Оригинальная статья / Original Paper

https://doi.org/10.30686/1609-9192-2023-5S-10-17

Анализ методов подготовки и преобразования информации, поступающей в хранилища данных для эффективного управления горнотехнической системой

В.Н. Захаров, Д.А. Клебанов✉, М.А. Макеев, Д.Н. Радченко

Институт проблем комплексного освоения недр им. академика Н.В. Мельникова Российской академии наук, г. Москва, Российская Федерация

✉ Klebanov_d@ipkonran.ru

Резюме: В статье оценены существующие методы сбора и работы с данными в промышленности, а также предложен альтернативный вариант работы с данными для управления горнотехническими системами на различных этапах их функционирования. Показано, что формируемый объем информации, поступающей с различной частотой и требующей специализированных методов обработки, структурирования и анализа, образует систему больших данных и служит для повышения эффективности реализации геотехнологических процессов. Приведена возможная унифицированная структура сбора данных от цифровых источников горнотехнической системы. Дан анализ возможных методов обработки данных для их аналитики и поиска неявных зависимостей и решения задач прогнозной аналитики. Предложены инструменты сбора и хранения данных для создания унифицированной системы анализа больших данных при управлении горнотехническими системами. Заявлено, что для создания унифицированной аналитической системы сбора цифровых данных горнотехнической системы необходимо использование типовых промышленных протоколов сбора и хранения данных, например MQTT, применяемых в качестве стандарта в промышленных системах интернета вещей – IIoT в соответствии с ISO ISO/IEC 20922:2016. Для решения задач хранения требуется применение типовой архитектуры брокера очередей, а также инструмента работы с временными рядами, необходимого для применения методов машинного обучения и работы с большими данными. Предложенный в статье подход к классификации данных с точки зрения скорости их получения позволяет стандартизировать принципы работы с данными. В связи с тем что объем передаваемых данных не зависит от частоты формирования информации от цифрового источника, предлагается передавать все формируемые данные от цифрового источника для последующего поиска неявных зависимостей между ними. Отмечено, что применение конкретных методов и алгоритмов анализа данных горнотехнической системы прежде всего зависит от поставленной задачи, часто формируемой в виде гипотезы, которая должна быть проверена за счет выявления неявных зависимостей между разными источниками данных. Для повышения эффективности управления горнотехнической системой на всех этапах освоения месторождений предлагается применять подход ELT, что может стать важным преимуществом с точки зрения управления технологическими процессами горнотехнической системы в будущем.

Ключевые слова: горнотехническая система, большие данные, архитектура информационных систем, системы диспетчеризации, обработка данных, методы обработки информации

Благодарности: Статья написана в рамках выполнения гранта Российского научного фонда №22617600142, https://rscf.ru/project/22617600142/

Для цитирования: Захаров В.Н., Клебанов Д.А., Макеев М.А., Радченко Д.Н. Анализ методов подготовки и преобразования информации, поступающей в хранилища данных для эффективного управления горнотехнической системой. *Горная промышленность*. 2023;(5S):10–17. https://doi.org/10.30686/1609-9192-2023-5S-10-17

Analysis of methods to prepare and transform information entering data repositories for effective management of the mining system

V.N. Zakharov, D.A. Klebanov ✉, M.A. Makeev, D.N. Radchenko

Institute of Comprehensive Exploitation of Mineral Resources of Russian Academy of Sciences, Moscow, Russian Federation

✉ Klebanov_d@ipkonran.ru

Abstract: TThe article assesses the existing methods of data collection and processing in the industry and proposes an alternative option of data processing to manage mining engineering systems at various stages of their operation. It is demonstrated that the formed volume of information that comes with different frequency rates and requires dedicated processing, structuring and analysis methods forms a system of big data and serves to enhance the efficiency of implementing geotechnical processes. A possible unified structure of data acquisition from digital sources of mining engineering system is presented. An analysis

is provided of possible methods to process data for analytical purposes and for searching implicit dependencies and solving problems of predictive analytics. Tools for data collection and storage are proposed to create a unified system of big data analysis for management of mining engineering systems. It is stated that in order to create a unified analytical system for collection of digital data from a mining system, it is necessary to use standard industrial protocols for data collection and storage, for instance MQTT, used as an industry standard in the Industrial Internet of Things (IIoT) systems in accordance with requirements of ISO/IEC 20922:2016. Storage requires the use of a conventional queue broker architecture, as well as a tool for working with the time series which is required to apply machine learning and big data methods. The approach to data classification in terms of the data acquisition speed proposed in the article makes it possible to standardize the data handling principles. Since the volume of transmitted data does not depend on the frequency of acquiring information from a digital source, it is proposed to transmit all the generated data from a digital source for subsequent search of implicit dependencies between the data. It is noted that application of specific methods and algorithms to analyze data of a mining system depends primarily on the task set which is often formed as a hypothesis to be tested by identifying implicit dependencies between different data sources. In order to improve the efficiency of managing a mining system at all the stages of field development, it is proposed to apply the ELT approach, which can be an important advantage in terms of controlling the technological processes of the mining system in the future.

