ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ АНАЛИЗА АНОМАЛЬНОГО ПОВЕДЕНИЯ ТЕХНОЛОГИЧЕСКИХ ПАРАМЕТРОВ, ПРИ КЛАССИФИКАЦИИ ТЕХНОЛОГИЧЕСКИХ ОПЕРАЦИЙ, ЦИКЛА СТРОИТЕЛЬСТВА СКВАЖИНЫ

А.А. ШИБАЕВ,
генеральный директор
hq@lamt-s.spb.ru

И.Л. ШРАГО,
председатель совета директоров
hq@lamt-s.spb.ru

И.А. ВАСИНКИН,
главный программист
hq@amt-s.spb.ru

А.С. ЧЕРНЫШОВ,
инженер-программист 1к
hq@amt-s.spb.ru

ООО «ЗАО АМТ»
г. Санкт-Петербург,
199106, РФ

УДК 550.8.072

В статье представлена работа по внедрению методов машинного обучения для решения задачи классификации технологических операций, в сфере геолого-технологических исследований (ГТИ). Подробно описываются этапы по обучению, тестированию и отбору моделей искусственного интеллекта на основе выбранной метрики. Показана методика поиска аномальных значений во временных файлах со станции ГТИ АМТ-301 с помощью методов искусственного интеллекта (ИИ) и описательных статистик.

Ключевые слова: внедрение ИИ в ГТИ, методы машинного обучения в бурении, классификация технологических операций бурения, деревья решений, метрики оценки моделей

APPLICATION OF MACHINE LEARNING METHODS IN THE TASK OF ANALYSIS OF ANOMAL BEHAVIOR OF TECHNOLOGICAL PARAMETERS, IN THE CLASSIFICATION OF TECHNOLOGICAL OPERATIONS, OF THE WELL CONSTRUCTION CYCLE

A.A. SHIBAEV,
I.L. SHRAGO,
I.A. VASINKIN
A.S. CHERNYSHOV,

ZAO AMT LLC
St. Petersburg, 199106,
Russian Federation

The article presents the work on the implementation of machine learning methods for solving the problem of classifying technological operations in the field of geological and technological research (GTI). The stages of training, testing and selection of artificial intelligence models based on the selected metric are described in detail. A technique for searching for anomalous values in temporary files from the GTI AMT-301 station using artificial intelligence (AI) methods and descriptive statistics is shown.

Keywords: implementation of AI in mud logging, machine learning methods in drilling, classification of technological drilling operations, decision trees, model evaluation metrics

ООО «ЗАО АМТ» уже более 30 лет успешно разрабатывает, производит и поставляет Заказчикам аппаратно-программные комплексы контроля технологических процессов при строительстве скважин и тренажерные комплексы для подготовки производственного персонала, к ведению работ по строительству и эксплуатации скважин.

Последняя разработка, станции контроля семейства АМТ-301 – это модульная архитектура, адаптивность структуры и широкий круг решаемых задач:

- безопасная работа оборудования во взрывоопасных средах;
- выполнение требований Правил промышленной безопасности (Ростехнадзор № 534 от 15.12.2020);
- автономный контроль параметров требований Правил промышленной безопасности геолого-технологических исследований и газового каротажа (ГТИ);
- решения для мобильных буровых установок;
- решения для контроля капитального ремонта скважин;
- решения для контроля траектории скважины;
- решения для контроля цементирования скважины;
- собственные проверенные методики.

Станция ГТИ АМТ-301 имеет целью осуществление оперативного геологического и технологического контроля бурения вертикальных, наклонно-направленных и горизонтальных нефтяных и газовых скважин, позволяет получить полную и объективную информацию по скважинам, необходимую для управления бурением и его оптимизации, а также для разведки и освоения месторождений.

К информационным системам мониторинга АМТ относятся система удаленного мониторинга строительства скважин в реальном времени (ИС МСРВ) и информационная система мониторинга состояния оборудования верхнего привода (ИС СВП РВ).

Успешность удаленного мониторинга скважин определяется следующими факторами:

- потребностью своевременного получения службами контроля недропользователя оперативной, полной и достоверной информации о выполняемых технологических процессах и состоянии скважины в реальном режиме времени;
- возможностью профессиональной обработки и интерпретации информации, поступающей на удаленные центры, с последующей выдачей грамотных рекомендаций по управлению процессами на буровой;
- возможностью оперативно сопоставлять и анализировать полученные данные с других ранее пробуренных скважин.

