



Федеральное государственное бюджетное учреждение
“Всероссийский научно-исследовательский институт
минерального сырья им. Н.М. Федоровского
(ФГБУ “ВИМС”)

ИНФОРМАЦИОННО- АНАЛИТИЧЕСКИЕ МАТЕРИАЛЫ:

**МИРОВЫЕ ДОСТИЖЕНИЯ
РАЗВИТИЯ ТЕХНОЛОГИЙ, МЕТОДОВ И
МЕТОДИК ПРОГНОЗИРОВАНИЯ И
ПОИСКОВ МЕСТОРОЖДЕНИЙ**

**СЕРИЯ:
ПОИСКОВЫЕ ТЕХНОЛОГИИ**

Москва 2025 г.

Обзор 3-4

ПРОГНОЗНО-МИНЕРАГЕНИЧЕСКИЕ ИССЛЕДОВАНИЯ И ПОИСКОВЫЕ РАБОТЫ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ.

РЕДАКТОР-СОСТАВИТЕЛЬ:
В.В. Коротков

СОДЕРЖАНИЕ:

	Стр.
I. МЕТОДОЛОГИЯ ПРОГНОЗИРОВАНИЯ МИНЕРАЛЬНЫХ РЕСУРСОВ НА ОСНОВЕ БОЛЬШИХ ГЕОЛОГИЧЕСКИХ ДАННЫХ.....	5
1. Введение.....	5
2. Большие геологические данные.....	6
3. Теории количественного прогнозирования на основе больших данных.....	9
4. Методы формирования больших данных.....	9
4.1. Разработка методов анализа больших текстовых данных.....	10
4.2. Разработка алгоритмов анализа пространственных данных.....	12
5. Проблемы прогнозирования минерагенического потенциала.....	14
5.1. Большие данные для прогнозирования на основе систем знаний...	14
5.2. Построение концептуальной модели на основе текстового анализа.....	15
5.3. Реализация технологии прогнозирования на основе пространственных данных.....	16
5.4. Использование и визуализация больших данных для геологического прогноза.....	16
6. Теоретические основы и методы прогнозирования минерального потенциала.....	17
6.1. Интеллектуальная обработка данных.....	18
6.2. Интеллектуальная информатизация.....	19
6.3. Разумно осведомленный метод.....	20
6.4. Интеллектуальная сервитизация.....	23
7. Вывод.....	24
II. СРАВНЕНИЕ МОДЕЛЕЙ МАШИННОГО ПРОГНОЗИРОВАНИЯ НА ОСНОВЕ АЛГОРИТМА ГИПЕРПАРАМЕТРИЧЕСКОЙ ОПТИМИЗАЦИИ	26
1. Введение.....	26
2. Модели машинного обучения и гиперпараметрическая оптимизация...	27
2.1. MLP-модель.....	27
2.2. Модель AdaBoost.....	28
2.3. Модель OCSVM.....	29
2.4. Алгоритм Bat.....	30
2.5. Алгоритм Firefly.....	31
2.6. Построение модели перспективности.....	32
3. Область исследования и извлечение информации.....	33
3.1. Геологическая обстановка.....	33
3.2. Извлечение металлогенической информации.....	34

4. Картирование перспективности.....	38
4.1. Выбор слоя карты.....	38
4.2. Машинное обучение в сочетании с гиперпараметрической оптимизацией.....	40
4.2.1. Модель MLP в сочетании с гиперпараметрической оптимизацией.....	42
4.2.2. Модель AdaBoost в сочетании с гиперпараметрической оптимизацией.....	43
4.2.3. Модель OCSVM в сочетании с гиперпараметрической оптимизацией.....	44
4.3. Картирование минерального потенциала.....	45
5. Результаты.....	46
6. Обсуждение.....	50
7. Выводы.....	51
 III. ПРИМЕНЕНИЕ ПРОЦЕССА АНАЛИТИЧЕСКОЙ ИЕРАРХИИ ПРИ ПРОГНОЗИРОВАНИИ МИНЕРАЛЬНОГО ПОТЕНЦИАЛА НА ОСНОВЕ ИНТЕГРИРОВАННОЙ ГЕОЛОГО-АЭРОГЕОФИЗИКО-ГЕОХИМИЧЕСКОЙ МОДЕЛИ.....	53
1. Введение.....	53
2. Процесс аналитической иерархии (АНР).....	53
3. Построение и вес системы оценочных индексов.....	56
3.1. Построение системы оценочных индексов.....	56
3.2. Определение весов оценочных индексов.....	58
3.3. Метод взвешивания для моделей прогнозирования.....	59
4. Применение.....	60
5. Выводы.....	63
 IV. ПРОГНОЗИРОВАНИЕ Fe-Ti-P-RZM ОРУДЕНЕНИЯ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ (<i>рифт Осло, Норвегия</i>).....	65
1. Введение.....	65
2. Геологические предпосылки.....	66
3. Данные.....	68
4. Методология и рабочий процесс.....	69
4.1. Стратегия выборки обучающих данных.....	69
4.2. Разработка объектов.....	70
4.3. Выбор объекта.....	74
4.4. Модель классификации.....	74
5. Результаты и обсуждение.....	76
5.1. Область 1: Kodal.....	77
5.2. Область 2: Сильян.....	78
5.3. Область 3: контакт между В1-базальтом и монцонит/сиенитом.....	79
6. Вывод.....	81
 V. МАШИННОЕ ОБУЧЕНИЕ ДЛЯ ОПРЕДЕЛЕНИЯ ГРАНИЦ ГЕОХИМИЧЕСКИХ АНОМАЛИЙ (<i>район Юаньбонанг, провинция Ганьсу, Китай</i>).....	82
1. Введение.....	82
2. Материалы и методы.....	83
2.1. Случайный лес (RF).....	84

