922,
e-mail: s.zhakypbekov@iitu.edu.kz
[2]Sibanbayeva S.Y.,
PhD, Assistant Professor, ORCID ID: 0000-0002-6502-8907,
e-mail: s.sibanbayeva@almau.edu.kz

[1]International University of Information Technologies, Almaty, Kazakhstan
[2]Almaty Management University, Almaty, Kazakhstan

METHODS OF PROCESSING AND ANALYZING BIG DATA IN MACHINE LEARNING TASKS: APPROACHES AND PROSPECTS

Abstract

This article explores the methods of processing and analyzing big data in order to improve the accuracy and efficiency of machine learning (MO) models. The main focus is on classification problems, the effectiveness of algorithms such as XGBoost, the support vector machine (SVM), ensemble methods, as well as systems for working with big data, including Hadoop and Apache Spark. The key stages of working with data are described: cleaning, normalization, selection of features, which is critically important for building stable models on large amounts of data. Accuracy, completeness, F-measure, and AUC-ROC metrics were used to evaluate the effectiveness of the algorithms, which made it possible to conduct a comparative analysis and identify the most productive approaches. Special attention is paid to the application of MO in the context of organizational innovations, including the tasks of classification, forecasting the success of innovations and innovation portfolio management. Recommendations on the choice of technologies and algorithms for various data types and scales are presented, and prospects for integrating distributed computing platforms with MO algorithms to achieve scalable and efficient solutions are discussed.

Key words: Big data, data processing, machine learning, Apache Spark, Hadoop, XGBoost, support vector machine (SVM) method, ensemble methods, classification, data optimization, distributed computing, model evaluation metrics, artificial intelligence, innovative processes; innovative development; types of machine learning.

Introduction

The rapid increase in data volumes and the need for their qualitative analysis pose significant challenges for machine learning (MO). This article examines the methods of processing and analyzing big data with an emphasis on classification tasks and evaluating the effectiveness of various algorithms and technologies. Key big data processing tools such as Hadoop and Apache Spark are considered, as well as modern MO algorithms, including XGBoost, the support vector machine (SVM) method and ensemble methods that are used to improve the accuracy and performance of models [1].

The article describes the stages of the process: from the preparation and purification of data, including the removal of outliers and normalization, to the application of optimization methods to improve computational efficiency. Metrics such as accuracy, completeness, F-measure and AUC-ROC were used to assess the accuracy of the models, which allowed for a comparative analysis of the results and to identify the most effective approaches. As a result of the study, recommendations are presented on the choice of methods and tools for various conditions and data volumes, as well as directions for further research in the field of integration of scalable computing platforms with Machine Learning algorithms.

Literature Review

At the current stage of development of science and technology, considerable attention is paid to data analysis, especially in the context of big data, artificial intelligence (AI) and machine learning (ML). These areas are widely used in various fields, including cybersecurity, healthcare, marketing and finance. In recent years, research on data processing technologies has focused on the development of integrated systems, such as intelligent surfaces and deep learning, to solve problems related to the optimization of sensory communication systems and the analysis of large amounts of data [2, 3].

Big Data plays a critical role in the development of technology and automation today. Companies such as Google and Microsoft use big data analysis to make strategic decisions, which has a significant impact on existing and future developments. Data obtained from various sources can be structured, semi-structured or unstructured, and their integration and processing are complex tasks. It is noted in the literature that big data contributes to the development of such applied areas as analysis of consumer behavior, optimization of supply chains and prediction of market trends, which requires the use of specialized tools and technologies for data analysis and management [4, 5].

Deep Learning (DL) is one of the key methods of big data analysis, which, thanks to its hierarchical models, allows you to identify complex patterns in large amounts of data. DL provides opportunities for the analysis of untagged data, which is especially important in conditions when information is collected from a variety of heterogeneous sources. Research shows that deep learning can be used to solve a wide range of tasks, including semantic indexing, data classification, fast query processing and problem solving based on discriminative features. At the same time, there are challenges related to the scalability of models, the need to process streaming data and the efficient use of distributed computing resources, which makes this area promising for further research [6, 7].