Отдельное место среди разработок ООО «ЗАО АМТ» занимает станция контроля

параметров бурения «ВОСТОК-6». Разработанная совместно с учеными федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский горный университет императрицы Екатерины II», созданная для работы в экстремальных условиях антарктического климата, СКПБ «ВОСТОК-6» использовалась при строительстве скважины и вскрытии поверхности подледникового озера «Восток». Она безотказно снабжала буровую бригаду оперативными данными, предупреждала о возможности возникновения нештатных ситуаций, и предупредила о достижении поверхности озера.

В настоящее время ООО «ЗАО АМТ» занимается внедрением методов машинного обучения в задачи геолого-технологических исследований, которые призваны улучшить качество результатов компьютерного анализа первичной (технологической, геологической и геохимической) информации.

Существует достаточно большое количество различных алгоритмов машинного обучения, различающихся по решаемым ими задачам, по принципу работы, по линейности и нелинейности. Например, модель может быть представлена одной прямой или гиперплоскостью, а может представлять собой дерево решений или другой набор правил. Нейронные сети также являются методами машинного обучения, однако, их архитектура наиболее сложна, и они формируют отдельный подраздел, называемый «глубокое обучение».

Самыми популярными методами машинного обучения являются методы, основанные на построении решающих деревьев. Одним из самых базовых алгоритмов является классическое дерево решений – Decision Tree Classifier (DTC) [1], которое можно применить на начальном этапе любого исследования для выявления и анализа каких-либо закономерностей, поскольку данный алгоритм имеет простую интерпретацию.

Более продвинутые алгоритмы используют ансамбли деревьев решений, например, Random Forest Classifier (RFC) [2] может взвешивать предсказания 100, 500 и 1000 разных деревьев. Используемая нами в дальнейшем, свободно распространяемая, библиотека sklearn [3], для языка Python, позволяет строить различные модели машинного обучения, настраивать параметры этих моделей, а также визуализировать их работу. Например, визуализация дерева решений в игрушечной задаче классификации сортов ириса приведена на рис. 1.

Ошибку неглубоких деревьев, дающих неточные предсказания, при обучении понижают посредством алгоритмов градиентного бустинга, применяемых в различных сферах анализа данных и являющихся наиболее продвинутыми при работе с табличными данными.

Существуют 3 наиболее популярные реализации градиентного бустинга: LightGBM [4], разработанный Microsoft; XGBoost [5]; CatBoost [6], разработанный Яндекс.

Апробирование и последующая реализация ограничилась моделями RFC, XGBoost и CatBoost в задаче определения технологических операций. Это связано с большой ее востребованностью у заказчиков сервиса ГТИ.

Для обучения и тестирования использовали исторические данные бурения за несколько лет – порядка 10 миллионов. Поскольку каждый класс операций имеет различное количество экземпляров во всем наборе, то для того, чтобы сбалансировать обучающую выборку, было выбрано по 20 тысяч данных на каждую

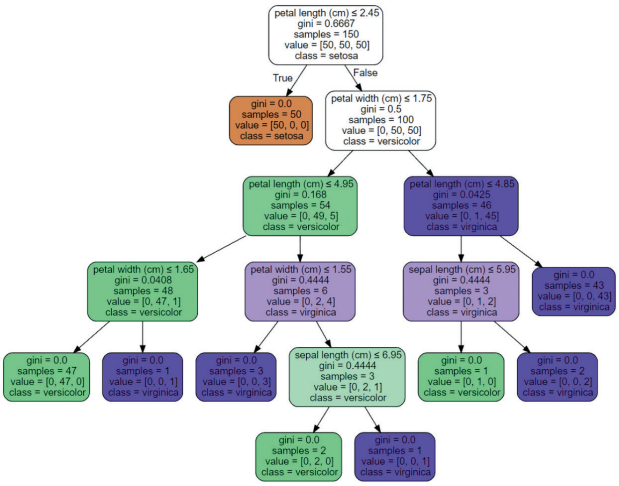

Рис.1. DTC при классификации сортов ириса

технологическую операцию. С учетом того, что данные в небольшом временном интервале слабо различаются между собой, около 8 тысяч данных пришлось на обучение и 12 тысяч на тестирование по каждому классу отдельно. Увеличение набора обучающей выборки приводило к переобучению на начальных этапах исследования.

Современные методы анализа данных позволили создать стратифицированную и репрезентативную выборку, с более низкой размерностью и без потери важной для нас информации. Ввиду большого объема данных, были устранены те, пропущенные значения в которых превышали 30 %. В оставшихся случаях пропущенные значения заменили на медианные.