Keywords: mining system, big data, information systems architecture, dispatching systems, data processing, information processing methodsд

Acknowledgments: The article was written within the framework of the Russian Science Foundation Grant No.22617600142, https://rscf.ru/project/22617600142/

For citation: Zakharov V.N., Klebanov D.A., Makeev M.A., Radchenko D.N. Analysis of methods to prepare and transform information entering data repositories for effective management of the mining system. *Russian Mining Industry.* 2023;(5S):10–17. https://doi.org/10.30686/1609-9192-2023-5S-10-17

Введение

Горнотехническая система по мере освоения участка недр является источником данных больших объемов, характеризующихся различной вариативностью и ценностью. Структура таких данных определяется перечнем применяемых систем автоматизации, горно-геологических и систем планирования, сейсмического и геомеханического контроля, 3D данные оборудования сканирования, показателями различного рода датчиков, установленных на оборудовании и других систем [1].

Современное горнодобывающее предприятие формирует до 150 ТБ информации в год, создавая ежедневно на их основе более 200 производных показателей по производительности, качеству, объемам, техническому состоянию и пр., что позволяет классифицировать собираемые и расчетные параметры как Большие данные [2], использование которых служит для повышения эффективности реализации геотехнологических процессов [3].

Различная вариативность и значительные объемы получаемых данных вызывают необходимость подготовки и преобразования собранной информации, что требует применения специализированных методов, позволяющих создавать дополнительную ценность для горнодобывающего предприятия и разрабатывать инновационные системы управления техникой и оборудованием, осуществлять оценку и прогнозирование рисков промышленной безопасности и экологии, выявляя нелинейные зависимости между различными переделами добычи и переработки. В настоящее время поставщики технологического оборудования и ИТ систем предлагают разные методы сбора и подготовки данных, которые реализуют конкретный функционал, при этом ограничивается возможность использования собранных данных для применения в более широких задачах управления горнотехническими системами. Например, отдельные системы высокоточной навигации и контроля параметров бурения собирают и хранят только данные о пробуренных скважинах, сетке бурения и состоянии бурового станка, необходимые для выявления целевых режимов работы буровых станков, однако телеметрия параметров бурения совместно с данными по производительности экскавации, мониторинга гранулометрического состава может быть использована для районирования блоков по крепости, моделирования развала, оптимизации измельчения на всех технологических этапах управления горными работами до процесса дробления при управлении процессами обогащения.

Таким образом, если сформулировать методологию по работе с данными при управлении горнотехнической системой на долгосрочном горизонте, собранную информацию возможно использовать более эффективно, при этом позволяя находить зависимости между технологическими процессами отработки массива, используя единые базы данных, не ограничиваясь внедрением отдельных систем, направленных на улучшение конкретного технологического процесса. Однако в таком случае необходима стандартизация методов сбора, извлечения, конвертации, оценки и сжатия данных, которая позволит унифицировать разработку информационных систем и предоставить компаниям готовый информационный ландшафт для внедрения технологий при управлении горнотехническими системами.

Анализ принципов работы с данными, классификация источников, типов, скорости формирования и объемов данных, генерируемых горнотехнической системой и внешней средой по мере освоения участка недр

Задача стандартизации методов обработки и использования данных является ключевым вызовом горнодобывающей отрасли, что отражено в отчетах[1] [2] одной из ведущих мировых некоммерческих организаций по исследованию применения технологий в горной отрасли – Global Mining Guidance Group (GMG), в которую входят представители недропользователей, научного сообщества и университетов,

1 Guideline for Sharing Open Data Sets in Mining (GMG13-AI-2022). Global Mining Guidelines Group (2022).

компании – поставщики техники и информационных систем, ведущие консалтинговые компании. Эксперты GMG заключают, что в настоящее время не разработаны единые принципы работы с данными, что влечет за собой проблемы использования данных, собираемых и обрабатываемых системами от различных поставщиков. Принципы работы с данными частично описаны в стандарте ISO 81346 [2], который определяет общие подходы к стандартизации обработки данных в промышленных системах, но не учитывает специфику работы горнотехнических систем.

Прежде всего в основу работы с данными горнотехнической системы должна быть заложена классификация имеющихся в ней источников информации. Классификация источников информации горнотехнической системы по признаку объекта-получения представлена в статье [2]. В работе [2] отмечено, что источники включают внутренние данные горнотехнической системы (отрабатываемые массивы, технологическая среда, оборудование) и внешние данные (социум – влияние человека, окружающая природная среда). При этом типы данных могут быть разделены на данные, получаемые непосредственно от взаимодействия с объектами (датчики), данные опосредованного наблюдения (видео и фото данные, данные от лидаров) и данные по результатам обработки [1]. Однако с точки зрения методов хранения и обработки информации данная классификация не позволяет сформировать единый унифицированный подход к формированию архитектуры построения информационных систем. Предлагаемая в статье [2] классификация типов данных учитывает способ хранения и дальнейшее применение алгоритмов обработки данных:

1. Временные ряды – двухмерные данные с указанием значения параметра, времени получения данных (показания датчиков, состояния оборудования, данные по закупленным запчастям и пр.).
2. Геопространственные данные – трехмерные данные, характеризующиеся наличием трех измерений и указанием характеристик трехмерной точки (координаты долгота, широта, получаемая от навигационных устройств тахеометром, либо расстояние, горизонтальный и вертикальный угол, получаемые от лазерных, радарных систем).
3. Фото, видео и звуковые данные, получаемые от стационарных, носимых систем или летательных аппаратов, систем аудиозаписи (диспетчерские переговоры, совещания, межцеховая связь, участковая связь).