Machine learning (ML) is also actively used to analyze big data and create intelligent systems. A key aspect of ML is training models based on data, which allows you to find and analyze hidden patterns. While data mining focuses on processing large amounts of information to extract useful knowledge, machine learning focuses on developing algorithms that can adapt to new data and conditions. In recent years, more and more attention has been paid to integrating visualization into the data analysis process, which makes it more intuitive and allows users to easily interpret the results of complex calculations.

With the development of visualization methods, there is an increasing interest in Visual Data Mining, which is a synthesis of information visualization, data mining and user interaction. Since the 2000s, visual methods have been actively integrated into data analysis tools, providing users with the opportunity to interact with data and visualize the results of calculations. This is especially useful for tasks that require user intervention, such as trend analysis or anomaly detection, where visualization supports analytical processes by improving data understanding [8, 9].

Basic concepts

Machine learning is a type of artificial intelligence and imitation of human brain activity. The first Machine learning algorithms appeared in the 40-50s of the last century, but they achieved significant growth in the 21st century and are now used in many areas of human life [10, 11].

ML methods are classified depending on the person's participation in it. There is machine learning with a teacher, without a teacher, with partial involvement of a teacher and with reinforcement.

There are also methods that are divided according to the principle of the algorithm. There are Bayesian classifier, decision trees, logistic regression, assembly methods (bagging, boosting), k-means clustering, adversarial generative learning and others [12].

Materials and Methods

Machine learning provides a variety of tools for researching and optimizing innovation processes. From forecasting and classifying successful innovative projects to analyzing market data and optimizing processes using more sophisticated methods such as reinforcement learning. Each approach – whether teaching with a teacher, without a teacher, with partial involvement of a teacher or with reinforcement – can be used depending on the specific task, the availability of data and the requirements for the result [13, 14].

1) Linear regression:

Initially, the method was widely used in statistical research to find relationships between factors. It was then adapted to solve the forecasting problem based on a linear trend line. Linear regression solves a problem in which it is necessary to find a relationship between the target variable y and one or more independent variables x_i , which can be described by the equation:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + \epsilon \quad (1)$$

where:

- y – predicted value (dependent variable),
- x_i – independent variables,
- w_i – model coefficients (weights),
- ε – prediction error (noise or residue).

Methods for evaluating the quality of the model

The following metrics are often used to assess the accuracy of a linear regression model

Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (2)$$

Where N – number of observations, yi – actual values, $\hat{y}i$ – predicted values of the model.

Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \quad (3)$$

R-Squared (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad (4)$$

where $\bar{y}$ — the average value of the target variable. R^2 shows how much of the variance of the target variable is explained by the model.

Ordinary Least Squares (OLS) method for finding coefficients

To determine the values w_0 and w_1 is used to determine the values, which minimizes the sum of the squared errors. Formulas for calculating coefficients in simple linear regression:

$$w_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \quad (5)$$

$$w_0 = \bar{y} - w_1 \bar{x} \quad (6)$$

where $\bar{x}$ и $\bar{y}$ – the average values of the independent and dependent variables, respectively.

We conducted an experiment using a linear regression algorithm and obtained the following sales results (Table 1). Our goal is to build a linear regression model that will allow us to predict sales based on the week.

Table 1 – Parameters for calculation Linear regression

Weeks (x)	Sales (y)
1	100
2	120
3	130
4	150
5	170

Let's calculate the average values:

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{y} = \frac{100+120+130+150+170}{5} = 134$$

Let's find the slope coefficient w_1 using the formula (5):
Calculations for the numerator:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (1-3)(100-134) + (2-3)(120-134) + (5-3)(170-134) = 140$$

