Применение методов машинного обучения и математических моделей для анализа больших данных

Application of machine learning methods and mathematical models for big data analysis

УДК 004
Получено: 19.10.2025 Одобрено: 22.11.2025 Опубликовано: 25.12.2025

Исаева А.Т.
Преподаватель кафедры «Технологии обучения математике, информатике и образовательный менеджмент», Ошский государственный университет, Кыргызская Республика, г. Ош
e-mail: Isaeva.aida.talaevna@gmail.com

Isaeva A.T.
Lecturer, Department of Technologies for Teaching Mathematics, Computer Science, and Educational Management, Osh State University, Osh, Kyrgyz Republic
e-mail: Isaeva.aida.talaevna@gmail.com

Келдибекова А.О.
Д-р пед. наук, профессор, заведующая кафедрой «Технологии обучения математике, информатике и образовательный менеджмент», Ошский государственный университет, Кыргызская Республика, г. Ош
e-mail: aidaoskk@gmail.com

Keldibekova A.O.
Doctor of Pedagogical Sciences, Professor, Head of the Department of Technologies for Teaching Mathematics, Computer Science, and Educational Management, Osh State University, Osh, Kyrgyz Republic
e-mail: aidaoskk@gmail.com

Аннотация
В условиях стремительного роста объемов информации проблема анализа больших данных (Big Data) приобретает особую актуальность. В данной статье исследуется симбиоз методов машинного обучения (МО) и фундаментальных математических моделей как основа для эффективного извлечения знаний из больших массивов информации. Цель работы — разработка и сравнительная оценка комплекса методов МО, подкрепленных математическим аппаратом, для задач классификации и кластеризации. На основе эксперимента с использованием набора данных UCI Machine Learning Repository проведен сравнительный анализ алгоритмов, включая логистическую регрессию, метод опорных векторов (SVM), случайный лес и многослойный перцептрон. Результаты показывают, что нейронные сети (Accuracy: 0.92, F1-мера: 0.89) и ансамблевые методы демонстрируют превосходство над классическими алгоритмами при работе с разнородными данными. Подчеркивается, что математические модели из областей

оптимизации, линейной алгебры и теории вероятностей являются неотъемлемым фундаментом, обеспечивающим корректность и эффективность алгоритмов МО. Делается вывод о целесообразности комплексного подхода, объединяющего вычислительную мощь МО и строгость математических моделей.
Ключевые слова: большие данные, машинное обучение, математические модели, классификация, кластеризация, нейронные сети, оптимизация.

Abstract
In the context of the rapid growth of information volumes, the problem of big data analysis is becoming particularly relevant. This article investigates the symbiosis of machine learning (ML) methods and fundamental mathematical models as a basis for effective knowledge extraction from large datasets. aree aim of the work is the development and comparative evaluation of a set of ML methods supported by a mathematical apparatus for classification and clustering tasks. Based on an experiment using a dataset from the UCI Machine Learning Repository, a comparative analysis of algorithms was conducted, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and Multilayer Perceptron. The results show that neural networks (Accuracy: 0.92, F1-score: 0.89) and ensemble methods outperform classical algorithms when working with heterogeneous data. It is emphasized that mathematical models from the fields of optimization, linear algebra, and probability theory are an integral foundation that ensures the correctness and efficiency of ML algorithms. The conclusion is made about the feasibility of an integrated approach combining the computational power of ML and the rigor of mathematical models.
Keywords: big data, machine learning, mathematical models, classification, clustering, neural networks, optimization.

Введение

Современный этап развития цифрового общества характеризуется внедрением машинного обучения в разные сферы жизни, в том числе и в образовательные процессы. Методы машинного обучения применяются педагогами для анализа данных, персонализации обучения, прогнозирования успеваемости студентов. Наблюдается экспоненциальный рост объемов данных, генерируемых пользователями, организациями и интеллектуальными системами. По оценкам экспертов, к 2025 г. объем мировых данных превысит 180 зеттабайт, что делает задачу их эффективной обработки и анализа одной из ключевых в научных и практических исследованиях [1].