Важно отметить, что получаемые данные различных типов могут формироваться с различной дискретностью и скоростью получения от различных источников, а также различаются методом получения – автоматически по данным автоматизированных систем (датчики, видеокамеры, тахеометры) или посредством ручного ввода данных персоналом предприятия (справочная информация, график работы водителей, проведенные ремонтные работы и пр.).

В части цифровой обработки сигналов все параметры, получаемые в информационных системах, являются дискретными, так как данные, получаемые даже в виде аналогового сигнала в ходе непрерывных измерений, при сохранении в базе данных представляют собой дискретные записи. Поэтому с точки зрения применения методов использования данных в информационных системах при управлении горнотехническими системами необходимо классифицировать данные с учетом скорости получения информации:

1. Непрерывные данные, получаемые в режиме реального времени от датчиков и устройств, например, видеокамер или радаров геомониторинга, с частотой от нескольких раз в секунду до нескольких минут.
2. Данные, получаемые автоматически от датчиков в режиме событий, например, датчик формирует событие только при превышении параметра или фото водителя, получаемое при отвлечении внимания.
3. Периодические данные, получаемые с разной частотой от ручных измерений (данные по объёму склада от маркшейдера, данные лабораторных исследований или ввода данных в ERP систему о покупке запасной части).

При этом объем информации от различных источников данных напрямую не зависит от скорости получения данных, так, например, маркшейдерская съемка, произведенная дроном, получаемая один раз в неделю, может превышать десятки гигабайт, а данные телеметрии о состоянии техники или производственных систем, таких как системы диспетчеризации горно-транспортного комплекса от 100 единиц техники, составляют не более 15 Гб в год. Классификация с точки зрения объема получаемой информации имеет значение для создания систем хранения данных, при этом методы обработки данных зависят в основном от их типа, а не объема получаемой информации.

Методология эффективного сбора и хранения больших данных на горнодобывающем предприятии, анализ существующих технических и инфраструктурных ограничений и путей развития технологий сбора и хранения данных

В настоящее время большинство используемых информационных систем на горнодобывающих предприятиях используют структурированные базы данных и связи между ними, в которых хранится информация, требуемая для реализации конкретного функционала отдельной системы. Так называемые Data Warehouse хранят не всю информацию, получаемую из исходных источников, а предварительно обработанную и структурированную [4], при этом используют технологию обработки данных ETL – сокращение от *extraction (извлечение),* transformation (преобразование) и *loading (загрузка)*. Это процесс сбора «сырых» данных из раздельных источников, последующей передачи в промежуточную базу данных для преобразования и загрузки подготовленных данных в единую целевую систему. Процесс обработки данных по принципу ETL представлен на рис. 1.

Инструменты технологии ETL используются для интеграции данных, чтобы удовлетворить требованиям систем управления реляционными базами данных и/или традиционных хранилищ данных с поддержкой OLAP (online analytical processing, аналитической онлайн-обработки) [4]. Инструменты OLAP и запросы (SQL) требуют, чтобы массивы данных структурировались и стандартизировались при помощи серии преобразований, выполняемых до того, как данные попадут в хранилище. Эта методика возникла в 1970-х, когда компании начали использовать множественные репозитории данных для работы с разными типами бизнес-информации. С ростом объёмов разрозненных баз данных происходил рост потребности в консолидации

2 ISO 81346-12:2018 Industrial systems, installations and equipment and industrial products – Structuring principles and reference designations. Available at: https://www.iso.org/standard/63886.html

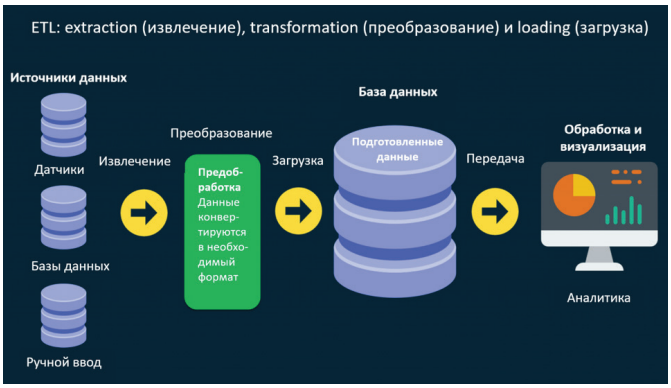

Рис. 1
Процесс обработки данных по принципу ETL

Fig. 1
Data processing according to the ETL principle

всех этих данных в централизованную систему. ETL возник как решение этой проблемы и стал стандартным методом интеграции данных. С конца 1980-х, когда появились хранилища данных, и до середины 2000-х ETL был основным способом создания баз данных, используемых как основа для бизнес-аналитики (business intelligence BI) [4].

С дальнейшим ростом объёмов и типов данных ETL становился не только довольно неэффективным, но более дорогостоящим и трудозатратным[3]. Благодаря взрывному росту количества источников информации и всё усиливающейся потребности обработки огромных массивов данных для целей бизнес-аналитики и аналитики больший данных стала возрастать популярность технологии ELT – альтернативы традиционному методу интеграции данных, отличающейся подходом: extraction (извлечение), loading (загрузка) и transformation (преобразование). Процесс обработки данных по принципу ELT представлен на рис. 2.