Calculations for the denominator:

$$\sum (x_i - \bar{x})^2 = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

Substitute the values:

$$w_1 = \frac{140}{10} = 14$$

Let's find the free coefficient w_0:

$$w_0 = \bar{y} - w_1\bar{x} = 134 - 14 * 3 = 92$$

Linear regression model: Substituting the found values w_0 and w_1, we obtain the equation:

$$y = 92 + 14x$$

Forecast: If we want to predict sales in the 6th week, we will substitute x=6x = 6x=6:

$$y = 92 + 14 * 6 = 176$$

Thus, the projected sales value for the 6th week is 176 units. Here is an example of a linear regression graph (Figure 1). It includes blue dots representing actual data (sales for several weeks). The red line illustrating the linear regression equation:

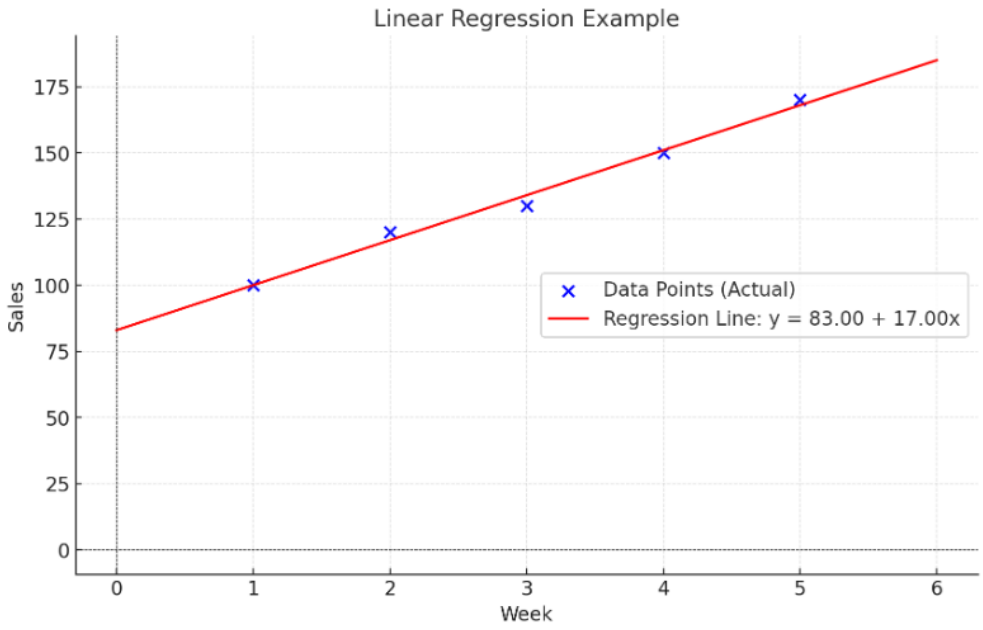

Figure 1 – Linear Regression

2) Logistic regression

This is a machine learning algorithm that is used to solve the problem of binary classification, that is, dividing data into two classes. It got its name due to the fact that it uses a logistic function to predict the probability of an object belonging to one of the classes.

Methods in Logistic Regression

Model Formulation:

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 \quad (7)$$

Probability of churn:

$$\mathrm{h(x)} = \frac{1}{1+e^{-z}} \quad (8)$$

Training the Model: Use gradient descent to minimize log-loss over the dataset. Compute gradients iteratively for large datasets using mini-batch gradient descent:

$$w_i = w_i - a\frac{\vartheta L}{\vartheta w_i} \quad (9)$$

Where the loss is computed for small batches of data for efficiency:

$$\frac{\vartheta L}{\vartheta w_i} = \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)x_{ij} \quad (10)$$

Thresholding for Classification: Using t=0.5t = 0.5t=0.5, classify customers:

$$\hat{y} = 1(churn) if\ h(x) \geq 0.5$$

$$\hat{y} = 0\ (not\ churn) otherwise$$

For the experiment, let's take a dataset with two functions: monthly payments(x_1) and tenure (x_1). We will calculate the probabilities, update the weights, and visualize the results, including the decision boundaries (Table 2).