2.2. График частичной зависимости (PDP).....	85
2.3. График накопленного локального эффекта (ALE).....	85
2.4. Анализ SHAP.....	86
3. Результаты и обсуждение.....	87
3.1. Регрессия с помощью RF-моделирования.....	88
3.2. Классификация с помощью радиочастотного моделирования.....	90
4. Выводы.....	94
 VI. ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОРФИРОВОГО ОРУДЕНЕНИЯ НА ОСНОВЕ МИКРОЭЛЕМЕНТОВ ЦИРКОНА (металлогеническая провинция Тетис).....	
1. Введение.....	96
2. Геологические условия.....	97
3. Методы.....	98
3.1. Данные и предварительная обработка.....	98
3.2. Алгоритмы машинного обучения.....	99
3.3. Оценка эффективности модели.....	101
3.4. Визуализация t-SNE.....	101
4. Результаты и обсуждение.....	102
4.1. Обычные геохимические индикаторы.....	102
4.2. Построение моделей машинного обучения.....	104
4.3. Применение модели на порфириковых месторождениях в центре Тетиса.....	107
4.4. Классификация типов порфириковых месторождений	108
4.5. Обсуждение.....	111
5. Выводы.....	112
 VII. МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ РУДНЫХ МЕСТОРОЖДЕНИЙ НА ОСНОВЕ МИКРОЭЛЕМЕНТОВ В МАГНЕТИТЕ (рудный район Осса-Морена, Южная Иберия, Португалия)..	
1. Введение.....	113
2. Классический подход.....	114
3. Машинное обучение (ML).....	116
3. Материалы и методы.....	
4. Тематическое исследование рудной зоны Осса-Морена.....	117
5. Материалы и методы.....	118
5.1. Данные и предварительная обработка.....	118
5.2. Уменьшение размеров.....	119
5.3. Машинное обучение.....	119
6. Результаты.....	121
6.1. Анализ данных.....	121
6.2. Классический подход.....	123
6.3. Уменьшение размеров (PCA).....	124
6.4. Машинное обучение.....	126
6.4.1. Моделирование эталонных наборов данных.....	127
6.4.2. Моделирование набора данных тематического исследования.....	129
7. Обсуждение.....	131
7.1. Уменьшение размеров с помощью PCA.....	131
7.2. Модели ML в тестовом наборе данных.....	132

7.3. ML применительно к набору данных тематического исследования.....	134
8. Выводы.....	135
<i>Источники:</i>	136

1. МЕТОДОЛОГИЯ ПРОГНОЗИРОВАНИЯ МИНЕРАЛЬНЫХ РЕСУРСОВ НА ОСНОВЕ БОЛЬШИХ ГЕОЛОГИЧЕСКИХ ДАННЫХ [4].

1. Введение

Поскольку глобальная информатизация вступает в продвинутую стадию, постепенно наступает эра больших данных. С быстрым развитием новейших технологий, таких как облачные вычисления и искусственный интеллект, технологии и приложения для обработки больших данных стали свидетелями прорывов в различных областях. McKinsey, всемирно известная консалтинговая компания, заявила следующее: “Данные вошли во все отрасли и сферы бизнеса, став важным производственным фактором. Открытие, добыча и применение больших данных будут способствовать новому росту производительности и технологическим инновациям”.

Геологические большие данные обладают следующими характеристиками: направление, пространство-время, разнообразие, полиморфизм, причинность, многоисточник, изменчивость, уникальность, комбинация, гетерогенность, нелинейность, выборка, случайность, размытость, редкость, ассоциация и зональность. Являясь ключевым компонентом геонаучных больших данных, геологические большие данные содержат неструктурированную и полуструктурированную информацию из многочисленных дисциплин, включая стратиграфию, седиментологию, геоморфологию, геотектонику, палеонтологию, минералогию, математическую геологию, геохронологию, картографию, металлогению и др. Будучи междисциплиной между геологической наукой и математикой, математическую геологию можно рассматривать как аспект геологической науки - “наука о данных”.

Областью прикладных исследований математической геологии является количественное прогнозирование и оценка минеральных ресурсов с самой длинной историей, широчайшими аспектами, наибольшим количеством участников и наиболее выдающимися достижениями. Как материальная основа выживания человечества, минеральные ресурсы являются не только основой для развития информационного общества, но и предпосылкой для строительства индустриализации, играя значительную стратегическую роль в содействии развитию национальной экономики. С ростом спроса на минеральные ресурсы становится все меньше открытых на поверхности и неглубоких руд, что усугубляет проблемы поисков. Страны по всему миру приступили к разведке «скрытых» рудных месторождений, успешность которых зависит от углубленного изучения металлогенических закономерностей и научных исследований по теории и методам прогнозирования и оценки минеральных ресурсов. Внедрение больших данных в области наук о земле является неизбежной тенденцией развития, чтобы перейти от “математической геологии” к “цифровой геологии”, а также для проведения

научной количественной оценки и анализа прогнозов поисковых работ. В связи с резким ростом объема и сложности больших геологических данных текущая новая задача в интеллектуальных прогнозировании и поисках заключается в том, как расширить применение интеллектуальных средств цифровых данных, текстовых, 2D и 3D пространственных данных, основанных на геологических знаниях и опыте, для обеспечения эффективности и точности прогнозирования и поисков. До сих пор лишь немногие предложили теорию и метод интеллектуального прогнозирования и поисков, основанные на больших геологических данных, и технологический процесс интеллектуального прогнозирования, основанный на больших данных, еще не создан.

С точки зрения текстового и пространственного интеллектуального анализа больших данных целью данного исследования является анализ проблем и решений интеллектуального прогнозирования поисковых работ на основе больших данных, чтобы обеспечить теоретическую и техническую поддержку для прогнозирования и поисков на основе больших данных.

2. Большие геологические данные.

Поскольку информационные технологии постоянно меняются в эпоху больших данных, применение больших геологических данных и искусственного интеллекта в последние годы стало предметом исследований, что означает, что идеи, технологии и методы больших данных в информационную эпоху постепенно успешно применяются в геологической области. В настоящее время геологические архивы, базы данных геологических исследований и базы данных геологических документов различных стран предоставляют многоуровневые и всесторонние геологические данные, включая текстовые и пространственные данные, а также модели интеллектуального анализа для неструктурированных, полу-структурированных и структурированных данных.