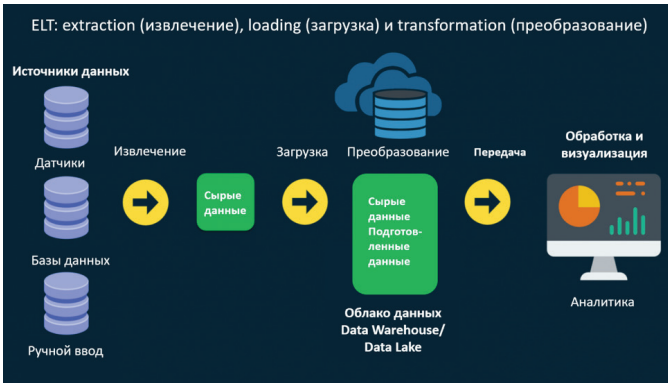

Рис. 2
Процесс обработки данных по принципу ELT

Fig. 2
Data processing according to the ELT principle

Исходные данные от всех источников информации загружаются в так называемые Data Lake – озера данных, которые принимают любые файлы всех форматов[4]. В этом случае источник данных тоже не имеет никакого значения. Уже потом, когда данные сохранены, с ними можно работать – извлекать по определенному шаблону в классические базы данных или анализировать и обрабатывать прямо внутри data lake (озеро данных). Ключевое отличие озер данных от обычных баз данных – это наличие структуры. В базах данных хранятся только четко структурированные данные, а в озерах – неструктурированные, никак не систематизированные и неупорядоченные данные. При этом процесс анализа не влияет на сами данные в озере – они так и остаются неструктурированными, чтобы их было также удобно хранить и использовать для других целей.

Озера данных часто используют для хранения важной информации, которая пока не используется в аналитике[5]. Например, при функционировании горнотехнических систем в озерах данных может храниться информация о структуре и свойствах горных массивов, информация, собираемая от систем телеметрии горного оборудования или собираемая сейсмодатчиками, а также другая информация, характеризующая элементы горнотехнической системы. Также озера данных используют для хранения данных, которые кажутся бесполезными ввиду отсутствия явных зависимостей между технологическими процессами на момент освоения месторождения, но, вероятно, могут стать важным преимуществом с точки зрения управления технологическими процессами горнотехнической системы в будущем. Для горнодобывающих предприятий такими данными могут являться показатели метеостанций, информация о геодинамической, солнечной, сейсмической активности, а также данные о профессиональных заболеваниях на длительном промежутке времени для конкретной горнотехнической системы.

Таким образом, для сбора данных и решения широкого спектра задач оптимизации работы горнодобывающего предприятия с развитием методов обработки больших данных преимуществом может служить использование систем ELT или Data Lake, однако ELT – относительно новая методология сбора и хранения, поэтому для неё существует меньше наработок и профессиональных компетенций. Такие инструменты и системы всё ещё находятся на ранней стадии развития, кроме того, поиск специалистов, знающих процесс внедрения и эксплуатации процесса ELT, намного сложнее.

Таким образом, Data Lake является наиболее перспективным способом сбора и хранения данных и позволяет накапливать «данные будущих периодов», не характеризующиеся явной ценностью в текущий период функционирования горнотехнической системы.

За счет того, что в «озере» сбор данных осуществляется из различных источников, они не проходят первичной трансформации и агрегации, при анализе эффективности функционирования горнотехнических систем открываются возможности формулирования неявных гипотез и проверки их подлинности. Например, для оптимизации логистических схем и эффективного управления цепочками поставок – от более детального планирования и прогнозирования объема продаж до поставок концентрата в нужном количестве, в нужное время с минимальными затратами. Исследование[6] показывает, что компании, внедрившие Data Lake, на 9% опережают своих конкурентов по выручке. Проецируя такие данные на горную промышленность, можно говорить о перспективе повышения конкурентоспособности горнодобывающих компаний на мировом рынке.

3 Angling for insight in today's data lake, October 2017, Michael Lock, Senior Vice President, Analytics and Business Intelligence. Available at: https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf

4 Angling for insight in today's data lake, October 2017, Michael Lock, Senior Vice President, Analytics and Business Intelligence. Available at: https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf

5 Essential Guide to Data Lakes. Available at: https://www.matillion.com/uploads/pdf/essential-guide-to-data-lakes-designing-data-lakes-to-optimize-analytics.pdf

6 Industry Priorities, Challenges, and Collaborative Approaches: Report on the 2022 GMG Mine Operator Roundtables (Report). Global Mining Guidelines Group (2022).

