Методы машинного обучения для автоматической обработки документов

С.А. Корчагин, Д.В. Сердечный, А.И. Окунев, Н.А. Андриянов

Финансовый университет при Правительстве Российской Федерации, Москва

Аннотация: Работа посвящена анализу методов машинного обучения для решения задач автоматической обработки документов. В исследовании рассмотрены такие методы, как классификация, извлечение информации, распознавание образов и обработка естественного языка и их применение в анализе текстовых данных. Проведен анализ существующих алгоритмов и моделей, включая линейные модели, деревья решений, методы опорных векторов и проведено сравнение их эффективности в зависимости от различных условий и параметров. Особое внимание уделяется проблемам, с которыми сталкиваются специалисты при использовании методов машинного обучения в работе с документами, такими как качество данных, необходимость предварительной обработки и настройка параметров моделей. Приводятся перспективы дальнейших исследований в данной области и примеры возможной интеграции современных методов машинного обучения для повышения эффективности и точности автоматической обработки документов в различных отраслях.

Ключевые слова: машинное обучение, автоматическая обработка документов, вычислительный эксперимент, искусственный интеллект, модели классификации, программный комплекс

Введение

В настоящее время наблюдается экспоненциальный рост количества информации, обрабатываемой в цифровом формате [1,2]. Автоматизация процессов извлечения, структурирования и анализа данных становится одной из ключевых задач для организаций различных масштабов и направлений деятельности [3]. С использованием современных алгоритмов машинного обучения и методов обработки естественного языка (NLP), представляется возможным выделение, классификация и извлечение информации из большого объема неструктурированных данных, значительно сокращая время и ресурсы, необходимые для выполнения рутинных задач. Однако выбор оптимального метода для конкретной задачи, а также понимание возможностей и ограничений различных инструментов, является критически важным для достижения успешных результатов. В исследовании проводится анализ существующих методов машинного обучения, применяемых для

автоматической обработки документов, рассмотрены преимущества и недостатки таких методов, а также приведены практические кейсы применения технологий искусственного интеллекта для решения указанных задач.

Методы машинного обучения

Рассмотрим основные методы машинного обучения, которые применяются для автоматической обработки документов. Каждый из этих методов обладает определенными преимуществами и недостатками, и их выбор зависит от типа задачи, объема данных и других факторов. Автоматическая обработка документов включает в себя идентификацию, классификацию и извлечение информации из текстов, что делает эффективное использование машинного обучения особенно важным в этой области.

Логистическая регрессия является классическим инструментом для решения задач классификации [4]. В контексте автоматической обработки документов логистическая регрессия может быть использована для классификации документов по категориям, например, для определения по группам счетов, договоров, отчетов.

Деревья решений — интерпретируемый метод машинного обучения, который используется как для классификации, так и для регрессии [5]. В задачах автоматической обработки документов деревья решений могут эффективно разделять документы на категории на основе их содержимого, позволяя быстро идентифицировать, например, юридические документы, финансовую отчетность или научные статьи. Метод случайного леса (Random Forest) [6], основанный на ансамбльном подходе, может использоваться для повышения точности классификации, комбинируя результаты множества деревьев и уменьшая риск переобучения на специфических данных.

Методы градиентного бустинга, такие как XGBoost [7], LightGBM [8] и CatBoost [9], становятся особенно актуальными для обработки больших объемов документов. С их помощью можно автоматизировать такие задачи, как анализ тональности текстов, определение тематики или выявление аномалий в содержании. Благодаря своей высокой точности и способности обрабатывать большие объемы данных, эти алгоритмы часто используются в системах, обрабатывающих данные из электронных писем, отчетов и других форматов.

Методы опорных векторов (SVM) — популярный инструмент для классификации, который работает путем нахождения гиперплоскости, разделяющей различные классы с максимальным зазором [10]. В контексте автоматической обработки документов SVM могут использоваться для различения спама и легитимных электронных писем, а также для классификации текстов по тематическим категориям. Данные методы особенно полезны, когда набор данных мал, но качественно аннотирован и требует высокой точности в классификации.

Нейронные сети позволяют обрабатывать и анализировать большие объемы данных. Так, например, сверточные нейронные сети (CNN) [11] могут использоваться для обработки изображений документов, например, для распознавания текстов в рамках систем оптического распознавания символов (OCR). Рекуррентные нейронные сети (RNN) [12] и их варианты, такие как Long Short-Term Memory (LSTM) [13], эффективны для обработки последовательных данных, таких как текстовая информация, что делает их хорошо применимыми для автоматического анализа текста и извлечения ключевых фраз в документах.