Геологические большие данные представляют собой тип пространственно-временных больших данных, которые в основном генерируются в ходе исследований и соответствующих процессов геологических научных исследований в области фундаментальной геологии, геологии полезных ископаемых, гидрогеологии, инженерной геологии, геологии окружающей среды; данные также поступают в результате эксплуатации и использования энергии и минеральных ресурсов, процессов мониторинга, предотвращения и контроля окружающей среды и стихийных бедствий, а также различных космических и наземных наблюдений с помощью дистанционного зондирования земли. Подходы к получению данных включают геофизику, геохимию, бурение, дистанционное зондирование, а также различные приложения,

такие как составление карт, аналитические вычисления, аналоговое моделирование, оценка прогноза и интеллектуальное управление, которые существуют во множестве форм, таких как тексты, графика, акустические изображения и образцы. Следовательно, эффективное получение, сортировка и извлечение эффективной геологической информации из массивных и сложных больших данных для лучшего использования является одним из основных направлений исследований многих ученых. С учетом того, что эта тема исследования находится в центре внимания, были достигнуты некоторые достижения в построении платформы и методах обработки данных для геологических больших данных.

Исследования в области больших геологических данных сильно зависят от построения платформ больших данных. Организация и внедрение взаимосвязей геологических и пространственно-временных больших данных между различными темами в значительной степени зависят от решений и общей архитектуры платформ больших данных, безопасного хранения больших данных, механизмов индексирования, планирования, методов и технологий обработки больших данных, мониторинга функций управления, а также суперкомпьютерных каналов данных. Страны по всему миру активно содействуют созданию и развитию платформ геологических больших данных. Например, в 2007 году Европейский союз (ЕС) опубликовал “Акт об инфраструктуре пространственной информации в Европейском сообществе”, позволяющий обмениваться экологическими пространственными данными по всему ЕС. Впоследствии Национальный научный фонд Соединенных Штатов официально запустил проект “Earth Cube” в 2012 году, который успешно объединил большой объем геологических данных, а также соответствующую информацию и знания. Проект “Earth Cube”, характеризующийся открытостью, прозрачностью и всесторонностью, существенно улучшает поток знаний и коммуникации. Питер Бауманн и др. из Бременского университета в Германии в 2015 году успешно разработали EarthServer для механизма анализа данных о Земле, который включает в себя такие функции, как доступ к данным, анализ данных, извлечение данных и извлечение информации, и создали открытый, кроссплатформенный и простой в эксплуатации механизм доступа к геологическим данным. В настоящее время многие страны создали свои собственные платформы обмена геологическими большими данными (такие как Geo Connections в Канаде и Geospatial One-Stop в Соединенных Штатах) и реализовали эффективный поиск, доступ и получение геопро пространственной информации, оказывая таким образом помощь секторам национального, провинциального и других уровней в принятии решений. В 2012 году Лаборатория геологических информационных технологий Министерства земельных ресурсов успешно создала GeoBDA, техническую экспериментальную платформу для изучения технологий геологических больших данных. В 2013 году распределенная

инфраструктура Hadoop была постепенно модернизирована до стандартной архитектуры хранения геологических больших данных. Одновременно разработчики использовали программную модель MapReduce для значительного повышения вычислительной эффективности массивных геологических наборов данных за счет параллельных вычислений. В 2014 году первая в Китае независимо разработанная платформа для сбора больших геологических данных начала пробную эксплуатацию в Академии геологической службы Тибета. Впоследствии, в 2016 году, Геологическая служба Китая выпустила интеллектуальную платформу пространственных данных. В 2018 году Китайская академия наук запустила стратегический научно-исследовательский проект, известный как CASEarth, целью которого является поддержка крупных научных открытий и макрорешений. В 2019 году был официально запущен проект Deep-time Digital Earth (DDE), который был интегрирован с идеями, представленными 254 учеными из 24 стран. Его пять научных целей включают реконструкцию системы знаний о Земле на основе больших данных, построение крупнейшей системы разнородных данных из нескольких источников, создание универсальной платформы обработки и анализа данных о Земле, решение основных научных проблем на основе больших данных и формирование платформы глубокой коммуникационной интеграции digital earth, чтобы создать первый “Геологический Google”.

В сочетании с требованиями практического применения технология анализа геологических больших данных широко применяется в таких областях, как базовая геологическая съемка, разведка и оценка ресурсов нефти и газа, интеллектуальные города, управление земельными ресурсами, исследование океана, дистанционное зондирование, а также прогнозирование. Поиски, разведка и оценка полезных ископаемых. В частности, интеллектуальный алгоритм анализа больших данных добился прорывов в различных областях, таких как извлечение геохимических аномалий, извлечение гравитационных и магнитных аномалий, каротажная литологическая идентификация, пространственное предсказание сейсмических и осадочных фаций, прогнозирование времени землетрясений, а также разведка и оценка минеральных ресурсов, оказав глубокое влияние на область геологии.

Минеральные ресурсы являются одной из важнейших материальных баз, поддерживающих развитие национальных экономик. Применение интеллектуального алгоритма анализа для оценки минеральных ресурсов является неизбежной тенденцией в развитии математической геологии; это также первый шаг к реализации интеллектуальных поисков. Совместное применение новой теории и новых технологий имеет большое значение для развития прогнозно-минерагенических, поисковых и разведочных работ.

3. Теории количественного прогнозирования на основе больших данных.

В эпоху больших данных прогнозирование и оценка геологических минеральных ресурсов постепенно вступили в новую стадию, переходя от цифровой разведки к интеллектуальной разведке. С точки зрения обмена большими данными было создано множество облачных платформ для больших данных, таких как облачные вычисления, облако данных и автоматическое чтение текста. Соответствующие методы анализа больших данных также оказывают помощь в количественной оценке минеральных ресурсов. Объединив метод прогностического мышления в эпоху больших данных с прогнозированием и оценкой минеральных ресурсов и основываясь на теории причинности, теории тенденций и теории взаимосвязи изменения информации были исследованы четыре теории: теория моделирования прогнозирования полезных ископаемых, теория корреляции прогнозирования полезных ископаемых, метод анализа тенденций и теория дифференциации. Основанный на моделях метод количественного прогнозирования и оценки минеральных ресурсов использует геологическую теорию для руководства анализом геологических данных и компьютерные технологии для интеллектуального анализа геологических больших данных, а также для количественного прогнозирования и оценки минеральных ресурсов. После систематического обобщения применения больших геологических данных при изучении металлогенической регулярности, металлогенических рядов и металлогенических систем пришли к выводу, что большие геологические данные могут предоставить новые идеи для металлогенической теории и прогнозирования.