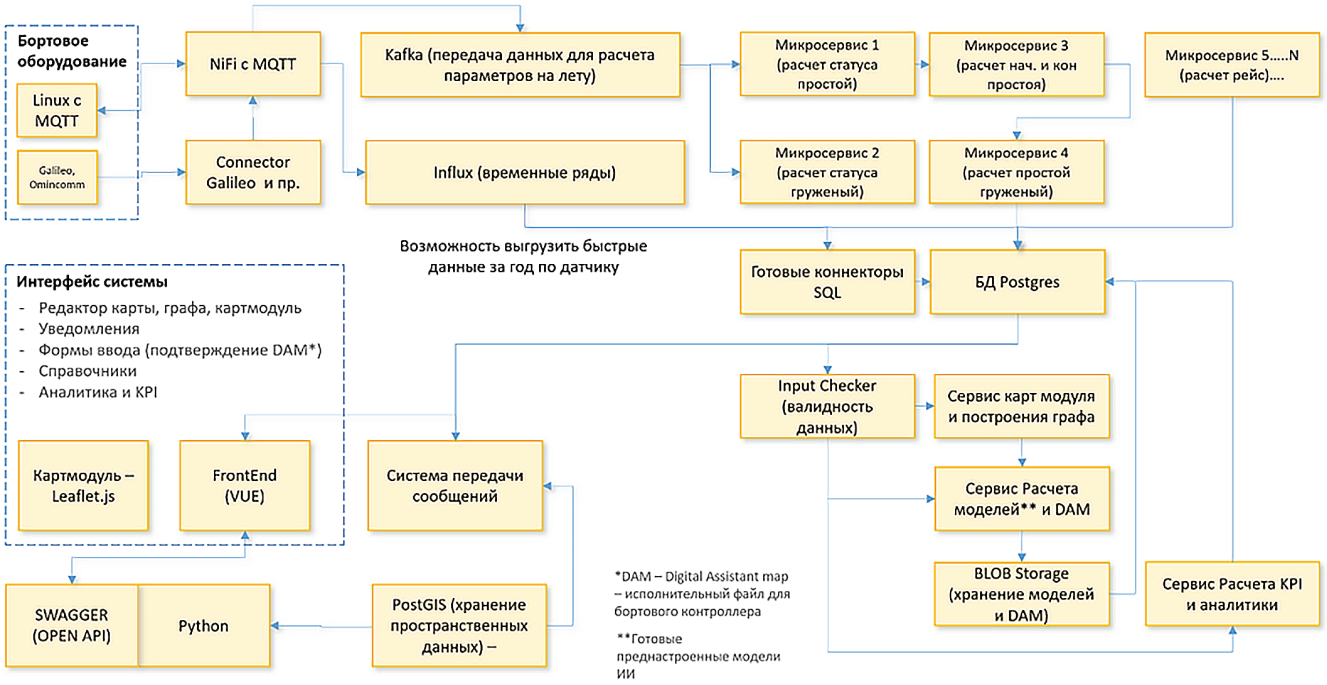

Рис. 3
Унифицированная аналитическая система сбора цифровых данных горнотехнической системы

Fig. 3
Unified analytical system to collect digital data of the mining system

Учитывая вышесказанное, унифицированная аналитическая система сбора цифровых данных горнотехнической системы, сформированная в соответствии с классификацией источников [2] и с учетом типов получаемых данных, представляется в виде схемы (рис. 3). Данная система отражает реализацию системы по принципу ETL для конкретного бизнес-приложения, получающего данные напрямую от устройств или из облака данных. Особенности создания облака данных описано в руководстве[7].

Унифицированная аналитическая система сбора цифровых данных горнотехнической системы подразумевает, что для сбора данных необходимо использование типовых промышленных протоколов сбора и хранения данных, например MQTT, применяемых в качестве стандарта в промышленных системах интернета вещей – IIoT в соответствии с ISO ISO/IEC 20922:2016[8].

Далее для решения задач хранения требуется применение типовой архитектуры в виде Kafka – брокер очереди сообщений[9], а также инструмента работы с временными рядами, необходимого для применения методов машинного обучения и работы с большими данными, Influx[10].

Кроме того, на основании полученной очереди сообщений, собранных от цифровых источников данных горнотехнической системы, должны быть организованы отдельные микросервисы, осуществляющие расчёт состояния или влияния одних данных технологического процесса на другие. Типовые инструменты хранения пространственных данных – POST Gis, реляционная база данных для хранения табличных данных – PostgreSQL.

Методы и алгоритмы обработки больших данных с учетом сознания интеллектуальных систем помощи принятия решений при функционировании горнотехнических систем

Вне зависимости от того, какую методологию сбора и хранения данных использует горнодобывающее предприятие на текущем этапе развития, возможно применять различные методы и алгоритмы обработки данных при управлении горнотехнической системой и создании интеллектуальных систем помощи принятия решений.

На основании анализа 4301 статей с 2016 по 2021 г. методы анализа больших данных были классифицированы на 2 группы [5]:

1. Контролируемое обучение (supervised learning).
2. Обучение без учителя (unsupervised learning).

Цель алгоритмов контролируемого обучения (обучения с учителем) состоит в том, чтобы спрогнозировать на основании нового набора данных известное заранее событие. В этом методе обучения набор входных и выходных данных изначально известен и найдена связь между ними при обучении системы. Главная цель обучения с учителем – моделирование зависимости между входными функциями и выходными данными целевого предсказания. Данные методы широко применяются в системах предсказания поломок оборудования [5], в которых собираются исторические данные и типы поломок, после которых строятся корреляционные зависимости. Другим примером может служить обучение нейросети обнаружению на фото людей, которые носят каску. Для этого используется несколько сотен размеченных изображений, на которых обведены люди с каской и без, после этого алгоритм позволяет с высокой точностью обнаружить отсутствие каски и выдать оповещение.

