УДК 004.8

# АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБРАБОТКИ И АНАЛИЗА БОЛЬШИХ ДАННЫХ: МАТЕМАТИЧЕСКИЕ МОДЕЛИ И ИХ ПРИМЕНЕНИЕ

***Сураева Е. А., студ., Озерина Л. Ю., студ.,***
***Благодатский П. В., к.э.н., Крючкова А. С., доц.***

*Российский университет транспорта,*
*г. Москва, Российская Федерация*

Реферат. *В статье рассматриваются основные алгоритмы машинного обучения, используемые для обработки и анализа больших данных. Уделено внимание математическим моделям, лежащим в основе этих алгоритмов, а также их практическому применению в различных областях, включая бизнес, медицину и социальные науки. Особое внимание уделяется таким методам, как регрессионный анализ, деревья решений, ансамблевые методы и нейронные сети, а также их эффективности в задачах предсказания и классификации. Исследуется также влияние качества данных и выбор алгоритмов на результаты анализа. В заключение подчеркивается значимость машинного обучения как инструмента для извлечения полезной информации из больших массивов данных и его роль в принятии обоснованных решений.*

Ключевые слова: машинное обучение, большие данные, математические модели, регрессионный анализ, деревья решений, ансамблевые методы.

В современном мире мы находимся в непрерывно растущем потоке информации, который породили бесчисленные источники данных, такие как социальные сети, устройства интернета вещей, электронная коммерция и много другое. Ускоренный темп жизни и постоянное взаимодействие человека с цифровыми технологиями создают огромные массивы данных, известные как большие данные. Эти данные высоко разнообразны, генерируются в реальном времени и часто находятся в неструктурированном виде. Их анализ и обработка стали ключевыми задачами для организаций всех типов, стремящихся извлечь ценные инсайты и повысить свою конкурентоспособность.

В этом контексте машинное обучение играет центральную роль. Это направление искусственного интеллекта предоставляет мощные инструменты для анализа и интерпретации больших массивов данных, позволяя не только идентифицировать закономерности, но и делать предсказания на основе имеющейся информации. Алгоритмы машинного обучения способны обучаться на данных, адаптироваться к новым условиям и самостоятельно делать выводы, что делает их незаменимыми в множестве отраслей. Будь то оптимизация бизнес-процессов, прогнозирование рыночного спроса или диагностика заболеваний в здравоохранении, возможность использовать машины для анализа данных кардинально изменила подход к решению сложных задач.

Для понимания горизонтов, которые открывает машинное обучение, важно сначала рассмотреть, что такое большие данные. Под большими данными понимаются объемы информации, которые настолько велики и сложны, что традиционные методы обработки данных не могут с ними справиться. Они характеризуются тремя основными «V» – объемом, разнообразием и скоростью генерирования [1].

Машинное обучение, в свою очередь, представляет собой набор методов и технологий, позволяющих компьютерам «учиться» из данных без явного программирования. Существует множество типов машинного обучения, включая обучение с учителем, обучение без учителя и обучение с подкреплением. Каждый из этих подходов предлагает разные методы для нахождения и использования скрытых закономерностей в данных. Например, методы обучения с учителем ориентированы на более структурированные данные, где имеются четкие метки, тогда как обучение без учителя может работать с данными, которые не имеют заранее определенных меток, помогая выявить скрытые группы или паттерны [2].

В основе машинного обучения лежат несколько ключевых математических понятий, включая линейную алгебру, теорию вероятностей и статистику. Эти области позволяют нам формулировать модели, которые могут обрабатывать и анализировать данные. Линейная алгебра, например, используется для работы с многомерными данными, где объекты

представляются в виде векторов, а их отношения друг к другу описываются матрицами. Теория вероятностей помогает оценивать неопределенность в данных и строить вероятностные модели, которые могут предсказывать вероятность событий.

Линейная регрессия – одна из самых простых и распространенных моделей в машинном обучении. Она предполагает линейную зависимость между независимыми переменными и зависимой переменной. Математически её можно описать уравнением:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n+\varepsilon,$$

где $y$ – зависимая переменная, $\beta_0$ – свободный член, $\beta_1,\beta_2,\ldots,\beta_n$ – коэффициенты, которые нужно найти, $x_1, x_2,\ldots, x_n$ – независимые переменные, а $\varepsilon$ – случайная ошибка.

Логистическая регрессия, в отличие от линейной, используется для задач классификации и оценивает вероятность принадлежности экземпляра к определенному классу. Модель представляет собой логистическую функцию:

$$P(y=1|x)=1/(1+exp(-(\beta_0+\beta_1x_1+\ldots+\beta_nx_n))).$$

Здесь $P(y=1|x)$ – вероятность того, что событийная переменная $y$ равна 1 при заданных $x$.

Деревья решений представляют собой еще один мощный инструмент в арсенале машинного обучения. Они работают по принципу разбиения пространства признаков на подмножества на основе значений атрибутов. Каждый узел дерева представляет собой условие на каком-то признаке, а ветви – результаты этих условий. Главной задачей является минимизация неопределенности, и для этого используется понятие энтропии или же критерий Джини [3].

При расчете энтропии в узле дерева можно использовать формулу:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i,$$

где $H(S)$ – энтропия подмножества $S$, $c$ – количество классов, а $p_i$ – вероятность появления класса $i$.