Основная идея больших данных - это не прогнозирование, а скорее, процесс перехода от известного к неизвестному - метод прогностического мышления, который включает в себя извлечение потенциальных знаний с помощью методов интеллектуального анализа. Чтобы применить прогностическое мышление, основанное на больших данных, к прогнозированию минеральных ресурсов, необходимо не только внедрить множество алгоритмов анализа данных, но и создать совершенную и эффективную теоретическую систему методов прогнозирования больших данных, чтобы продвинуть прогнозирование минеральных ресурсов к интеллектуализации.

4. Методы формирования больших данных.

Интеллектуальный анализ данных - сложный итеративный процесс. При использовании огромных объемов больших данных различные алгоритмы искусственного интеллекта, представленные машинным обучением, предоставляют широкое пространство для применения интеллектуального анализа больших данных. В настоящее время искусственный интеллект,

машинное обучение и глубокое обучение - это три «горячих» слова, используемых в текущих исследованиях, которые иногда используются взаимозаменяемо. Однако глубокое обучение появилось не внезапно. Существует тесная связь между глубоким обучением, искусственным интеллектом, большими данными и машинным обучением (рис. 1).

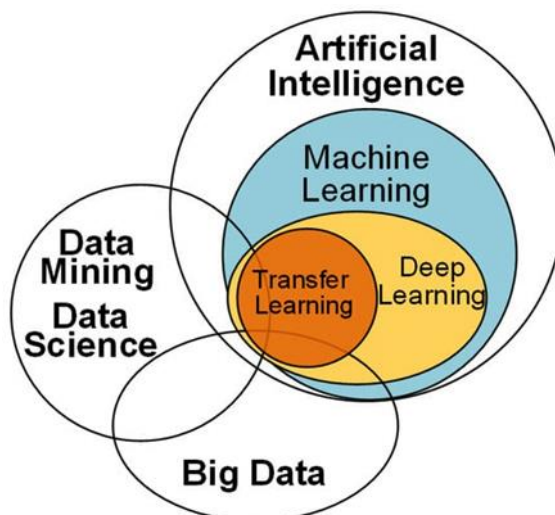


Рис. 1. Взаимосвязь между глубоким обучением и другими смежными актуальными темами.

Несомненно, текущий интеллектуальный анализ больших данных при прогнозировании и поисках далек от уровня искусственного интеллекта. Применение методов интеллектуального анализа больших данных в науках о Земле остается в зачаточном состоянии. Обычные методы анализа не могут полностью выявить глубокие корреляции, скрытые в различных геологических данных. Благодаря развитию технологии интеллектуального анализа данных можно анализировать разнородные и пространственные геологические данные. С помощью пространственного, корреляционного, классификационного, регрессионного и кластерного анализов геологических данных должны выполняться: прогнозирование, поисковые, разведочные и добычные работы. Существует множество средств интеллектуального анализа больших геологических данных. Геологические большие данные содержат неструктурированную, полуструктурированную и структурированную информацию, включая текстовые, пространственные, мультимедийные и другие разнообразные и массивные данные. Учитывая спрос на количественное прогнозирование минеральных ресурсов, в этом материале обобщен прогресс исследований с двух точек зрения: интеллектуальные и пространственные методы анализа больших текстовых данных.

4.1. Разработка методов анализа больших текстовых данных.

Для геологических исследований в эпоху больших данных необходимо всесторонне собирать различные данные геологических изысканий, включая полуструктурированные данные, а также неструктурированную литературу, технические и геологические отчеты,

научные статьи и т.п. Обработка этих накопленных данных и извлечение из них геологических знаний имеет большое значение. В условиях постоянно растущего объема текстовых данных вопрос о том, как разумно классифицировать и организовать такие геологические текстовые данные, а также понимать, анализировать и визуализировать их семантические ассоциации, стал основной проблемой при преобразовании больших геологических данных из данных в знания.

Интеллектуальный анализ текста был впервые предложен Фельдманом в 1995 году. В таблице 1 показаны ключевые технические процессы интеллектуального анализа текста, включая сбор набора данных, предварительную обработку текста, интеллектуальный анализ данных и визуализацию.

Табл. 1.

Ключевые технические процессы и содержание текстового майнинга.

Technical Process	Technology/Content
Dataset	Text databases, web pages, documents, e-books, etc.
Text preprocessing	Noise removal, word segmentation, parts-of-speech tagging, stop-word removal, feature representation, feature extraction, etc.
Data mining	Named entity recognition, word frequency analysis, sentiment analysis, automatic summarization, semantic network, similarity analysis, classification, clustering, association rules, intelligent retrieval, regression, trend analysis, etc.
Visualization	Knowledge and conclusions, graphical interface, command line, word cloud, knowledge graph, etc.

Обобщенный метод сегментации word не может соответствовать реальным требованиям. Тексты по геологии характеризуются различиями в описании, разнообразием компиляций и геологической экспертизой. Текстовое представление обычных векторов признаков не может интуитивно выражать профессиональную геологическую семантическую информацию, не может соответствовать разнообразным грамматическим характеристикам геологических текстов и не может углубляться в поиск информации в геологических текстах. Автоматизированные технологии, такие как сбор, предварительная обработка, и визуальный анализ массивных геологических текстовых данных, сложны в реализации.

Было предложено глубокое обучение, обеспечившее новое направление для интеллектуального анализа текстовых данных. С развитием глубокого обучения методы нейросетей достигли выдающихся результатов в семантическом анализе обработки естественного языка (NLP), поиске запросов, моделировании классификации предложений и других традиционных методах. В качестве одного из наиболее надежных алгоритмов классификации с глубоким обучением сверточные нейронные сети (CNN) успешно применяются в NLP. Sun завершила многозначную классификацию настроений Weibo с помощью контролируемого обучения многозначной классификации, основанного на модели CNN. Впоследствии успешно применили алгоритм M-CNN для обнаружения эмоций с

несколькими метками. Исследователи предприняли многочисленные попытки извлечь структурированную информацию из неструктурированных сетей для создания баз знаний и преобразования текстовой информации в семантические сети. Представление Google в виде графа знаний предоставляет новую идею для визуального представления NLP. Граф знаний как важнейшая семантическая сеть, представляет собой форму организации данных или продукт, который выражает сущности, концепции и их семантические отношения с помощью ориентированных графов. Они широко используются в области геологии. Успешно применен граф знаний для представления ключевой информации в геологическом неструктурированном тексте. Были использованы методы сетевого анализа для представления и извлечения информации, скрытой в данных о типе месторождения, местоположении и наблюдениях. Был предложен метод построения концептуальной модели поисков, основанный на металлогенической теории, с использованием основанных на текстовых технологиях обнаружения больших данных.