Обучение без учителя используется для входных данных

7 Essential Guide to Data Lakes. Available at: https://www.matillion.com/uploads/pdf/essential-guide-to-data-lakes-designing-data-lakes-to-optimize-analytics.pdf

8 ISO/IEC 20922:2016 ISO/IEC 20922:2016(en) – Information technology – Message Queuing Telemetry Transport (MQTT) v3.1.1 Available at: https://www.iso.org/obp/ui/#iso:std:iso-iec:20922:ed-1:v1:en

9 Apache Kafka Guide. Available at: https://docs.cloudera.com/documentation/enterprise/6/6.1/PDF/cloudera-kafka.pdf

10 InfluxDB Documentation. Release 5.3.1, 2022. Available at: https://buildmedia.readthedocs.org/media/pdf/influxdb-python/latest/influxdb-python.pdf

без соответствующей выходной переменной[11]. Эти алгоритмы обнаруживают скрытые закономерности в данных. Кластеризация является одним из основных видов алгоритмов обучения без учителя. Примером данного метода может служить система обнаружения дефектов конвейерной ленты или постороннего предмета на ленте, когда невозможно заранее обучить систему на сотнях подобных дефектов (часто невозможно собрать статистику и разметить изображения).

В рамках контролируемого обучения чаще применяют следующие алгоритмы анализа данных[12]:

- **Линейные алгоритмы**. Самые простые алгоритмы обучения с учителем моделируют линейное соотношение между входными признаками и выходной переменной, значение которой мы хотим предсказать, – линейная регрессия, логистическая регрессия.
- **Алгоритмы на основе соседства точек (neighborhood)**. Алгоритмы этой группы – «ленивые ученики», поскольку в данном случае обучение разметке новых точек основывается на их близости к уже размеченным точкам. В отличие от линейной или логистической регрессии модели, основанные на соседстве точек, не обучаются предсказанию меток для новых точек. Вместо этого они предсказывают метки для новых точек исключительно исходя из того, насколько новые точки удалены от существующих размеченных точек. Ленивое обучение (lazy learning) также относится к методам обучения на примерах (instance-based learning) или непараметрическим методам.
- **Алгоритмы на основе деревьев решений.** Вместо того чтобы использовать линейный метод, можно построить дерево решений, в котором все примеры сегментируются или стратифицируются по отдельным областям на основании имеющихся меток. По завершении сегментирования каждая область соответствует определенному классу меток (для задач классификации) или диапазону предсказанных значений (для задач регрессии). Этот процесс аналогичен тому, как если бы мы поручили ИИ автоматически создать правила, ориентированные на получение наилучших решений или предсказаний. К основным алгоритмам дерева решений относятся одиночное дерево решений, бэггинг, случайные леса, бустинг.
- **Алгоритмы опорных векторов**. Вместо того чтобы строить деревья для разделения данных, можно использовать алгоритмы, предназначенные для создания гиперплоскостей, разделяющих данные на основании имеющихся признаков. Соответствующий подход получил название метод опорных векторов (support vector machine – SVM). SVM не гарантирует идеального разделения (не все точки в определенной области гиперпространства обязаны иметь одну и ту же метку), но расстояние между пограничными точками, с которыми ассоциирована некоторая метка, и пограничными точками, с которыми ассоциирована другая метка, должны быть как можно большими. Кроме того, границы не обязаны быть линейными – можно обеспечить более гибкое разделение данных, используя нелинейные ядра.
- **Нейронные сети**. Обучение представлениям данных можно проводить с использованием нейронных сетей, которые состоят из входного слоя, нескольких скрытых слоев и выходного слоя. Входной слой использует признаки, тогда как выходной слой пытается добиться соответствия переменной отклика (ответной переменной). Скрытые слои представляют собой вложенную иерархию абстрактных понятий – каждый слой (или понятие) пытается понять, как предыдущий слой соотносится с выходным слоем. Опираясь на эту иерархию, нейронная сеть может обучаться сложным понятиям путем их формирования на основе более простых понятий. Нейронные сети – один из наиболее мощных подходов к аппроксимации функций, однако с ним сопряжены такие проблемы, как переобучение и трудность интерпретации.

В рамках обучения без учителя используются следующие алгоритмы анализа данных[13]:

- **Снижение размерности**. Алгоритмы снижения размерности транслируют оригинальное многомерное пространство входных данных в пространство более низкой размерности, фильтруя наименее релевантные признаки и сохраняя как можно большее количество признаков, представляющих интерес. Снижение размерности позволяет эффективнее выявлять шаблоны и решать крупномасштабные, вычислительно трудоемкие задачи (чаще всего связанные с обработкой изображений, видео, речи и текста). Существуют две основные разновидности алгоритмов снижения размерности: линейная проекция и нелинейное снижение размерности, которые включают в себя алгоритмы анализа главных компонент, сингулярное разложение и случайную проекцию.
- **Обучение на многообразиях**. Результаты линейной проекции и нелинейного снижения размерности можно улучшить за счет применения нелинейного преобразования вместо линейного – такой подход известен как обучение на многообразиях (manifold learning) или нелинейное снижение размерности. Среди алгоритмов выделяют Isomap, стохастическое вложение соседей с t-распределением, словарное обучение.
- **Анализ независимых компонент**. Одной из общих проблем, порождаемых неразмеченными данными, становится наличие множества независимых сигналов, которые вложены в имеющиеся в нашем распоряжении признаки. Используя анализ независимых компонент (independent component analysis – ICA), можно разделить эту смесь сигналов на независимые компоненты. После такого разделения можно реконструировать любой из оригинальных признаков, образуя линейные комбинации сгенерированных индивидуальных компонент. Метод ICA широко применяется для обработки сигналов (например, для идентификации отдельных голосов в аудиозаписи из шумной записи).
- **Латентное размещение Дирихле**. Обучение без

