DOI: 10.18137/RNU.V9187.24.03.P.45 УДК 004.85:343.721:004.78

К.И. Лихоузов

ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ НА ПЛАТФОРМЕ HADOOP ДЛЯ ЭФФЕКТИВНОГО АНАЛИЗА И КЛАССИФИКАЦИИ МОШЕННИЧЕСКИХ ОПЕРАЦИЙ В БАНКОВСКОЙ СФЕРЕ

Аннотация. Материалом для исследования послужили последние научные работы в области машинного обучения, больших данных, алгоритмов в этой сфере и финансов. Были применены ряд методов анализа: монографический метод позволил глубоко погрузиться в тему, оценочный – эффективно анализировать данные, а рефлексивный метод дал возможность критически оценить полученные результаты. Использование машинного обучения в финансовых системах является революционным подходом в обнаружении и предотвращении мошенничества. Метод не только решает сложные задачи, связанные с несбалансированными данными и изменчивостью паттернов, но и позволяет обрабатывать информацию с большой скоростью и точностью. В результате применение таких технологий не только улучшает, а кардинально меняет способы защиты финансовых учреждений от мошеннических атак, обеспечивая высокий уровень безопасности для операций и клиентов.

Ключевые слова: машинное обучение, банковская сфера, мошеннические операции, алгоритмы, Hadoop.

K.I. Likhouzov

APPLICATION OF MACHINE LEARNING ALGORITHMS ON HADOOP PLATFORM FOR EFFICIENT ANALYSIS AND CLASSIFICATION OF FRAUDULENT TRANSACTIONS IN THE BANKING INDUSTRY

Abstract. The article uses the latest research work in the field of machine learning, big data, algorithms in this area and financial activities as research material. A number of analytical methods were applied in the process: the monographic method allowed to dive deeply into the topic, the evaluative method allowed effectively analyze the data, and the reflective method allowed to critically evaluate the findings. The use of machine learning in financial systems is a revolutionary approach to fraud detection and prevention. This method not only solves complex problems associated with unbalanced data and pattern variability, but also allows information to be processed with incredible speed and accuracy. As a result, the use of such technologies is not just improving, but revolutionizing the way financial institutions protect themselves from fraudulent attacks, ensuring the highest level of security for operations and customers.

Keywords: machine learning, banking, fraudulent transactions, algorithms, Hadoop.

Введение

В современной банковской сфере объемы данных постоянно растут. В связи с этим обнаружение и предотвращение мошеннических операций становится критически важной задачей для обеспечения безопасности финансовых институтов и их клиентов. Традиционные методы анализа данных и обнаружения мошенничества уже не всегда способны эффективно реагировать на современные угрозы. Актуальность применения платформы Hadoop для анализа и классификации мошеннических операций в банковской сфере обусловлена рядом ключевых факторов. Во-первых, объемы данных в банковской сфере посто-

Лихоузов Кирилл Игоревич
аспирант кафедры информационных систем и технологий, Московский информационно-технологический университет – Московский архитектурно-строительный институт, Москва.
Сфера научных интересов: информационные технологии, искусственный интеллект. ORCID: 0009-0002-8790-0871, SPIN-код: 2651-0565, Researcher ID: rid59685.
Электронный адрес: beakir93@gmail.com

янно растут, включая информацию о транзакциях, клиентах, активности счетов и др. Во-вторых, современные мошеннические схемы становятся всё более сложными и изощренными, что требует разработки более точных и адаптивных методов анализа. В-третьих, в условиях жестких требований к безопасности и конфиденциальности банки нуждаются в интегрированных и высокоэффективных системах обнаружения мошенничества.

Исследования по вопросу распознавания мошеннических действий

Во многих научных исследованиях затрагивается вопрос распознавания мошеннических действий. Например, П. Крылов [1] и П.В. Слипенчук [2] предложили методы и алгоритмы для выделения характеристик пользовательской активности и классификации их на два класса. В области исследования мошеннических операций А. Хусеинович [3] предложил методологию, включающую в себя применение наивного байесовского классификатора, дерева решений C4.5, а также усовершенствованный подход через бэггинг, используя дерево решений C4.5 как основу для усиления предсказательной способности. Исследователи С. Дханхада, Э. Мохаммед и Б. Фара [4] взяли за основу комплексный подход, применив несколько расширенных алгоритмов, таких как стекинг, случайный лес и градиентный бустинг. При сравнении результатов было обнаружено, что применение стекинг-классификатора в сочетании с логистической регрессией метаклассификатора демонстрирует превосходные результаты по сравнению с другими рассмотренными методами. Й. Сахин и Е. Думан [5] сконцентрировались на анализе эффективности искусственных нейронных сетей и логистической регрессии, применяемых к реальным датасетам. Их исследование позволило выявить более высокую точность предсказаний, достигаемую с помощью нейронных сетей в сравнении с моделями, основанными на логистической регрессии. Благодаря машинному обучению банки могут реализовывать гибкие и эффективные стратегии, выгодные как бизнесу, так и клиентам [6].