Для количественного прогнозирования минеральных ресурсов текстовые данные содержат множество ключевой информации, которая не только помогает извлекать поисковую информацию из разнообразных, массивных, гетерогенных геологических текстовых больших данных, но также предоставляет новые идеи и направления для построения поисковых моделей.

4.2. Разработка алгоритмов анализа пространственных данных.

Сталкиваясь со все более сложными задачами поисков, количественное прогнозирование постепенно перешло от легко идентифицируемых неглубоких рудных залежей к «скрытому» оруденению. Традиционный метод прогнозирования 2D не может удовлетворить требованиям поисков скрытых объектов. Трехмерное прогнозирование подразумевает увеличение размеров данных и повышение их актуальности. Другими словами, машинное обучение все больше интегрируется с прогнозированием. Из-за сложных геологических условий поисков геологические данные часто проявляют нелинейные характеристики. Машинное обучение может лучше представлять сложную, нелинейную взаимосвязь между минерализацией и элементами доказательств. В настоящее время алгоритмы машинного обучения в основном включают искусственные нейронные сети, алгоритмы опорных векторов, случайного леса, экстремального обучения и моделей максимальной энтропии.

Однако информация в 3D имеет многоисточниковые и мультимодальные характеристики, которые создают проблемы для ее классификации при прогнозировании. Поэтому обычным алгоритмам машинного обучения не хватает необходимой производительности. Уместно и важно использовать глубокое обучение при прогнозировании, что также является значимым исследовательским экспериментом для применения интеллектуальных алгоритмов Big data в

геологических исследованиях. Глубокое обучение - это особый метод машинного обучения, который включает CNN, рекуррентные нейронные сети (RNN), многоуровневое автоматическое кодирование, а также многослойные рекуррентные нейронные сети с обратной связью и полные CNN. В настоящее время алгоритмы глубокого обучения достигли хороших результатов при извлечении 2D аномалий, таких как геохимические, гравитационные, магнитные и комплексное прогнозирование, в частности, за счет использования алгоритма CNN.

Однако 2D интеллектуальный алгоритм может извлекать только поверхностные особенности; он игнорирует 3D пространственные особенности при прогнозировании и сталкиваются со следующими проблемами: (1) акцент делается на количественном извлечении аномалий; важность пространственно-распределенных объектов и корреляция между различными условиями поиска игнорируются; (2) коэффициент использования скважин зависит от того, была ли найдена руда - метод взвешивания доказательств работает только на положительных примерах, как можно использовать отрицательные пробы?; (3) подход предполагал сосредоточение внимания на важных факторах, контролирующих руду, при игнорировании других, казалось бы, несущественных факторов, даже в обычных алгоритмах машинного обучения.

Следовательно, необходимо создать интеллектуальную систему поисков на основе как геологического фона, так и больших данных, чтобы повысить эффективность прогнозирования. В дополнение к извлечению взаимосвязи между пространственными позициями на изображении, как это делается с помощью 2D CNN, 3D сверточные нейронные сети (3D-CNN) могут дополнительно извлекать соответствующую информацию во временной области в виде непрерывных изображений (широко используемых в медицинской диагностике, распознавании действий на языке жестов, распознавании выражения, распознавании сигналов мозга, а также во многих других областях). Для прогнозирования поисковых работ пространственная информация в данных из областей геологии, геохимии и геофизических исследований, полученных различными методами мониторинга, содержит много аномалий рудообразования. Поэтому применение 3D-CNN для распознавания особенностей глубинного пространства и извлечения аномалий имеет большое значение.

Как извлечь эффективную информацию из такой множественной, массивной, разнородной информации для анализа геологоразведочных условий? Как извлечь особенности временных рядов и пространственную корреляцию на основе трехмерной геологической модели? Как повысить интеллектуальность процесса прогнозирования? Как реализовать сочетание геологической экспертизы и интеллектуальных алгоритмов? Таковы проблемы

интеллектуального анализа данных пространственных 3D поисков. Для решения проблемы асимметричной информации данных при прогнозировании также стоит рассмотреть перспективы применения трансфертного обучения.

5. Проблемы прогнозирования минерагенического потенциала.

Текущий интеллектуальный анализ данных при прогнозировании остается на стадии цифровой “искусственной разведки”. Существует определенный пробел в реализации интеллектуальной разведки. С наступлением эпохи больших данных не существует теории прогнозирования на основе больших данных, основанной на знаниях экспертов-геологов, с “геологической линией”, “технической линией” и “линией данных” в качестве основной линии, и продвижение интеллектуального прогнозирования. Требуется ряд технологий интеллектуального анализа для поддержки обнаружения больших данных, хранения обновлений больших данных, анализа характеристик больших данных и комплексного количественного прогнозирования, тем самым способствуя потоку и оптимизации трех основных элементов “данных”, “модели” и “знаний”. Что касается перехода от поиска данных и информатизации к прогнозированию, то основные проблемы заключаются в следующем.

5.1. Большие данные для прогнозирования на основе систем знаний.

Для интеллектуального анализа данных при прогнозировании экспертные базы знаний являются основой для исследований. Для полного раскрытия базовых данных и для выполнения классификации, хранения, статистики существует настоятельная необходимость реализовать интеграцию новых технологий и новых методов. Эффективные методы обнаружения больших данных могут предоставить не только базовые данные для интеллектуального анализа в исследуемых районах, но и решения для построения больших баз данных о полезных ископаемых. Для новых технологий получения данных необходимо преодолеть четыре основные трудности: регионализм, всесторонность, профессионализм и многоязычие. Что касается регионализма - как получить эффективные данные из 233 стран и регионов по всему миру? Что касается полноты - обычный метод запроса к базе данных с сопоставлением строк, такой как использование ключевых слов и тематических слов, не может точно и всесторонне запрашивать массивные данные, связанные с полезными ископаемыми. Как получить соответствующие данные более всесторонне и систематически в соответствии со спросом? Что касается профессионализма - как точно получить список целевых полевых данных с различных веб-сайтов академических конференций, учреждений по сбору геологических данных, научно-исследовательских институтов, поисковых систем, крупных издательств, журналов по геонауке и тд., а затем обновлять их в режиме реального времени?