11 Основы обучения без учителя. 2020. Режим доступа: http://www.williamspublishing.com/PDF/978-5-907144-99-6/part.pdf

12 Основы обучения без учителя. 2020. Режим доступа: http://www.williamspublishing.com/PDF/978-5-907144-99-6/part.pdf

13 Основы обучения без учителя. 2020. Режим доступа: http://www.williamspublishing.com/PDF/978-5-907144-99-6/part.pdf

учителя также может быть использовано для объяснения набора данных путем изучения факторов, которые обусловливают сходство отдельных частей набора между собой. Это требует обучения ненаблюдаемым элементам набора данных – подход, получивший название латентное размещение Дирихле (latent Dirichlet allocation – LDA).

- **Кластеризация**. После редуцирования набора оригинальных признаков до набора меньшего размера можно приступить к поиску интересующих нас закономерностей путем группирования схожих примеров. Этот процесс, называемый кластеризацией, можно реализовать с помощью целого ряда алгоритмов обучения без учителя и использовать в таких задачах, как сегментирование рынка. Среди алгоритмов выделают Метод k-средних, Иерархическая кластеризация, DBSCAN (density-based spatial clustering of applications with noise – основанная на плотности пространственная кластеризация для приложений с шумами).
- **Извлечение признаков**. Обучение без учителя позволяет обучаться новым представлениям оригинальных признаков данных – это называется извлечением признаков (feature extraction). Данный подход может применяться для эффективного снижения размерности данных путем создания редуцированного подмножества оригинальных признаков. А кроме того, это позволяет генерировать новые признаки с целью повышения производительности в задачах обучения с учителем. Среди алгоритмов выделяют автокодировщики, извлечение признаков путем контролируемого обучения сетей прямого распространения.
- **Глубокое обучение без учителя (unsupervised deep learning)**. Представления постепенно улучшаются за счет обновления весов различных узлов с использованием градиента функции ошибки на каждой итерации тренировки. Подобное обновление весов требует интенсивных вычислений. Среди алгоритмов выделяют предварительное обучение без учителя, ограниченные машины Больцмана, глубокие сети доверия, генеративно-состязательные сети.
- **Обработка последовательных данных с помощью обучения без учителя**. Обучение без учителя позволяет также обрабатывать последовательные данные, например, временные ряды. Один из таких подходов предполагает обучение скрытым состояниям марковской модели. Алгоритмы скрытой марковской модели включают обучение вероятному следующему состоянию при условии, что известна последовательность ранее встречавшихся частично наблюдаемых состояний и полностью наблюдаемых выходов. Эти алгоритмы широко применяются для решения задач, связанных с обработкой речи, текста и временных рядов.
- **Обучение с подкреплением с использованием обучения без учителя (inforcement learning)** – третья из основных методологий машинного обучения, в соответствии с которой агент определяет свое оптимальное поведение (действия) в условиях окружения на основе обратной связи (получаемого вознаграждения). Эта обратная связь называется сигналом подкрепления. Целью агента является максимизация накапливаемого вознаграждения.

Приведенные методы анализа данных также применимы и для анализа Больших Данных, формируемые горнотехнической системой. Выбор того или иного алгоритма зависит от решаемой задачи, так, для широкого спектра производственных и применения технологий компьютерного зрения для детекций аномалий применяются нейросети, а для поиска влияния одних переменных на события, соотносящиеся по времени, часто используют кластеризацию.

Обсуждение и выводы

Предложенный в статье подход к классификации данных с точки зрения скорости их получения: непрерывные данные, данные от датчиков в режиме событий и периодические данные, получаемые с разной частотой, позволит стандартизировать принципы работы с данными. В связи с тем что объем передаваемых данных не зависит от частоты формирования информации от цифрового источника, предлагается передавать все формируемые данные от цифрового источника для последующего поиска неявных зависимостей между ними.

Отмечено, что применение конкретных методов и алгоритмов анализа данных горнотехнической системы прежде всего зависит от поставленной задачи, часто формируемой в виде гипотезы, которая должна быть проверена за счет выявления неявных зависимостей между разными источниками данных. В этой связи для выявления таких зависимостей и поиска оптимизационных моделей для повышения эффективности управления горнотехнической системой на всех этапах освоения месторождений предлагается применять подход ELT, в основу которого входят озера данных, что может стать важным преимуществом с точки зрения управления технологическими процессами горнотехнической системы в будущем.

В зависимости от наличия исходных данных и решаемой задачи могут быть применены различные методы обучения данных – контролируемое обучение и обучение без учителя, в рамках которых существуют различные алгоритмы искусственного интеллекта, применяемые в зависимости от прикладной задачи и имеющихся данных. Комплексирование этих методов и алгоритмов к собранной выборке исторических данных может повысить эффективность управления при отработке месторождений твердых полезных ископаемых.