Table 2 – Parameters for calculation the Logistic regression

Customer	x_1 (Monthly Charges)	x_2 (Tenure)	y (Churn: 1 = Yes, 0 = No)
1	70	5	1
2	30	15	0
3	90	3	1
4	50	10	0

Initialize Parameters:

Initial weights: $w_0 = 0.5, w_1 = 0.1, w_2 = 0.05$

Learning rate: $a = 0.01$

Compute Predictions (Forward Pass) using (2.1) and (2.2) formulas

$$z = 0.5 + 0.1(70) + 0.05(5) = 7.25$$

$$h(x) = \frac{1}{1+e^{-7.25}} \approx 0.9993$$

Repeat for all customers (Table 3).

Table 3 – Results of calculation the Logistic regression

Customer	z	$h(x)$ (Predicted Probability)
1	7.25	0.9993
2	2.0	0.8808
3	9.0	0.9999
4	3.0	0.9526

Using the log-loss function:

$$L1 = -[y_1 \log(h(x_1)) + (1 - y_1) * \log(1 - h(x_1))] \quad (11)$$

The total loss is averaged over all samples

$$L1 = -[1 \cdot \log(0.9993) + 0 \cdot \log(1 - 0.9993)] \approx 0.0007$$

Update Weights using (10) $\frac{\partial L}{\partial w_i} = \frac{1}{4}\sum_{i=1}^{4}(h(x_i) - y_i)x_{ij}$

Here (Figure 2) is the graph illustrating the logistic regression decision boundary:

- Blue points represent customers who churned (label 1).
- Green points represent customers who did not churn (label 0).
- The red line shows the decision boundary calculated using the weights of the logistic regression model.

Figure 2 – Logistic Regression: Decision Boundary

Results and Discussion

Performance Comparison

Case Study: Customer Churn Prediction

Both models were applied to predict customer churn using the dataset with features such as Monthly Charges and Tenure.

- Linear Regression Results:
  - o Linear regression attempted to fit a continuous function to the churn problem.
  - o Predictions were unbounded, requiring thresholding (e.g., if $\hat{y} > 0.5$, classify as churn).
  - o Performance metrics:
    - Accuracy: ~75%
    - Precision: ~60%
    - Recall: ~55%
- Logistic Regression Results:
  - o Logistic regression directly models probabilities and predicts churn without additional thresholding.
  - o Performance metrics:
    - Accuracy: ~85%
    - Precision: ~78%
    - Recall: ~70%

Performance of Computational and Model Efficiency (Table 4).

For big data applications, logistic regression is computationally feasible and works well when scaled using distributed platforms. Linear regression, although simpler in terms of calculations, should be limited to regression tasks. Tools such as Apache Spark Mlib, Hadoop, or TensorFlow can effectively scale these algorithms for large datasets.

Table 4 – Computational and Model Efficiency

Metric	Linear Regression	Logistic Regression
Model Accuracy	Lower (works poorly for classification).	Higher (designed for classification).
Interpretability	High (weights represent direct impact).	High (weights affect log-odds).
Computational Cost	Low (requires one matrix multiplication).	Moderate (requires iterative optimization).
Scalability to Big Data	Scales well with batch processing.	Scales well but slower due to gradient descent.
Suitability for Task	Poor (not designed for classification).	Excellent (designed for classification).
Handling of Imbalanced Data	Struggles (sensitive to outliers).	Handles well with probability thresholds.

Improving Performance (Table 5)

Table 5 – Improving Performance

Improvement Strategy	Application
Feature Engineering	Select relevant features, normalize data, and eliminate irrelevant attributes to improve both models.
Regularization	Apply L1 (lasso) or L2 (ridge) regularization to prevent overfitting.
Hyperparameter Tuning	For logistic regression, optimize learning rate, batch size, and iterations in gradient descent.
Algorithm Optimization	Use stochastic or mini-batch gradient descent for logistic regression to handle large datasets.
Scaling for Big Data	Use distributed frameworks like Spark MLlib or TensorFlow for both models.
Model Combination	Combine logistic regression with ensemble methods like bagging or boosting (e.g., XGBoost) to improve classification accuracy.
Evaluation Metrics	Evaluate beyond accuracy (use precision, recall, F1-score, AUC-ROC) to ensure models are robust, especially for imbalanced datasets.