Принципы и виды машинного обучения

Принципы машинного обучения многочисленны и разнообразны. В Таблице 1 представлены ключевые из них, основанные на понятиях статистического анализа и вероятностного предсказания.

Таблица 1

Принципы машинного обучения

Принцип	Характеристика
Обучение на основе опыта	Модели машинного обучения «обучаются» на основе набора данных (тренировочного набора), который используется для настройки их параметров. Например, модель для классификации спама может быть обучена на основе тысячи примеров электронных писем, являющихся спамом или не являющихся таковыми
Обобщение	Обученная модель машинного обучения должна быть способна «обобщать» свои знания на новые, ранее не виденные данные. В случае классификатора спама, он должен быть способен правильно определить, является ли новое письмо спамом или нет
Адаптация и улучшение со временем	Со временем, по мере поступления новой информации, модели машинного обучения способны прогрессивно совершенствоваться. Это особо актуально в динамично развивающихся сферах, где необходимо, чтобы модель эффективно адаптировалась к появляющимся тенденциям и образцам

Источник: [7].

В современном понимании машинное обучение обычно разделяется на два основных типа – индуктивное и дедуктивное. Индуктивное обучение заключается в поиске закономерностей и шаблонов в данных для создания модели прогнозирования, дедуктивное обучение основано на использовании заранее известных экспертных знаний для создания модели [8]. Для целей машинного обучения применяются различные подходы. Например, метод обучения с учителем предполагает использование данных, где каждому примеру уже присвоена определенная метка или категория, что помогает модели обучаться на конкретных примерах [9]. В случае обучения без учителя данные представляются в виде векторов-признаков, но без каких-либо меток. Модель пытается самостоятельно найти структуру или взаимосвязи в данных. Существует также частичное обучение, предполагающее использование данных, в которых только некоторые примеры имеют метки [10]. Наконец, в обучении с подкреплением модель учится на опыте, реагируя на различные ситуации и получая обратную связь от окружающей среды в виде вознаграждения за успешные действия [11].

На Рисунке представлена структура экосистемы Hadoop. Она выстроена в виде слоев, каждый из которых выполняет определенные функции в рамках системы обработки и хранения больших данных:

1. Основа – инфраструктура хранения:

• HDFS представляет собой файловую систему, спроектированную для размещения данных на множестве серверов, обеспечивая надежность и доступность;

• HBase, построенная на основе HDFS, является базой данных типа columnarstorage, что приводит к ускорению доступа к большим массивам данных.

2. Управление ресурсами:

• MapReduce определяется как методика программирования с поддержкой фреймворка, предназначенная для обработки данных в рамках распределенных систем;

• YARN обозначает средство для управления ресурсами, позволяя использовать разнообразные процессы обработки данных в рамках всей экосистемы Hadoop.

3. Инструментарий высокого уровня:

• Hive – инструментарий, облегчающий взаимодействие с данными в HDFS при помощи SQL-запросов для анализа и управления;

• Pig предлагает платформу для обработки данных, используя специализированный скриптовый язык PigLatin;

• Mahout занимается предоставлением алгоритмов машинного обучения, масштабируемых под большой объем данных;

• Avro – фреймворк, который упрощает сериализацию структурированных данных;

• Sqoop служит мостом для эффективного перемещения данных между Hadoop и традиционными реляционными базами данных.

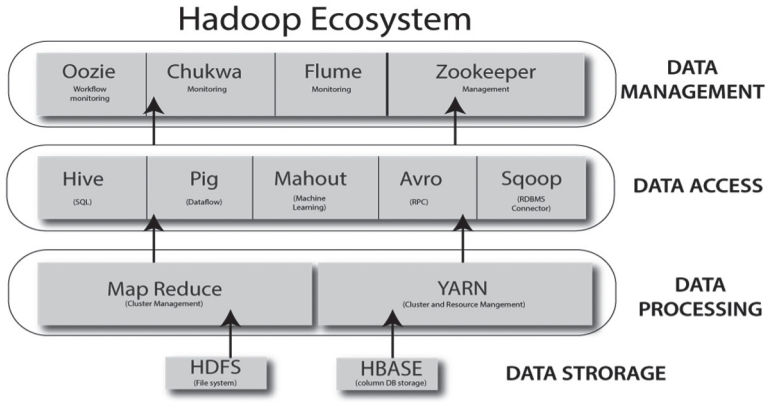

Рисунок. Экосистема Hadoop

Источник: Hadoop Ecosystem. URL: https://www.geeksforgeeks.org/hadoop-ecosystem/ (дата обращения: 09.03.2024).