Что касается многоязычия - как преодолеть языковые барьеры для максимальной интеграции между “большой наукой” и “большими данными”, а также для обмена трансграничными данными о полезных ископаемых? Нынешнему применению технологии обнаружения больших данных в области прогнозирования не хватает реальных решений и исследовательских моделей. Не существует теоретической системы или технического метода для обнаружения больших данных, которые действительно интегрировали бы экспертные знания; это в определенной степени ограничивает его применение в междисциплинарных областях.

5.2. Построение концептуальной модели на основе текстового анализа.

Построение концептуальной модели является ключевым звеном в количественном прогнозировании минеральных ресурсов. Концептуальная поисковая модель относится к текстовому представлению модели месторождения, основанной на различных поисковых методах с определяющими руду характеристиками, признаками и комбинациями данных (исключая генезис и допущения), включая модели месторождений, региональные поисковые модели и локальные поисковые модели, которые могут использоваться для прогнозирования. Концептуальная модель предназначена для конкретной исследуемой области. Прогнозирование требует всестороннего учета различных факторов, таких как мнения экспертов, геологический фон и тип поисковых работ, с помощью определенного набора правил. Среди них наиболее влиятельным фактором прогнозирования является опыт и знания. Исследовательские достижения экспертов по минеральным ресурсам в основном соответствуют большим объемам неструктурированных текстовых больших данных, таких как технические отчеты и т.п. Текстовые большие данные содержат важные знания и опыт. Перед лицом различных методов анализа NLP необходимо решить следующие проблемы: как применить такую технологию для разумной классификации, хранения и извлечения полезной информации из массивных текстовых данных, как осуществить углубленный анализ, понимание и визуализацию семантических корреляций в геологическом тексте и как построить концептуальную модель для прогнозирования и поисков.

Методы, которые могут извлекать важную информацию, такую как время поисков, предыстория, генетический тип минерализации, включают не только простую обработку текста, но и знания экспертов по минеральным признакам и эффективным средствам НЛП. Как интегрировать прогностическое мышление, основанное на больших данных, и методы поисков полезных ископаемых, управляемые системой знаний, в количественное прогнозирование? Как использовать предыдущие знания для оптимизации интеллектуальной модели поисков? Как построить региональную концептуальную поисковую модель исследуемой территории и реализовать автоматическое расширение базы моделей месторождений? Решение этих

вопросов - единственный способ продвинуть интеллектуальное прогнозирование поисковых работ. Однако существующая система методов требует дальнейшего совершенствования.

5.3. Реализация технологии прогнозирования на основе пространственных данных.

С наступлением эпохи больших данных оценка минеральных ресурсов постепенно переходит от цифровых поисков к интеллектуальным. Несмотря на бесконечное количество теоретических методов интеллектуального анализа больших данных, большинство из них не реализовали технических инноваций. В настоящее время интеллектуальные поиски, основанные на больших данных, находятся в зачаточном состоянии. Как интегрировать идею интеллектуальных поисков на основе больших данных в реальный спрос на интеллектуальные поисковые методы? Как построить теорию прогнозирования и техническую базу, способствующую процессу интеллектуальных поисков? Как расширить значение геологических больших цифровых данных на неструктурированные, полу-структурированные и структурированные данные, такие как текстовые данные по, данные 2D-изображений и даже данные 3D-модели, особенно в интеллектуальном анализе пространственных данных? Это основные технические трудности при использовании больших данных.

Кроме того, прогнозирование требует всестороннего учета различных факторов, таких как мнение эксперта, геологический фон и метод поисковых работ, с помощью определенного набора правил. Как объединить знания экспертов-геологов с искусственным интеллектом, чтобы реализовать органичное сочетание моделей, основанных на данных, и знаний, систематический путь эволюции, объединенный с предшествующими научными знаниями, и скачок к искусственному интеллекту? Как отслеживать автоматический сбор и обновление больших геологических данных и улучшать процесс принятия системных решений с помощью механизма самооптимизации модели? Это ключевые аспекты, которые необходимо реализовать при преобразовании моделей прогнозирования поисковых работ в рамках четвертой научной парадигмы.

5.4. Использование и визуализация больших данных для геологического прогноза.

Как правило, совместное использование и визуализация больших данных являются конечным звеном для реализации вспомогательного процесса принятия решений с помощью интеллектуального анализа. Эффективное использование больших геологических данных может решить научные проблемы, такие как прогнозирование, поиски, разведка и оценка ресурсов. Суть больших данных заключается в потоке данных. Поэтому обмен данными играет жизненно важную роль. Текстовые, пространственные, технологические и результирующие данные, полученные в процессе прогнозирования, часто имеют сложную структуру и большие объемы. В дополнение к разнообразию, массе и неоднородности больших данных, они

обладают уникальными характеристиками, такими как пространство–время, изменчивость, причинность и корреляция. В настоящее время развитая технология интеграции многомерных данных платформы виртуальной реальности (VRP) может реализовывать только автономный обмен. Отсутствует платформа, которая может максимально упростить обмен и повторную датификацию больших данных, отличных от конфиденциальных; это одна из основных нерешенных проблем.

Данные, генерируемые при интеллектуальном прогнозировании на основе больших геологических данных, имеют не только простой уровень, но также включают информационный уровень, уровень знаний и даже уровень принятия решений, такой как график экспертных решений, дерево сетевой структуры и интеллектуальная модель прогнозирования. В настоящее время не существует такой системы, которая позволяла бы эффективно хранить все промежуточные данные, включая базовые данные, модели алгоритмов, экспертные знания и распространенное программное обеспечение, и управлять ими. Поэтому важно и необходимо создать интеллектуальную платформу прогнозирования, которая объединяет цифровое выражение, количественную корреляцию, виртуальное моделирование и количественный прогноз с базой данных (DB), базой моделей (MB) и базой знаний (KB) в качестве основы и с пространственно-временным облачным хранилищем и пространственно-временной базой данных, базой знаний и графом знаний, а также с высокопроизводительными вычислениями, виртуальной визуализацией и разработкой приложений в качестве поддержки.