Для контрольного тестирования использовались данные по скважине, не используемые при обучении.

Размерность, используемого моделью пространства признаков была ограничена списком самых необходимых: глубина скважины, вес на крюке, давление нагнетания, обороты бурового инструмента, расстояние между долотом и забоем.

Впоследствии, после завершения этапа апробации и отладки кода, планируется возврат к полноразмерному пространству признаков для возможного улучшения классификатора.

После подготовки данных для обучения была определена метрика для оценивания качества модели.

В задачах машинного обучения, для оценки качества моделей и сравнения различных алгоритмов, используются специальные метрики, такие как точность (precision) и полнота (recall). Точность можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а полнота показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм.

Отдельно по этим метрикам невозможно оценить качество получившейся модели, а только при совокупном их использовании. Удобной для интерпретации метрикой, учитывающей ошибки первого и второго рода, которая вычисляется как гармоническое среднее между точностью и полнотой – (f1) (1). f1 позволяет найти оптимальный баланс между этими двумя метриками.

$$f1=2/(recall^{-1} + precision^{-1}) \quad (1)$$

Для первого выявления закономерностей применили подход RFC по той причине, что он дает предсказания более качественно, чем DTC. Значение метрики f1 составило 0,82 на тестовых данных.

Для повышения качества предсказания использовался подбор параметров модели таких как глубина деревьев, количество деревьев и других с помощью метода случайного поиска. Поскольку полный перебор всевозможных параметров может быть слишком большой, а время работы алгоритма увеличивается пропорционально N!, где N количество подбираемых параметров, то случайный поиск перебирает параметры, оценивая наилучший рост значения метрики на кросс валидации. После этого значение метрики f1 выросло и составило 0,87 на тестовых данных.

На следующем этапе были использованы реализации градиентного бустинга XGBoost и CatBoost, поскольку данные модели являются более продвинутыми. XGBoost и CatBoost являются конкурентами, постоянно совершенствуются и показывают достаточно сравнимую точность при работе.

Ввиду того, что применение метода случайного поиска улучшило качество модели и значение метрики f1 при переборе параметров модели, то он же использовался и для обучения алгоритмов градиентного бустинга. По итогам применения новых моделей значения метрик f1 составили 0,927 и 0,930 на тестовых данных с использованием XGBoost и CatBoost соответственно.

После этого, поскольку величина метрики перестала реагировать на улучшение модели, полученный результат был протестирован на «сырых» данных, упомянутой ранее дополнительной скважины, содержащей около 1,5 миллиона записей. Значение метрики, при этом, составило 0.891 и 0.895 для XGBoost и CatBoost, соответственно. Тем самым мы подтвердили, что XGBoost и CatBoost практически равносильно справляются с задачей классификации, но CatBoost имеет более гибкую настройку обучения и тестирования, более простую интеграцию в различные системы, а также больше возможностей по визуализации.

Несмотря на достаточно высокое значение метрики f1, результат классификации все равно является недостаточно хорошим, потому что не достиг желаемых 0,95.

Попытка использовать разные агрегации данных за различные временные промежутки и значение предыдущей технологической операции для увеличения показателя метрики f1 моделей результатов не улучшила.

Результат тестирования обученных моделей в приложении «прогнозно-аналитические модели определения технологических операций (искусственный интеллект)», изображен на рис. 2.

Фактор, способный существенно снижать качество модели – это различного рода аномальные ситуации. Большое количество различных аномалий может негативно влиять как на обучение, поскольку модель изначально обучается на ошибочных данных, так и на тестирование, поскольку такие значения могут не соответствовать определенным им правилам и понижать значение метрики.

Например, методы, основанные на деревьях решений, разделяют данные в узлах на более простые части путем сравнения с пороговым значением. Данные, которые не могут быть объяснены деревом решений, т.е. различные пограничные случаи, случаи, выпадающие

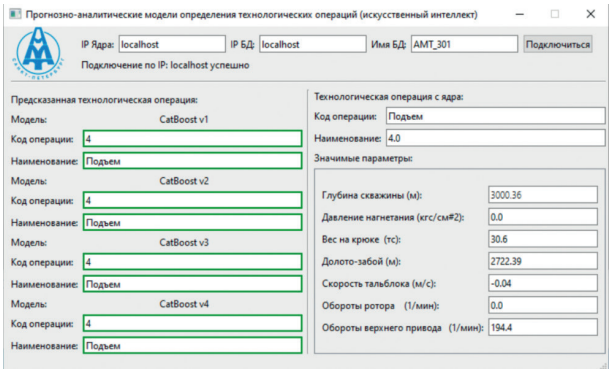

Рис. 2. Прогнозно-аналитические модели определения технологических операций (искусственный интеллект)

из общего распределения – выбросы или случаи, когда одинаковые серии данных относятся к разным классам, также считаются аномальными. Однако применение методов машинного обучения позволяет достаточно быстро находить такого рода ситуации, поскольку мы можем наблюдать за поведением моделей на данных.