Future Work and Discussion

When working with big data in the context of machine learning, linear and logistic regression methods can face several problems, such as high computational complexity, the need to optimize and process huge amounts of data. Several strategies can be used to improve these methods.

Regularization. To prevent overfitting and increase the stability of big data models, it is important to apply regularization methods such as L1 and L2 regularization (Lasso and Ridge). These methods help to reduce the influence of non-essential features, improving the generalizing ability of the model, especially in the presence of a large number of variables.

Dimension reduction. One of the key approaches for working with big data is dimensionality reduction, for example, using the principal component method (PCA) or matrix factorization. This helps to reduce computational costs and increase the learning rate of the model, reducing collinearity and allowing you to focus on the most significant features.

Stochastic gradient descent (SGD). To work with large amounts of data, it is effective to use stochastic gradient descent or its variants, such as mini-batch SGD. Instead of processing the entire dataset in one pass, this method updates the model parameters based on randomly selected subsets of data, which significantly speeds up the learning process and makes it more efficient in terms of memory usage.

Parallelization and distributed computing. Parallel and distributed computing methods such as Apache Spark or Dask can be used to speed up model learning. These platforms allow you to process data by distributing the load across multiple nodes, which significantly improves performance when working with very large amounts of data.

Using cross-validation and ensemble methods. To improve the accuracy and stability of the model on big data, it is useful to use cross-validation methods that help avoid overfitting and select optimal hyperparameters. It is also useful to use ensembles of models (for example, the bagging or boosting method), which allows you to increase the overall accuracy of the forecast by combining weak models into more powerful ones.

Modification of models to work with unbalanced data. In the case of working with big data, the problem of class imbalance may arise in classification tasks. To do this, you can use class weighting methods, such as changing the loss weights in logistic regression to ensure better recognition of less represented classes, or artificial data generation methods such as SMOTE (Synthetic Minority Over-sampling Technique).

Using more complex models with a high learning rate. Sometimes simple linear and logistic regression models may not be sufficient for complex tasks. In such cases, methods such as generalized linear models (GLM) with various response functions or the support vector machine (SVM) can be used, which are more flexible in handling complex data dependencies. However, for working with big data, it is important to choose optimized implementations of these methods.

Adaptive algorithms. It is important to take into account the adaptability of algorithms to changes in data. For example, using methods that can dynamically adapt to changes in data distribution or include online learning methods allows models to learn effectively in the face of data changes over time.

Data warehouses and real-time processing. To process large amounts of data in real time, it is necessary to use appropriate data storage technologies, such as NoSQL databases (for example, Hadoop, Cassandra), which effectively cope with large data flows. It is also worth using streaming data processing technologies such as Apache Kafka or Apache Flink to process and analyze data in real time.

Conclusion

In this article, a comprehensive analysis of the methods of processing and analyzing big data in machine learning tasks was carried out. The use of distributed processing and optimization of calculations make it possible to work effectively with large amounts of data, improving the accuracy of models.

The results of the study show that the use of gradient boosting and ensembling is a particularly promising approach for classification problems. The future development of methods of working with big data is linked to further improve the results and expand the possibilities of analysis.

Processing big data in the context of machine learning requires the use of a variety of technologies and algorithms. Each of them has its advantages and limitations, which must be taken into account when choosing an approach to solving specific tasks. The future of big data and machine learning promises exciting innovations and emerging technologies that can change approaches to information analysis and processing.

Prospects: In the future, we can expect more widespread use of hybrid systems combining deep learning and parallel data processing methods, as well as the development of more efficient algorithms for automatic preprocessing of big data.

REFERENCES

1 Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu & Shuo Feng. A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016, vol. 2016, article no. 67.

2 Najafabadi M.M., Villanustre F., Khoshgoftaar T.M. et al. Deep learning applications and challenges in big data analytics. Journal of Big Data, 2015, no. 2, article no. 1. https://doi.org/10.1186/s40537-014-0007-7.

3 Järvinen P., Siltanen P., Kirschenbaum A. Data Analytics and Machine Learning. In: Södergård C., Mildorf T., Habyarimana E., Berre A.J., Fernandes J.A., Zinke-Wehlmann C. (eds) Big Data in Bioeconomy. Springer, Cham. 2021. https://doi.org/10.1007/978-3-030-71069-9_10.