6. Теоретические основы и методы прогнозирования минерального потенциала.

Большие данные, характеризующиеся их огромным масштабом, неоднородностью, мультимодальностью, сложными корреляциями и динамичным возникновением, требуют эффективных вычислительных моделей и методов. Более того, исследования в области наук о Земле в целом остаются “многопарадигмальной” смесью, где основная часть остается в первой парадигме “логических рассуждений”, в то время как некоторые вошли во вторую парадигму “теоретического моделирования”. Несколько дисциплин исследовали третью парадигму “вычислительного моделирования”, которая постепенно продвигается к четвертой парадигме “анализа данных”; для этого требуется объединить специфические характеристики науки о Земле и предложить новую теорию для анализа больших геологических данных в прогнозных и поисковых целях. Теория прототипа интеллектуального прогнозирования и методы геологических больших данных, предложенные в этом исследовании (рис. 2), пытаются преобразовать “логические рассуждения” в первой парадигме в правила ассоциации графов знаний и преобразовать “теоретическое моделирование” во второй парадигме в правила

рассуждений в модели алгоритма, которые успешно сочетают геологические знания и “искусственный интеллект” в “четвертой парадигме”.

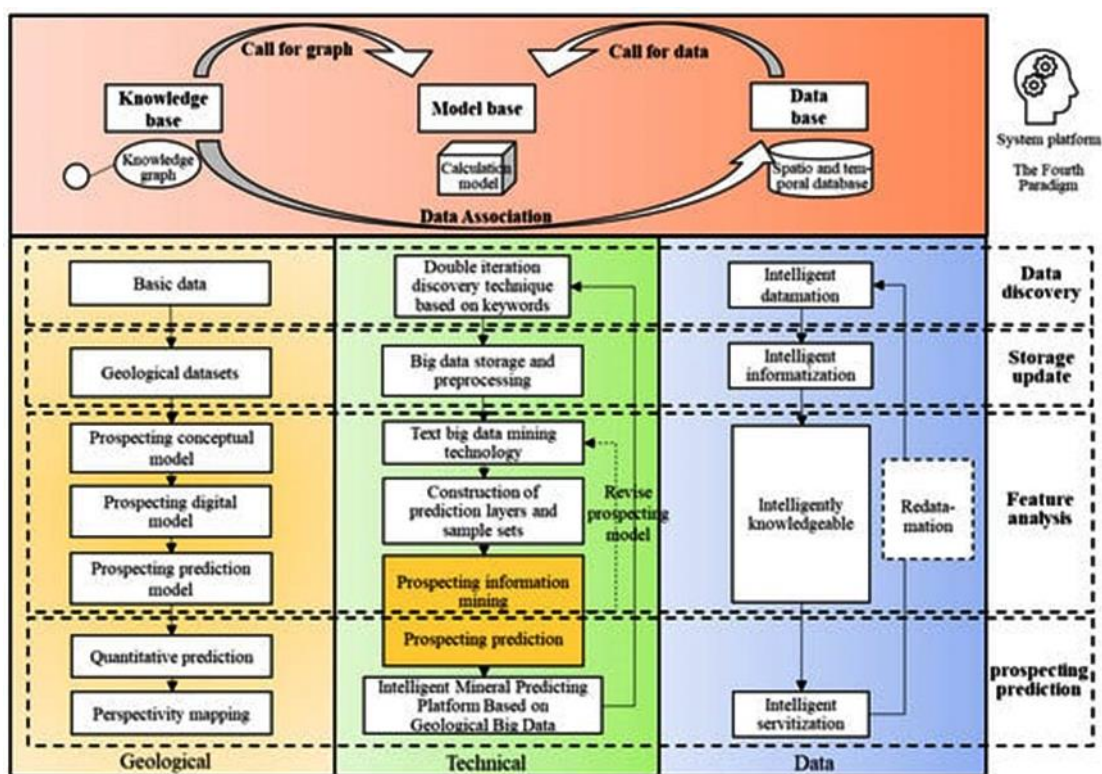


Рис. 2. Методы интеллектуальных поисков на основе больших геологических данных.

В частности, теория и метод интеллектуальной поисков на основе геологических больших данных подразумевает рассмотрение “геологических”, “технических” и “данных” как трех основных направлений. В отличие от обычного метода количественного прогнозирования, он использует гибридную модель прогнозирования поисковых работ на основе больших данных. Гибридная схема состоит из схемы, основанной на знаниях, основанной на геологической модели, которая, в свою очередь, основана на технической модели, и модели, основанной на данных. Суть теории заключается в содействии потоку, вызову, циркуляции и оптимизации трех ключевых факторов: “знаний”, “модели” и “данных”, и в предварительном виде она представляет собой прототип механизма интеллектуальной увязки, который переводит этап, основанный на данных и информации, в интеллектуальный этап. С целью количественного прогнозирования на основе больших геологических данных анализ данных можно разделить на следующие четыре части.

6.1. Интеллектуальная обработка данных.

Теоретический метод интеллектуального обнаружения и сбора больших геологических данных обеспечивает теоретическую основу для процесса их интеллектуальной обработки. Полнота, профессионализм и эффективность соответствующих данных, необходимых для

прогнозирования и поисков, напрямую определяют точность результатов интеллектуального анализа в эпоху больших данных, что также является основой для проведения интеллектуальных прогнозов. Предыдущие прогнозы часто основывались на существующих геологических данных; однако наступление эпохи больших данных значительно расширило наши источники. Следовательно, необходимо создать совершенную интеллектуальную систему обнаружения и обновления данных, основанную на уникальных характеристиках больших геологических данных, чтобы эффективно просматривать и извлекать полуструктурированные, структурированные и неструктурированные данные. Благодаря интеллектуальному обнаружению, интеллектуальному отбору, интеллектуальному сопоставлению и интеллектуальному обновлению данных они могут быть автоматически классифицированы и сохранены, предоставляя подробные текстовые и пространственные данные для процессов информатизации и накопления знаний.

6.2. Интеллектуальная информатизация.