Список литературы

1. Рыльникова М.В., Цупкина М.В., Кирков А.Е. Технологии сбора и обработки больших данных – основа повышения достоверности первичной информации о массивах горных пород при освоении месторождений полезных ископаемых и техногенных образований. Известия Тульского государственного университета. *Науки о Земле*. 2023;(1):308–327. https://doi.org/10.46689/2218-5194-2023-1-1-308-327

2. Захаров В.Н., Каплунов Д.Р., Клебанов Д.А., Радченко Д.Н. Методические подходы к стандартизации сбора, хранения и анализа данных при управлении горнотехническими системами. *Горный журнал*. 2022;(12):55–61. https://doi.org/10.17580/gzh.2022.12.10
3. Захаров В.Н., Гвишиани А.Д., Вайсберг Л.А., Дзеранов Б.В. Большие данные и устойчивое функционирование горнотехнических систем. *Горный журнал*. 2021;(11):45–52. https://doi.org/10.17580/gzh.2021.11.06
4. Nambiar A., Mundra D. An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*. 2022;6(4):132. https://doi.org/10.3390/bdcc6040132
5. Rahmani A.M., Azhir E., Ali S., Mohammadi M., Ahmed O.H., Yassin Ghafour M., Hasan Ahmed S., Hosseinzadeh M. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. *PeerJ Computer Science*. 2021;7:e488 https://doi.org/10.7717/peerj-cs.488

References

1. Rylnikova M.V., Tsupkina M.V., Kirkov A.E. Technologies of big data collection and processing – the basis for increasing the reliability of primary information about rock massifs in the development of mineral deposits and technogenic formations. Izvestiya Tulskogo gosudarstvennogo universiteta. *Nauki o Zemle*. 2023;(1):308–327. (In Russ.) https://doi.org/10.46689/2218-5194-2023-1-1-308-327
2. Zakharov V.N., Kaplunov D.R., Klebanov D.A., Radchenko D.N. Methodical approaches to standardization of data acquisition, storage and analysis in management of geotechnical systems. *Gornyi Zhurnal*. 2022;(12):55–61. (In Russ.) https://doi.org/10.17580/gzh.2022.12.10
3. Zakharov V.N., Gvishiani A.D., Vaisberg L.A., Dzeranov B.V. Big Data and sustainable functioning of geotechnical systems. *Gornyi Zhurnal*. 2021;(11):45–52. (In Russ.) https://doi.org/10.17580/gzh.2021.11.06
4. Nambiar A., Mundra D. An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*. 2022;6(4):132. https://doi.org/10.3390/bdcc6040132
5. Rahmani A.M., Azhir E., Ali S., Mohammadi M., Ahmed O.H., Yassin Ghafour M., Hasan Ahmed S., Hosseinzadeh M. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. *PeerJ Computer Science*. 2021;7:e488 https://doi.org/10.7717/peerj-cs.488

Информация об авторах

Валерий Николаевич Захаров – академик Российской академии наук, директор, Институт проблем комплексного освоения недр им. академика Н.В. Мельникова Российской академии наук, г. Москва, Российская Федерация; https://orcid.org/0000-0002-9309-2391, Scopus ID 56438797200

Дмитрий Алексеевич Клебанов – кандидат технических наук, заведующий лабораторией №3.2, Институт проблем комплексного освоения недр им. академика Н.В. Мельникова Российской академии наук, г. Москва, Российская Федерация; https://orcid.org/0000-0002-3289-9212, Scopus ID 55922194400, e-mail: Klebanov_d@ipkonran.ru

Михаил Андреевич Макеев – научный сотрудник лаборатории №3.2, Институт проблем комплексного освоения недр им. академика Н.В. Мельникова Российской академии наук, г. Москва, Российская Федерация; https://orcid.org/0000-0003-0941-7606, Scopus ID 57270771800

Дмитрий Николаевич Радченко – кандидат технических наук, доцент, заведующий лабораторией №1.1, Институт проблем комплексного освоения недр им. академика Н.В. Мельникова Российской академии наук, г. Москва, Российская Федерация; https://orcid.org/0000-0003-1821-3840, Scopus ID 6507269210

Информация о статье

Поступила в редакцию: 03.10.2023
Поступила после рецензирования: 22.11.2023
Принята к публикации: 02.12.2023

Information about the authors

Valerii N. Zakharov – Corresponding Member of RAS, Dr. Sci. (Eng.), Professor, Director, Institute of Comprehensive Exploitation of Mineral Resources Russian Academy of Sciences, Moscow, Russian Federation; https://orcid.org/0000-0002-9309-2391, Scopus ID 56438797200

Dmitry A. Klebanov – Cand. Sci. (Eng.), Head of Laboratory of Intelligent Systems and Digital Technologies, Institute of Comprehensive Exploitation of Mineral Resources of Russian Academy of Sciences, Moscow, Russian Federation; https://orcid.org/0000-0002-3289-9212, Scopus ID 55922194400, e-mail: Klebanov_d@ipkonran.ru

Mikhail A. Makeev – Research Associate, Laboratory No.3.2, Institute of Comprehensive Exploitation of Mineral Resources Russian Academy of Sciences, Moscow, Russian Federation https://orcid.org/0000-0003-0941-7606, Scopus ID 57270771800

Dmitry N. Radchenko – Cand. Sci. (Eng.), Associate Professor, Head of the Laboratory of Theoretical Fundamentals for Mining Systems Design, Institute of Comprehensive Exploitation of Mineral Resources Russian Academy of Sciences, Moscow, Russian Federation; https://orcid.org/0000-0003-1821-3840, Scopus ID 6507269210

Article info

Received: 03.10.2023
Revised: 22.11.2023
Accepted: 02.12.2023