Алгоритмы машинного обучения

Для определения типов мошеннических действий в Hadoop используются два основных метода – бинарная и многоклассовая классификации. Бинарная классификация определяет, является ли транзакция мошеннической или нет, в то время как многоклассовая классификация выявляет конкретные типы мошенничества. Представим несколько алгоритмов, которые можно использовать для улучшения точности прогнозов в банковской сфере (см. Таблицу 2).

Таблица 2

Обзор алгоритмов машинного обучения в банковском секторе

Наименование алгоритма	**Описание**	**Примеры применения в банковской сфере**
Случайный лес	Использует множество деревьев решений, результаты которых объединяются для повышения точности	Определение мошенничества с кредитными картами, кредитный скоринг
Градиентный бустинг	Постепенно улучшает слабые модели, создавая более сильную предсказательную модель	Оценка риска дефолта по кредитам, автоматизация внутреннего аудита
Логистическая регрессия	Идеально подходит для бинарной классификации объектов	Классификация транзакций как мошеннических или законных

Окончание таблицы 2

Наивный байесовский классификатор	Применяет вероятностный подход к классификации	Категоризация клиентов по типам для таргетированных маркетинговых акций
Машины опорных векторов	Эффективны при большом количестве признаков	Определение клиентов, склонных к досрочному погашению кредитов

Источник: Машинное обучение против фрода. URL: https://www.osp.ru/os/2017/02/13052223/ (дата обращения: 09.03.2024).

Отказоустойчивость – одно из основных преимуществ использования Hadoop в финансах. На фундаментальном уровне большие финансовые данные передаются через отдельные узлы и в процессе передачи копируются на другие узлы кластера. Это гарантирует, что при маловероятном сценарии сбоя копия данных всегда будет готова к использованию. Кроме того, распределенная архитектура без NameNode обеспечивает страховку как от единичных, так и от множественных сбоев.

При использовании алгоритмов машинного обучения при выявлении аномалий и мошенничества могут возникнуть ряд проблем (см. Таблицу 3).

Таблица 3

Проблемы и решения в обнаружении мошенничества с использованием алгоритмов машинного обучения

Проблема	Описание	Решение
Недостаток данных	Необходимость в большем объеме репрезентативных данных для эффективной работы алгоритмов машинного обучения	Пополнение набора данных аномалий и мошеннической активности, обучение на обновленных данных
Обработка данных в реальном времени	Трудности с обработкой больших объемов данных в режиме реального времени, особенно при высокой нагрузке	Параллельная обработка данных; оптимизация алгоритмов для работы с большими объемами информации
Развитие новых методов мошенничества	Постоянная необходимость адаптации алгоритмов к новым видам мошенничества и обходу существующих систем	Регулярное обновление и обучение алгоритмов машинного обучения, использование новейших методов и приемов
Сложность интерпретации результатов	Проблемы с выявлением ложных срабатываний или пропуска реальных аномалий, затрудняющие их интерпретацию	Использование нескольких алгоритмов и анализ результатов вместе, применение дополнительных методов проверки

Источник: таблицы 3 и 4 составлены автором.

Совершенствование алгоритмов обнаружения аномалий включает контекстно-ориентированный подход, использование гибридных алгоритмов и адаптивное обучение для повышения точности и эффективности (см. Таблицу 4).

Таблица 4

Развитие подходов к обнаружению аномалий: особенности и примеры использования

Подход	Характеристика	Примеры применения

Окончание таблицы 4

Контекстно-ориентированный	Адаптация к изменяющейся среде и учет особенностей для точного определения аномалий и оценки их опасности	Мониторинг финансовых транзакций для выявления аномальных операций в реальном времени
Гибридный	Комбинация различных методов машинного обучения для повышения точности и эффективности обнаружения аномалий	Обнаружение аномалий в сетевом трафике путем сочетания статистических методов с методами глубокого обучения
Адаптивное обучение	Обновление моделей и стратегий обнаружения аномалий в соответствии с изменениями данных и появлением новых угроз	Использование алгоритмов обнаружения аномалий в системах безопасности для адаптации к новым видам атак

Обсуждение результатов исследования и заключение

В борьбе с мошенническими операциями в финансовых системах ключевое значение приобретает машинное обучение. Решив такие проблемы, как дисбаланс данных, их качество, отработка функционала и дрейф концепций, финансовые организации могут создавать надежные модели обнаружения ложных схем. Надежность моделей повышают эффективные методы предварительной обработки данных, включая их очистку, масштабирование и отбор признаков.