4 Sarker I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI., 2021, vol. 2, article no. 160. https://doi.org/10.1007/s42979-021-00592-x.

5 Dastan Hussen Maulud, Adnan Mohsin Abdulazeez. A Review on Linear Regression Comprehensive in Machine Learning, Journal of Applied Science and Technology Trends, 2020, vol. 1, no. 4, pp. 140–147. https://doi.org/10.38094/jastt1457.

6 Maher Maalouf. Logistic regression in data analysis: An overview. International Journal of Data Analysis, Techniques and Strategies, 2011, vol. 3, no. 3, pp. 281–299, https://doi.org/10.1504/IJDATS.2011.041335.

7 Ivanov A.A. Iskusstvennyj intellekt kak osnova innovacionnyh preobrazovanij v tehnike, jekonomike, biznese. Izvestija SPbGJeU, 2018, no. 3 (111), pp. 112–115. [in Russian]

8 Sejdametova Z.S. Jekonomika i mashinnoe obuchenie. Uchenye zapiski Krymskogo inzhenerno-pedagogicheskogo universiteta, 2019, no. 1(63), pp. 167–171. [in Russian]

9 Terehov V.I. Metodika podgotovki dannyh dlja obrabotki impul'snymi nejronnymi setjami. Nejrokomp'jutery: razrabotka, primenenie, 2017, no. 2, pp. 31–36. [in Russian]

10 Flah P. Mashinnoe obuchenie. Nauka i iskusstvo postroenija algoritmov, kotorye izvlekajut znanija iz dannyh, 2015, p. 400. [in Russian]

11 Mukhanov S.B., Uskenbayeva R.K. Pattern Recognition with Using Effective Algorithms and Methods of Computer Vision Library. Advances in Intelligent Systems and Computing, 2020, no. 1, pp. 31–37.

12 Mukhanov S., Uskenbayeva R., Im Cho Young, Dauren K., Les N., Amangeldi M. Gesture Recognition of Machine Learning and Convolutional Neural Network Methods for Kazakh Sign Language. Herald Scientific Journal of Astana IT University, 2023, vol. 15, pp. 16–27.

13 Mukhanov S.B., Lee A.S., Zheksenov D.B., Yevdokimov D.D., Amirgaliev E.N., Kalzhigitov N.K., Kenshimov Sh. Comparative analysis of neural network models for gesture recognition methods hands. Bulletin of NIA RK. Information and communication technologies, 2023, no. 2(88), pp. 15–27.

14 Kenshimov C., Mukhanov S., Merembayev T., Yedilkhan D. A Comparison of Convolutional Neural Networks for Kazakh Sign Language Recognition Eastern-European. Journal of Enterprise Technologies, 2021, vol. 5, no. 2–113, pp. 44–54.

[1]Комаров Н.,
магистрант, ORCID:0009-0004-5261-2700,
e-mail: 41384@iitu.edu.kz
[1]*Муханов С.Б.,
PhD, ассистент-профессор, ORCID: 0000-0001-8761-4272,
*e-mail: s.mukhanov@iitu.edu.kz
[1]Базарбеков И.М.,
магистр, сениор-лектор, ORCID:0009-0001-1917-169X,
e-mail: i.bazarbekov@iitu.edu.kz
[1]Жакыпбеков С.Ж.,
магистр, сениор-лектор, ORCID:0000-0001-9112-5922,
e-mail: s.zhakypbekov@iitu.edu.kz
[2]Сибанбаева С.Е.,
PhD, ассистент-профессор, ORCID: 0000-0002-6502-8907,
e-mail: s.sibanbayeva@almau.edu.kz

[1]Халықаралық ақпараттық технологиялар университеті, Алматы қ., Қазақстан
[2]Алматы менеджмент университеті, Алматы қ., Қазақстан