Методы кластеризации, такие как K-means [14] и DBSCAN [15], позволяют группировать документы на основе сходства содержимого. Автоматическая обработка документов может включать в себя

использование кластеризации для выявления групп схожих документов, что может помочь в организации файлов, классификации множества текстов или в поиске релевантной информации. Эти алгоритмы могут также применяться в контексте тематического моделирования, чтобы обнаруживать скрытые темы в наборах данных.

Для обработки больших объемов данных и повышения эффективности их последующей классификации часто применяются методы снижения размерности, такие как Principal Component Analysis (PCA) и t-Distributed Stochastic Neighbor Embedding (t-SNE) [15]. В области автоматической обработки документов эти алгоритмы позволяют уменьшить размерность векторных представлений текстов, сохраняя при этом их наиболее существенные характеристики. Это упрощает процесс анализа и визуализации данных, а также помогает улучшить производительность алгоритмов классификации и кластеризации.

Каждый из перечисленных методов имеет свои преимущества и недостатки, и выбор подходящего подхода зависит от специфики задачи. Важным аспектом является проведение вычислительных экспериментов и тестирование различных моделей для достижения максимально возможной точности и эффективности автоматической обработки документов. Далее мы рассмотрим конкретный пример применения приведенных методов в возможных сценариях обработки документов и продемонстрируем их практическое значение для оптимизации бизнес-процессов организации.

Программная реализация и проведение вычислительных экспериментов

В рамках исследования была решена задача автоматической классификации документов. В качестве примера, использовались документы 4 типов:

– научные статьи;

– документы юридического характера (договоры, акты);

– резюме соискателей;

– техническая документация.

Для обучения моделей было использовано 800 документов (по 200 документов в каждом классе). Выборка была разделена на тестовую – 20% и обучающую – 80%.

На первом этапе все документы были преобразованы в единый формат (JSON) удобный для анализа. Далее проводилась предобработка данных: очистка данных (удаление стоп-слов, знаков препинания и специальных символов), нормализация (лемматизация или стемминг), токенизация (разделение текста на отдельные слова или фразы) и удаление дубликатов.

На следующем шаге документы были представлены в виде векторов, после чего были выбраны алгоритмы машинного обучения. В рамках исследования мы выбрали следующие алгоритмы классификации, описанные в предыдущем разделе: логистическая регрессия, SVM, Random Forest, K-means. Фрагмент программного кода для выполнения указанных алгоритмов приведен на рисунке 1. Для реализации моделей машинного обучения использовалась библиотека Scikit-Learn. Основной метрикой оценки качества работы моделей была точность (accuracy) – доля правильно классифицированных экземпляров к общему числу экземпляров. Результаты работы моделей для тестовой выборки приведены на рисунке 2.

```python
import numpy as np
import pandas as pd
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.cluster import KMeans
from sklearn.metrics import classification_report, silhouette_score

# Разделение данных на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Логистическая регрессия
logreg = LogisticRegression(max_iter=1000)
logreg.fit(X_train, y_train)
y_pred_logreg = logreg.predict(X_test)

print("Логистическая регрессия:\n", classification_report(y_test, y_pred_logreg))

# SVM
svm = SVC()
svm.fit(X_train, y_train)
y_pred_svm = svm.predict(X_test)

print("SVM:\n", classification_report(y_test, y_pred_svm))

# Случайный лес
rf = RandomForestClassifier(n_estimators=100)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

print("Случайный лес:\n", classification_report(y_test, y_pred_rf))

# K-means
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(X)
```

Рис. 1. – Фрагмент программного кода реализации алгоритмов машинного обучения

Проведя анализ используемых методов: логистической регрессии, SVM, Random Forest и алгоритма K-means, мы получили наглядное представление о том, как различные алгоритмы классификации справляются с задачей. Как видно из графика, лучше всего сработал алгоритм Random Forest. Данный алгоритм, основанный на методах ансамблевого обучения обеспечил самую высокую точность в 84 %.