Преимуществом этого метода является то, что он не требует предварительной обработки данных и предоставляет простые интерпретируемые правила для определения аномалий.

Для начала, при тестировании классификатора, обученного решать определенную задачу, основываясь на различных метриках и ошибках, определяются проблемные ситуации. Это могут быть различного рода ошибки и неточности в данных, высокие или низкие значения с датчиков и даже подлог данных. После обнаружения выбросов или аномальных значений применяются методы сводных статистик. Сводные статистики являются классическими статистическими методами, которые используются для обнаружения аномалий, основываясь на анализе распределения данных.

Методы сводных статистик используются для анализа распределения данных и поиска значений, которые выбиваются за пределы нормального диапазона. Они основываются на вычислении стандартных статистических показателей, таких как среднее, стандартное отклонение, медиана, минимальное и максимальное значения, квартили и т.д. Значения, которые находятся далеко от среднего или имеют большое стандартное отклонение, могут считаться аномальными. Одним из простых, но очень наглядным, способов анализа аномальных значений является график BoxPlot.

График BoxPlot позволяет очень компактно и наглядно представлять порядковые статистики одномерного закона распределения: квартили, медиану, наблюдаемые минимальное и максимальное значение выборки, а также отображать выбросы (рис. 3.).

В связи с неуспешными попытками улучшить качество классификации модели было принято решение реализовать алгоритм метода сводных статистик для детального анализа входных данных, на предмет обнаружения аномалий. Это позволило выявить, что одним из факторов, понижающих точность модели на тестах, явилась неучтенная дополнительная настройка по определению отрыва инструмента от забоя в процессе бурения. Отрыв долота и перемещение в интервале меньше указанного оператором значения «Долото-Забой» технологическая операция определяется как Подрыв, и только после превышения данного расстояния как Проработка (вверх или вниз, с вращением

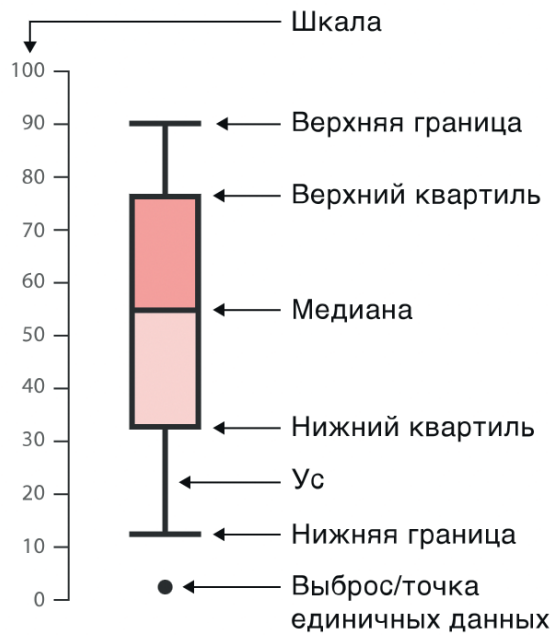

Рис. 3. BoxPlot

или без), до момента постановки долота на забой. Также было замечено, что модель допускает ошибки при попытке различать Проработку и Промывку с вращением и без.

Для этого по каждому из классов были построены графики BoxPlot на основании распределения параметра вращение. Притом, что при Проработке и Промывке допускается минимальное значение вращения, график показывал наличие высоких значений вращения в тех условиях, где их не должно быть. Примеры аномально высоких и аномально низких значений показаны на рис. 4 и рис. 5.

Подобные большие скачки могут отражать такие ситуации, как неправильное обращение с оборудованием, выставление ошибочно больших минимально рабочих значений, а также неисправность в работе датчика.