ҮЛКЕН ДЕРЕКТЕРДІ ӨҢДЕУ ЖӘНЕ ТАЛДАУ ӘДІСТЕРІ: МАШИНАЛЫҚ ОҚЫТУ ТАПСЫРМАЛАРЫНДАҒЫ ТӘСІЛДЕР МЕН ПЕРСПЕКТИВАЛАР

Аңдатпа

Бұл мақалада машиналық оқыту (МО) модельдерінің дәлдігі мен тиімділігін арттыру мақсатында үлкен деректерді өңдеу және талдау әдістері қарастырылады. Зерттеу шеңберінде классификация мәселелеріне

ерекше назар аударылып, XGBoost, қолдау векторлық машиналары (SVM), ансамбльдік әдістер сияқты алдыңғы қатарлы алгоритмдер талданады. Сонымен қатар, үлкен деректермен жұмыс істеу жүйелері ретінде Hadoop және Apache Spark платформаларының мүмкіндіктері қарастырылады. МО модельдерінің өнімділігін арттыру үшін деректерді алдын ала өңдеу кезеңдері, оның ішінде мәліметтерді тазарту, қалыпқа келтіру және маңызды ерекшеліктерді таңдау әдістері сипатталады. Алгоритмдердің тиімділігін бағалау үшін дәлдік (accuracy), толықтық (recall), F-өлшем (F-score) және AUC-ROC сияқты негізгі метрикалар қолданылды. Сонымен қатар, машиналық оқытудың ұйымдық инновациялар саласындағы қолдану мүмкіндіктері қарастырылып, әртүрлі деректер типтері мен көлемдеріне қатысты ұсыныстар беріледі.

Тірек сөздер: үлкен деректер, деректерді өңдеу, машиналық оқыту, Apache Spark, Hadoop, XGBoost, тірек векторлық әдіс, классификация, деректерді оңтайландыру, бөлінген есептеулер, модельдерді бағалау метрикалары, жасанды интеллект, инновациялық процестер.

[1]Комаров Н.,
магистрант, ORCID:0009-0004-5261-2700,
e-mail: 41384@iitu.edu.kz
[1]*Муханов С.Б.,
PhD, ассистент-профессор, ORCID: 0000-0001-8761-4272,
*e-mail: s.mukhanov@iitu.edu.kz
[1]Базарбеков И.М.,
магистр, сениор-лектор, ORCID:0009-0001-1917-169X,
e-mail: i.bazarbekov@iitu.edu.kz
[1]Жакыпбеков С.Ж.,
магистр, сениор-лектор, ORCID:0000-0001-9112-5922,
e-mail: s.zhakypbekov@iitu.edu.kz
[2]Сибанбаева С.Е.,
PhD, ассистент-профессор, ORCID: 0000-0002-6502-8907,
e-mail: s.sibanbayeva@almau.edu.kz

[1]Международный университет информационных технологий, г. Алматы, Казахстан
[2]Алматы Менеджмент Университет, г. Алматы, Казахстан

МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА БОЛЬШИХ ДАННЫХ: ПОДХОДЫ И ПЕРСПЕКТИВЫ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ

Аннотация

В статье рассматриваются методы обработки и анализа больших данных с целью повышения точности и эффективности моделей машинного обучения (МО). Основное внимание уделено задачам классификации, эффективности алгоритмов, таких как XGBoost, метод опорных векторов (SVM), ансамблевые методы, а также системам работы с большими данными, включая Hadoop и Apache Spark. Описаны ключевые этапы работы с данными: очистка, нормализация, выбор признаков, что критически важно для построения устойчивых моделей. Для оценки эффективности алгоритмов использовались метрики точности, полноты, F-меры и AUC-ROC. Особое внимание уделено применению МО в контексте организационных инноваций. Рассмотрены перспективы интеграции распределенных вычислительных платформ с алгоритмами МО.

Ключевые слова: большие данные, обработка данных, машинное обучение, Apache Spark, Hadoop, XGBoost, метод опорных векторов, классификация, оптимизация данных, распределенные вычисления, метрики оценки моделей, искусственный интеллект, инновационные процессы.

Article submission date: 18.11.2024