Мониторинг в режиме реального времени на основе потоковых систем и систем оповещения обеспечивает своевременное обнаружение и предотвращение фактов мошенничества. Оптимальное приложение для борьбы с мошенничеством должно быть мощным, быстрым и точным и адаптироваться к различным ситуациям. Чтобы достичь этой цели, приложение должно уметь обрабатывать различные данные о транзакциях и подписях, одновременно поддерживая базу данных в актуальном состоянии. В этом плане платформа на базе Hadoop является чрезвычайно производительным приложением для машинного обучения, способным поддерживать множество различных алгоритмов машинного обучения. Платформы, функционирующие на основе Hadoop, обладают функционалом высокоточного анализа данных, тем самым обеспечивая эффективные средства для противодействия финансовым мошенничествам в банковской сфере. Достигается это благодаря возможности проведения аналитики в реальном времени, которая позволяет своевременно выявлять и предотвращать возможные злоупотребления.

Дальнейшие направления совершенствования и разработки алгоритмов машинного обучения в рамках Hadoop обусловлены следующими аспектами:

• внедрение контекстно-зависимого распознавания для более точной идентификации аномальных активностей и событий;

• применение гибридных алгоритмов, сочетающих различные подходы, с целью повышения эффективности обнаружения подозрительных операций;

• концепция адаптивного обучения, допускающая корректировку моделей и механизмов обнаружения в соответствии с текущими тенденциями и изменениями в данных;

• использование обширных наборов данных для формирования более надежных предсказательных моделей;

• развитие вычислительных алгоритмов для обеспечения возможности обнаружения отклонений и аномалий в максимально приближенном к реальному времени режиме.

Такой подход укрепляет позиции Hadoop как мощной платформы борьбы с мошенничеством, делая адаптивные механизмы обнаружения аномалий одним из приоритетных

направлений в области данных.

Литература

1. *Крылов П.* Схемы хищений в системах ДБО и пять уровней противодействия им // Расчеты и операционная работа в коммерческом банке. 2018. № 3 (145). С. 46–59. URL: http://www.reglament.net/bank/raschet/2018_3/get_article.htm?id=5669 (дата обращения: 04.06.2024).
2. *Слипенчук П.В.* Алгоритм извлечения характерных признаков из данных пользовательских активностей // Вопросы кибербезопасности. 2019. № 1 (29). С. 53–58. EDN YZFWPZ. DOI: 10.21681/2311-3456-2019-1-53-58
3. *Husejinovic A.* Credit card fraud detection using naive Bayesian and c4.5 decision tree classifiers // Credit card fraud detection using naive Bayesian and C. 2020. Vol. 8. No. 1. P. 1–5. DOI: 10.21533/pen.v%25vi%25i.300
4. *Dhankhad S., Mohammed E., Far B.* Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study // IEEE International Conference on Information Reuse and Integration (IRI). Salt Lake City, UT, USA, July 06–09, 2018. P. 122–125. DOI: https://doi.org/10.1109/IRI.2018.00025
5. *Sahin Y., Duman E.* Detecting credit card fraud by ANN and logistic regression // International Symposium on Innovations in Intelligent Systems and Applications. Istanbul, Turkey, June 15–18, 2011. P. 315–319. DOI: https://doi.org/10.1109/INISTA.2011.5946108
6. *Татаринцев М.А., Никитин П.В., Горохова Р.И., Долгов В.И.* Сравнительный анализ технологий машинного обучения для задач кредитного скоринга // Фундаментальные исследования. 2023. № 1. С. 49–54. DOI: http://doi.org/10.17513/fr.43419
7. *Кривчук М.А.* Практическая ценность использования нейросетей в личных целях // Обществознание и социальная психология. 2023. № 9-3 (39). С. 59–62. EDN ARANTM.
8. *Судаков В.А., Семенов Т.П.* Методы машинного обучения при расчёте скоринга клиентов банка // Международный журнал информационных технологий и энергоэффективности. 2023. Т. 8. № 3 (29). С. 22–25. EDN RCIRND. URL: http://openaccessscience.ru/index.php/ijcse/article/view/297/268 (дата обращения: 04.06.2024).
9. *Загайнов М.А., Костенков Е.А., Кузнецов Д.С.* Машинное обучение. Области применения технологии и перспективы развития // Colloquium-Journal. 2019. № 16–1(40). С. 49–50. URL: https://colloquium-journal.org/wp-content/uploads/2022/05/Colloquium-journal-2019-40-1.pdf (дата обращения: 04.06.2024).
10. *Згонникова А.О., Прокопенко А.А.* Машинное обучение и обучение на протяжении всей жизни // Новые научные исследования : Труды VIII Международной научно-практической конференции. Пенза, 27 августа 2021 г. Пенза : Наука и Просвещение, 2022. С. 22–24. EDN SCYXYS.
11. *Богданов А.В., Тхуреин К., Пья С.Ко.Ко., Чжо За.* Сравнение производительности инструментов для обработки больших данных // Современные наукоемкие технологии. 2020. № 6–1. С. 9–14. EDN TWRWGY. DOI: https://doi.org/10.17513/snt.38064