На основе проделанной работы и тестирования реализованных алгоритмов можно сделать вывод о применимости искусственного интеллекта в задаче определения технологических операций, что позволит отказаться от традиционных алгоритмов, использующих много ручных настроек граничных значений параметров, одновременно с улучшением качества ее решения.

Следующим шагом будет ориентация разработанного программного обеспечения на другие задачи сервиса ГТИ.

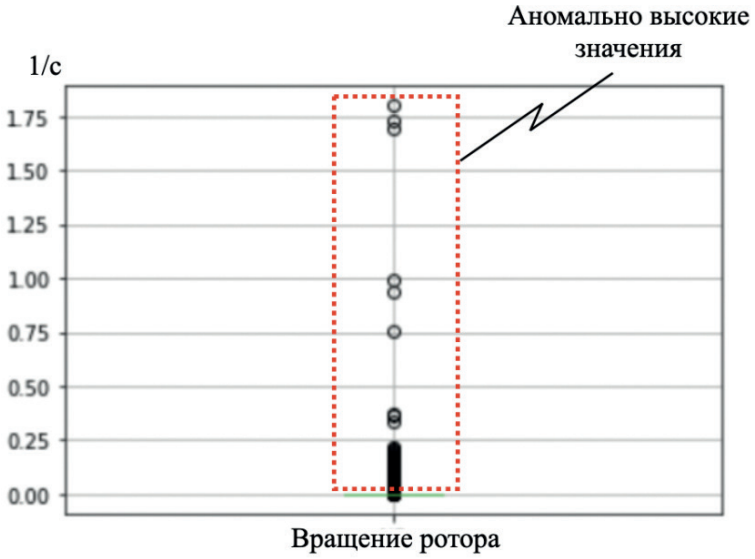

Рис. 4. Аномально высокие значения оборотов при проработке без вращения

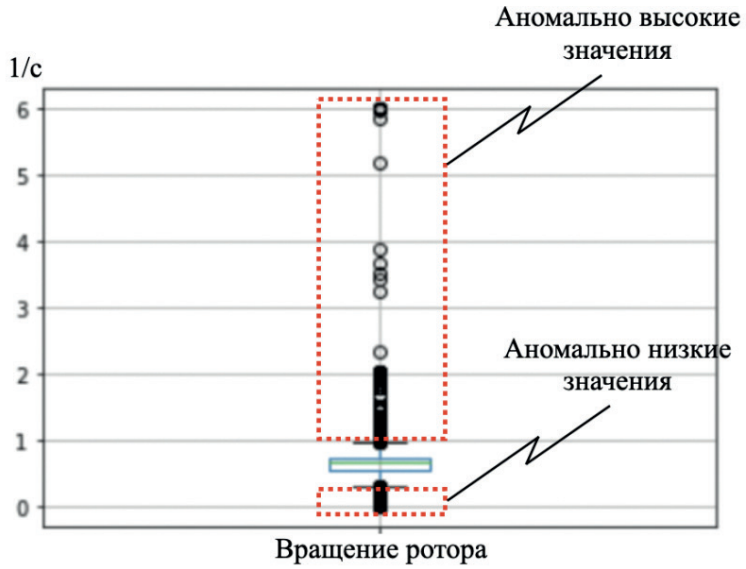

Рис. 5. Аномально высокие и аномально низкие значения оборотов при проработке с вращением

Литература

1. *Decision Trees.* [Online]. Available: https://scikit-learn.org/stable/modules/tree.html
2. *RandomForestClassifier.* [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.
3. *RandomForestClassifier.html.sklearn* [Online]. Available: https://scikit-learn.org.
4. *LightGBM. Welcome to LightGBM's documentation!* [Online]. Available: https://lightgbm.readthedocs.io/en/latest/.
5. *XGBoost. XGBoost Documentation.* [Online]. Available: https://xgboost.readthedocs.io/en/stable/.
6. *CatBoost. CatBoost is a high-performance open source library for gradient boosting on decision trees* [Online]. Available: https://catboost.ai.

References

1. Decision Trees. Available at: https://scikit-learn.org/stable/modules/tree.html
2. RandomForestClassifier. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.
3. RandomForestClassifier.html.sklearn. Available at: https://scikit-learn.org.
4. LightGBM. Welcome to LightGBM's documentation! Available at: https://lightgbm.readthedocs.io/en/latest/.
5. XGBoost. XGBoost Documentation. Available at: https://xgboost.readthedocs.io/en/stable/.
6. CatBoost. CatBoost is a high-performance open source library for gradient boosting on decision trees. Available at: https://catboost.ai.