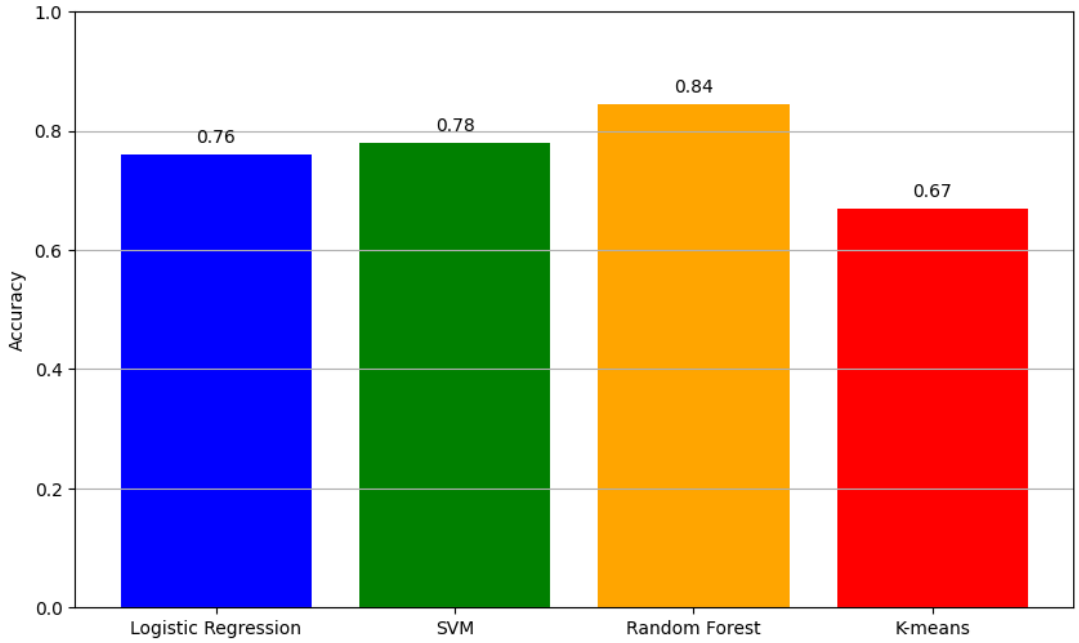

Рис. 2. – Сравнение точности моделей классификации

Заключение

Результаты исследования показали, что современные алгоритмы, такие как логистическая регрессия, метод опорных векторов (SVM), случайный лес, K-means показали свою эффективность при решении задач, связанных с классификацией документов.

Перспективы дальнейших исследований в области машинного обучения для автоматической обработки документов обширны. Во-первых, увеличение объема данных, создаваемых в цифровом формате, требует разработки новых методов, способных работать с большими и сложными наборами данных, включая неструктурированные данные. Такие методы, как глубокое обучение, могут быть более эффективно внедрены для извлечения значимых паттернов из сложных источников информации, например, из изображений и видео.

Во-вторых, важно продолжать работу в области интерпретируемости и объяснимости моделей. Понимание механизмов, лежащих в основе таких

решений, становится критически важным. Это поможет повысить доверие пользователей и соответствовать правовым и этическим стандартам.

Таким образом, применяя современные алгоритмы машинного обучения для автоматической обработки документов, организации могут повысить свою эффективность, уменьшить затраты и автоматизировать работу сотрудников. Будущее этой области обещает быть многообещающим, и дальнейшие исследования в данном направлении будут способствовать созданию более производительный и универсальных систем.

Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финуниверситета.

Литература

1. Панишев И. А. Полнота предоставляемой информации как ключевой аспект успешной проверки контрагента на предмет его финансовой устойчивости и благонадежности // Сибирская финансовая школа. – 2024. – №. 1. – С. 81-86.

2. Montina A., Wolf S. Generalized Gleason theorem and a finite amount of information about the measurement context // Physical Review A. – 2023. – V. 107. – №. 1. – P. 012218.

3. Domova V., Vrotsou K. A model for types and levels of automation in visual analytics: a survey, a taxonomy, and examples // IEEE Transactions on Visualization and Computer Graphics. – 2022. – V. 29. – №. 8. – pp. 3550-3568.

4. Бегунков В. И., Ковалев М. Я. Классификация займов с использованием логистической регрессии // Информатика. – 2023. – Т. 20. – №. 1. – С. 55-74.

5. Горшенин А. Ю., Грицай А. С., Денисова Л. А. Применение машинного обучения деревьев решений для краткосрочного

прогнозирования электропотребления // Известия Тульского государственного университета. Технические науки. – 2023. – №. 11. – С. 226-231.

6. Aria M., Cuccurullo C., Gnasso A. A comparison among interpretative proposals for Random Forests //Machine Learning with Applications. – 2021. – V. 6. – P. 100094.

7. Korchagin, S. A., Gataullin, S. T., Osipov, A. V., Smirnov, M. V., Suvorov, S. V., Serdechnyi, D. V., & Bublikov, K. V. Development of an optimal algorithm for detecting damaged and diseased potato tubers moving along a conveyor belt using computer vision systems //Agronomy. – 2021. – V. 11. – №. 10. – P. 1980.