Анализ больших данных, характеризуемых объемом, скоростью и разнообразием, невозможен без применения машинного обучения и математических моделей, так как традиционные статистические методы не справляются с проблемами масштаба. В исследованиях отмечается, что методы машинного обучения позволяют выявлять скрытые закономерности и строить предсказательные модели, в то время как математические модели формируют теоретическую основу, обеспечивая корректность и надежность полученных результатов [2], [3].

Актуальность исследования обосновывается возрастающей потребностью в интеллектуальных системах, способных анализировать большие массивы разнородных данных для поддержки принятия решений в сферах: финансы, здравоохранение [4] и управление.

Цель исследования — разработка и оценка комплекса методов машинного обучения и математических моделей, применимых для анализа больших данных.

Задачи исследования:

- изучить теоретические основы анализа больших данных и математические принципы, лежащие в основе МО;
- провести обзор и сравнительный анализ современных методов машинного обучения для классификации и кластеризации;

- реализовать эксперимент по анализу выбранного набора данных с применением различных алгоритмов;
- сравнить эффективность алгоритмов на основе метрик точности и сформулировать выводы.

Научная новизна исследования состоит в комплексном рассмотрении методов машинного обучения и математических моделей в едином подходе, подкрепленном экспериментальным сравнением эффективности алгоритмов на реальном наборе данных.

Обсуждение и результаты исследования

Классификация, регрессия, кластеризация и обучение с подкреплением относятся к распространенными типами задач в машинном обучении.

1.1. Большие данные и математический аппарат машинного обучения.

Термин Big Data обозначает массивы информации, соответствующие модели 3V: Volume (объем), Velocity (скорость) и Variety (разнообразие). Эти особенности делают необходимым использование машинного обучения (МО), которое, в свою очередь, базируется на строгом математическом фундаменте [5]:

- *Линейная алгебра* используется для операций с матрицами признаков, снижения размерности (PCA, SVD) и в архитектурах нейронных сетей.
- *Теория вероятностей и математическая статистика* лежат в основе байесовских методов, оценки uncertainty и валидации моделей.
- *Оптимизация* является ядром процесса обучения; методы градиентного спуска и его модификации (Adam, RMSprop) решают задачу минимизации функции потерь [3].

1.2. Обзор ключевых методов машинного обучения (МО).

Понимание принципов работы методов машинного обучения, их применимости к разным типам задач способствует правильному выбору эффективного подхода к их решению. Рассмотрим подробнее наиболее распространенные методы МО.

Обучение с учителем (Supervised Learning) используется для задач классификации и регрессии:

- *Логистическая регрессия:* статистическая модель, основанная на функции логит-преобразования.
- *Метод опорных векторов (SVM):* алгоритм, находящий оптимальную разделяющую гиперплоскость в пространстве признаков.
- *Ансамблевые методы (Случайный лес):* мощные алгоритмы, комбинирующие множество простых моделей (деревьев) для повышения точности и устойчивости к переобучению [6].
- *Нейронные сети (Многослойный перцептрон):* нелинейные модели, способные аппроксимировать сложные зависимости благодаря архитектуре из множества слоев [7].
- *Обучение без учителя (Unsupervised Learning).* Применяется для выявления скрытых структур.
- *Метод k-средних (k-means):* алгоритм кластеризации, минимизирующий внутри кластерную дисперсию.

Для верификации эффективности методов авторами был использован открытый набор данных "Bank Marketing" из репозитория UCI Machine Learning Repository [8], содержащий сведения о клиентах банка (≈45 000 записей, 20 признаков). Этапы обработки данных включали:

- *Очистка данных:* удаление дубликатов и записей с пропусками.
- *Кодирование категориальных признаков:* применен метод one-hot encoding.
- *Масштабирование признаков:* нормализация числовых значений до диапазона [0,1].