References

1. Krylov P. (2018) Theft schemes in RBS systems and five levels of counteraction to them. *Raschety i operatsionnaya rabota v kommercheskom banke* [Settlements and operational work in a commercial bank]. No. 3 (145). Pp. 46–59. URL: http://www.reglament.net/bank/raschet/2018_3/get_article.

htm?id=5669 (accessed 04.06.2024). (In Russian).
2. Slipenchuk P.V. (2019) Algorithm for extracting characteristic features from user activity data. *Voprosy kiberbezopasnosti* [Cybersecurity Issues]. No. 1 (29). Pp. 53–58. DOI: 10.21681/2311-3456-2019-1-53-58 (In Russian).
3. Husejinovic A. (2020) Credit card fraud detection using naive Bayesian and c4.5 decision tree classifiers. *Credit card fraud detection using naive Bayesian and C*. Vol. 8. No. 1. Pp. 1–5. DOI: 10.21533/pen.v%25vi%25i.300
4. Dhankhad S., Mohammed E., Far B. (2018) Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In: *IEEE International Conference on Information Reuse and Integration (IRI)*. Salt Lake City, UT, USA, July 06–09, 2018. Pp. 122–125. DOI: https://doi.org/10.1109/IRI.2018.00025
5. Sahin Y., Duman E. (2011) Detecting credit card fraud by ANN and logistic regression. *International Symposium on Innovations in Intelligent Systems and Applications.* Istanbul, Turkey, June 15–18, 2011. Pp. 315–319. DOI: 10.1109/INISTA.2011.5946108
6. Tatarintsev M.A., Nikitin P.V., Gorokhova R.I., Dolgov V.I. (2023) Comparative analysis of machine learning technologies for credit scoring tasks. *Fundamental'nye issledovaniya* [Fundamental Research]. No. 1. Pp. 49–54. DOI: 10.17513/fr.43419 (In Russian).
7. Krivchuk M.A. (2023) Practical value of using neural networks for personal purposes. *Social Studies and Social Psychology*. No. 9-3 (39): Pp. 59–62. URL: https://www.elibrary.ru/ARANTM (In Russian).
8. Sudakov V.A., Semenov T.P. (2023) Machine learning methods in calculation of bank customer scoring. *International Journal of Information Technology and Energy Efficiency*. Vol. 8. No. 3 (29). Pp. 22–25. URL: http://openaccessscience.ru/index.php/ijcse/article/view/297/268 (accessed 04.06.2024). (In Russian).
9. Zagainov M.A., Kostenkov E.A., Kuznetsov D.S. (2019) Machine Learning. Areas of technology application and prospects of development. *Colloquium-Journal*. No. 16-1 (40). Pp. 49–50. URL: https://colloquium-journal.org/wp-content/uploads/2022/05/Colloquium-journal-2019-40-1.pdf (accessed 04.06.2024). (In Russian).
10. Zgonnikova A.O., Prokopenko A.A. (2022) Machine learning and lifelong learning. In: Gulyaev G.Yu. (Ed) *Novye nauchnye issledovaniya* [New Scientific Research] : Proceedings of the VIII International Scientific and Practical Conference. Penza, August 27, 2021. Penza : Nauka i Prosveshchenie Publ. Pp. 22–24. URL: https://naukaip.ru/wp-content/uploads/2022/08/MK-1482.pdf (accessed 04.06.2024). (In Russian).
11. Bogdanov A.V., Thurein K., Pya S.Ko.Ko, Zho Za. (2020) Performance comparison using spark and Hadoop for big data processing. *Modern High Technologies*. No. 6-1. Pp. 9–14. URL: https://s.top-technologies.ru/pdf/2020/6-1/38064.pdf (accessed 04.06.2024). (In Russian).