8. Kamyshova G., Osipov, A., Gataullin, S., Korchagin, S., Ignar, S., Gataullin, T., Terekhova N. & Suvorov, S. Artificial neural networks and computer vision's-based phytoindication systems for variable rate irrigation improving //IEEE Access. – 2022. – V. 10. – pp. 8577-8589.

9. Беспалова Н. В., Корчагин С.А., Сердечный Д.В., Селиверстов В.В. Анализ зарубежного опыта применения интеллектуальных методов в задачах защиты объектов критической информационной инфраструктуры финансового сектора // Инженерный вестник Дона. – 2024. – №. 5.
URL: ivdon.ru/ru/magazine/archive/n5y2024/9196

10. Андриянов Н. А., Дементьев В. Е., Ташлинский А. Г. Обнаружение объектов на изображении: от критериев Байеса и Неймана–Пирсона к детекторам на базе нейронных сетей EfficientDet //Компьютерная оптика. – 2022. – Т. 46. – №. 1. – С. 139-159.

11. Gajjar H., Sanyal S., Shah M. A comprehensive study on lane detecting autonomous car using computer vision //Expert Systems with Applications. – 2023. – V. 233. – P. 120929.

12. Корчагин С.А. Система компьютерного зрения для автоматической классификации нанокомпозитов // Информационно-измерительные и управляющие системы. – 2023. – Т. 21. – № 5. – С. 16-26.

13. Schneider, S., Taylor, G. W., Kremer, S. C., & Fryxell, J. M. Getting the bugs out of AI: advancing ecological research on arthropods through computer vision //Ecology Letters. – 2023. – V. 26. – №. 7. – pp. 1247-1258.

14. Lund B., Ma J. A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering //Performance Measurement and Metrics. – 2021. – V. 22. – №. 3. – pp. 161-173.

15. Проневич О. Б., Клокова А. П. Анализ UMAP–метода снижения размерности исходных данных в машинном обучении для прогнозирования отказов в локомотивном комплексе //Надежность. – 2022. – Т. 22. – №. 4. – С. 53-62.

References

1. Panishev I. A. Sibirskaya finansovaya shkola. 2024. V. 1. pp. 81-86.

2. Montina A., Wolf S. Physical Review A. 2023. V. 107(1). P. 012218.

3. Domova V., Vrotsou K. IEEE Transactions on Visualization and Computer Graphics. 2022. V. 29(8). pp. 3550-3568.

4. Begunkov V. I., Kovalev M. Ya. Informatika. 2023. V. 20(1). pp. 55-74.

5. Gorshenin A. Yu., Gritsay A. S., Denisova L. A. Izvestiya Tulskogo gosudarstvennogo universiteta. Tekhnicheskiye nauki. 2023. V. 11. pp. 226-231.

6. Aria M., Cuccurullo C., Gnasso A. Machine Learning with Applications. 2021. T. 6. P. 100094.

7. Korchagin, S. A., Gataullin, S. T., Osipov, A. V., Smirnov, M. V., Suvorov, S. V., Serdechnyi, D. V., & Bublikov, K. V Agronomy. 2021. V. 11(10). P. 1980.

8. Kamyshova G., Osipov, A., Gataullin, S., Korchagin, S., Ignar, S., Gataullin, T., Terekhova N. & Suvorov, S. IEEE Access. 2022. V. 10. pp. 8577-8589.

9. Bespalova N.V., Korchagin S.A., Serdechny D.V., Seliverstov V.V. Inzhenernyj vestnik Dona. 2024. №5. URL: ivdon.ru/ru/magazine/archive/n5y2024/9196

10. Andriyanov N. A., Dementyev V. E., Tashlinskiy A. G. Kompyuternaya optika. 2022. V. 46(1). pp. 139-159.

11. Gajjar H., Sanyal S. Shah M. Expert Systems with Applications. 2023. V. 233. P. 120929.

12. Korchagin S.A. Informatsionno-izmeritelnyye i upravlyayushchiye sistemy. 2023. V. 21(5). pp. 16-26.

13. Schneider, S., Taylor, G. W., Kremer, S. C., & Fryxell, J. M. Ecology Letters. 2023. V. 26(7). pp. 1247-1258.

14. Lund B., Ma J. Performance Measurement and Metrics. 2021. V. 22(3). pp. 161-173.

15. Pronevich O. B.. Klokova A. P. Nadezhnost. 2022. T. 22(4). pp. 53-62.

Дата поступления: 9.01.2025

Дата публикации: 25.02.2025