Для классификации были применены алгоритмы: Логистическая

регрессия, SVM, Случайный лес, Многослойный перцептрон. Для кластеризации использовались метод k-средних и иерархическая кластеризация. Оценка качества классификации проводилась по метрикам Accuracy, F1-мера и AUC-ROC.

Результаты оценки точности моделей классификации приведены в табл.

Таблица

Сравнительный анализ алгоритмов классификации

Алгоритм	Accuracy	F1-мера	AUC-ROC
Логистическая регрессия	0.81	0.79	0.84
SVM	0.85	0.82	0.87
Случайный лес	0.89	0.86	0.91
Нейронная сеть	**0.92**	**0.89**	**0.94**

Экспериментальные данные демонстрировали, что нейронные сети и ансамблевые методы (случайный лес) показывают более высокую точность по сравнению с классическими алгоритмами при анализе больших объемов разнородных данных. Это объясняется их способностью моделировать нелинейные и сложные зависимости в данных.

В задаче кластеризации метод k-средних показал лучшее время работы и интерпретируемость результатов для большого набора данных по сравнению с иерархической кластеризацией, выявив 4 четких сегмента клиентов.

Полученные результаты согласуются с современными тенденциями в анализе данных, приведенных в исследованиях [6], [7]. Однако выбор алгоритма должен быть обусловлен не только точностью, но и вычислительной эффективностью. Для задач, требующих быстрого ответа, более предпочтительными могут оказаться SVM или логистическая регрессия, в то время как для сложных прогнозных моделей оправдано применение нейронных сетей и градиентного бустинга. Авторы исследований считают, что математические модели оптимизации и линейной алгебры являются критически важным фундаментом, обеспечивающим работу всех этих алгоритмов на больших данных [3], [5].

Заключение

В ходе исследования сделаны выводы, подтверждающие высокую эффективность применения методов машинного обучения и математических моделей для анализа больших данных:

- ✓ определены ключевые особенности больших данных, и обоснована необходимость использования машинного обучения, подкрепленного математическим аппаратом;
- ✓ проведён сравнительный анализ алгоритмов классификации, показавший,

что нейронные сети (Accuracy: 0.92) и случайный лес (Accuracy: 0.89)

обеспечивают наилучшие результаты по метрикам точности и устойчивости;

- ✓ сочетание математических моделей (оптимизация, линейная алгебра) и методов машинного обучения является оптимальным подходом к анализу больших данных.

Перспективы дальнейших исследований связаны с применением более сложных архитектур нейронных сетей (например, глубоких и рекуррентных), а также с анализом данных в реальном времени с использованием потоковой обработки и изучением методов повышения интерпретируемости моделей (Explainable AI).

Литература

1. Исаков Р.Ж., Абдыкалыков А.А. Возможности применения искусственного интеллекта в диагностике заболеваний в Кыргызстане. Вестник Кыргызско-Российского Славянского университета. 2022. № 22(5). С. 124–130.
2. Bottou L., Curtis F.E., & Nocedal J. Optimization Methods for Large-Scale Machine Learning. SIAM Review. 2018. № 60(2). Pp. 223–311.
3. Chen M., Mao S., & Liu Y. Big Data: A Survey. Mobile Networks and Applications. 2019. №19(2). Pp. 171–209.
4. Dua D. and Graff C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. 2019. [Электронный ресурс]. URL: http://archive.ics.uci.edu/ml.
5. Deisenroth M.P., Faisal A.A. & Ong C.S. Mathematics for Machine Learning. Cambridge University Press. 2020. 398 p.
6. Goodfellow I., Bengio, Y. & Courville A. Deep Learning. MIT Press. 2016. 800 p.
7. Murphy K.P. Probabilistic Machine Learning: An Introduction. MIT Press. 2022. 864 p.
8. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., & Gulin A. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems. 2018. P. 31.