Ансамблевые методы объединяют несколько моделей для получения более точного прогноза. Random Forest использует множество деревьев решений для создания предсказания, делая акцент на голосовании между различными деревьями. Каждый «лес» обучается на случайных подмножествах данных, что обеспечивает устойчивость к переобучению и увеличивает точность.

Градиентный бустинг, с другой стороны, строит деревья последовательно, каждое из которых пытается исправить ошибки предыдущих моделей. Это достигается за счет минимизации функции потерь с помощью градиентного спуска, что делает метод очень мощным и гибким для различных типов задач.

Нейронные сети представляют собой один из наиболее прогрессивных подходов в машинном обучении, который основан на имитации работы человеческого мозга. Простая нейронная сеть состоит из входного слоя, одного или нескольких скрытых слоев и выходного слоя. Каждый нейрон выполняет простую вычислительную задачу, и результаты передаются следующему слою [4].

С глубокими нейронными сетями (глубоким обучением) возникает возможность работать с большими объемами данных и выявлять сложные паттерны. Эти модели применяются в таких областях, как обработка естественного языка, компьютерное зрение и многое другое. Обучение нейронных сетей осуществляется с использованием алгоритмов обратного распространения ошибки, что позволяет оптимизировать веса нейронов для уменьшения ошибки предсказания.

Машинное обучение находит применение в самых разных сферах, от бизнеса до медицины и социальных наук.

В бизнесе алгоритмы машинного обучения используются для анализа потребительского поведения, что позволяет компаниям лучше понимать своих клиентов и адаптировать предложения. Например, с помощью кластеризации можно сегментировать клиентов по их поведению при покупках, а предсказательные модели позволяют прогнозировать спрос на товары.

В медицине машинное обучение применяется для диагностики заболеваний и построения персонализированных лечебных планов. С помощью анализа медицинских изображений и данных о пациентах можно выявлять патологии, которые могли бы остаться

незамеченными при традиционном подходе [5].

Социальные науки тоже не остаются в стороне – те же методы используются для анализа социальных сетей и предсказания общественного мнения. Благодаря таким алгоритмам, как анализ тональности текста, исследователи могут оценивать общественные настроения и выявлять ключевые темы в дискуссиях.

Машинное обучение также активно внедряется в финансы, где его используют для предсказания рыночных трендов и оценки риска. В экологии его применяют для мониторинга изменений климата и распределения биологических видов, а в транспорте – для оптимизации логистических процессов и управления движением [6].

В заключение нашей статьи о применении алгоритмов машинного обучения для обработки и анализа больших данных, мы можем с уверенностью утверждать, что современные математические модели становятся краеугольным камнем в этой области. Эти модели не только позволяют эффективно обрабатывать огромные объемы информации, но и извлекать из них ценные инсайты, которые могут существенно повлиять на различные сферы – от бизнеса до медицины и науки.

Используя методы, основанные на теории вероятностей, линейной алгебре и статистике, мы смогли продемонстрировать, как алгоритмы машинного обучения способны адаптироваться к постоянно меняющимся данным и условиям. Например, применяя регрессионные и кластеризационные модели, разработчики и исследователи могут выявлять скрытые зависимости и паттерны, которые ранее были недоступны для анализа. Эти подходы открывают новые горизонты для прогнозирования и принятия обоснованных решений

Список использованных источников


1. Джункеев, У. Прогнозирование инфляции в России на основе градиентного бустинга и нейронных сетей // Деньги и кредит. – 2024. – Т. 83. № 1. – С. 53–76.
2. Петров, С. В. Использование метода «случайный лес» при построении моделей надежности // Информатика, моделирование, автоматизация проектирования (ИМАП–2022). XIV Международная научно–практическая конференция студентов, аспирантов и молодых ученых: сборник научных трудов. – Ульяновск, 2022. – С. 84–88.
3. Стрельников, В. Г., Трунов, А. С. Применение метода логистической регрессии для задачи классификации текстов судебных решений // Телекоммуникации и информационные технологии. – 2017. – Т. 4. № 2. – С. 75–78.
4. Трифонова, О. Н. Анализ методов поиска идей для решения проблем в бизнесе методом построения дерева проблем и дерева решений // Фундаментальные и прикладные исследования в современном мире. – 2015. – № 9 (2). – С. 131–135.


## *3.2 Экология и химические технологии*


УДК 677.027.4


### КРАШЕНИЕ ШЕРСТЯНОЙ ПРЯЖИ КОРНЯМИ SANGUISÓRBA


***Горохова А. В., студ, Скобова Н. В., к.т.н., доц.***

*Витебский государственный технологический университет,*
*г. Витебск, Республика Беларусь*



Реферат. *Рассмотрены способы подготовки растительного сырья (корней и корневищ кровохлебки (лат.Sanguisórba)) к экстрагированию: ультразвуковая обработка частей растения и ферментная отварка корней в среде нейтральных и кислых целлюлаз в сочетании с пектиназой. Проведен сравнительный анализ спектров волн растворов после экстрагирования, установлены значимые различия в применяемых способах подготовки сырья.*



Ключевые слова: экотехнологии, природные красители, шерстяная пряжа, корни кровохлебки.