Теоретические методы интеграции, хранения и автоматического обновления многомерных массивных больших данных не только обеспечивают теоретическую и методологическую поддержку информатизации данных, но также обеспечивают мощную базовую поддержку для интеллектуальной сервитизации и интеллектуальной повторной обработки. Нормальная работа интеллектуальной системы неотделима от интеллектуального потока геологических данных. В настоящее время интеграция геологических данных обычно вращается вокруг интеграции пространственного местоположения, игнорируя значимость и ценность самих данных. Таким образом, основываясь на теоретической и методологической основе интеграции пространственных данных (рис. 3), включая интеграцию источников данных, геометрическую интеграцию и интеграцию абстрактных диапазонов, данное исследование полностью использует ограничения, связанные с различными геологическими значениями и знаниями, для создания многомерной системы интеграции данных и управления на основе единого базового географического пространства. Соответственно, пространственные данные (такие как геология, структура, минерализованные точки, геофизика, геохимия и дистанционное зондирование), текстовые данные (такие как отчеты о поисках, разведке и т.п.), модели алгоритмов (такие как модели алгоритмов интеллектуального майнинга) и экспертные знания (такие как основы концептуальных моделей разведки, графики знаний и логические деревья) могут быть унифицированы и интегрированы в системную платформу. Таким образом, обмен данными может быть достигнут с помощью платформы интеллектуального прогнозирования, что в конечном итоге удовлетворит существующие потребности.

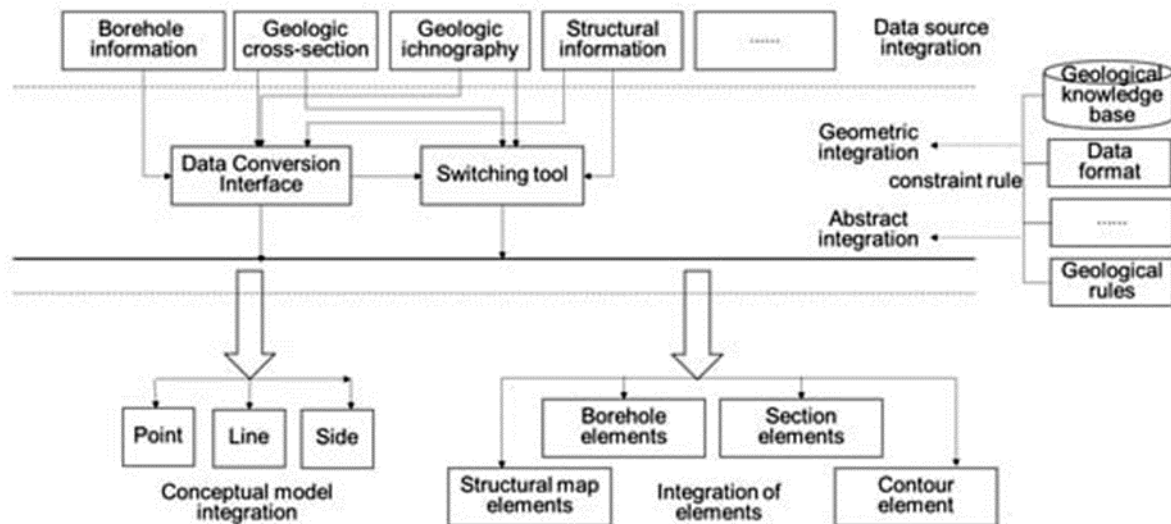


Рис. 3. Теоретическая основа метода интеграции данных из нескольких источников.

6.3. Разумно осведомленный метод.

Метод анализа металлогенических признаков в рамках больших данных обеспечивает теоретическую основу для использования больших данных в процессе принятия интеллектуальных решений, целью которых является построение концептуальной модели прогнозирования минеральных ресурсов. Анализ и извлечение металлогенических признаков в рамках больших геологических данных включают в себя не только извлечение ключевой металлогенической информации из текстов, но также анализ и извлечение металлогенических признаков из многомерного пространственного распределения данных. На основе извлечения текстовых признаков из больших данных строится концептуальная модель, которая служит руководством для построения модели прогнозирования на основе анализа признаков пространственных данных. Анализ металлогенических объектов включает в себя множество структурированной, полу-структурированной и неструктурированной информации; методы анализа включают метод NLP, метод с использованием искусственного интеллекта, метод геостатистики, метод нелинейного анализа и метод анализа многомерного пространства. Модели прогнозирования поисковых работ можно разделить на модели, основанные на знаниях, данных и гибридных моделях. Основной задачей металлогенического прогнозирования является разработка технологии прогнозирования, основанной на металлогенических закономерностях. В этом отношении методы интеллектуального анализа больших данных, такие как глубокое обучение, представляют собой модели, которые могут извлекать полезные свойства из массивных данных и формировать модель прогнозирования от известного к неизвестному. Однако у них есть недостатки в том, что они фокусируются только на корреляциях, а не причинно-следственных связях, и подчеркивают только результаты, а не процессы. Следовательно, лучше интегрировать геологические знания в металлогеническую

закономерность или металлогенические модели с помощью интеллектуального метода металлогенического прогнозирования на фоне больших данных.

Интеллектуальное накопление знаний относится к процессу формирования неизвестных и потенциальных знаний посредством интеллектуального глубокого анализа текстовых и пространственных данных на основе гибридных моделей. Этот прогресс не только объединяет графики знаний и логическую структуру знаний экспертов-геологов с концептуальными моделями поисков сотен рудных месторождений, но также обеспечивает поток, вызов, циркуляцию и оптимизацию носителей “знаний”. Носители “знаний” здесь относятся к новым узлам знаний, новым логическим структурам и новым концептуальным моделям разведки в процессе поисков. Методы интеллектуального анализа данных относятся к процессу поиска и извлечения неизвестной в настоящее время, но ценной потенциальной информации из больших данных путем объединения современной технологии больших данных и технологии искусственного интеллекта с традиционными методами анализа данных. Интеллектуальный анализ данных включает в себя множество алгоритмов, таких как наивный байесовский анализ, нейронная сеть, K окрестностей, дерево решений и другие алгоритмы машинного обучения, а также анализ ассоциаций, кластерный анализ, классификационное дерево регрессии и машинный алгоритм опорных векторов, основанный на теории статистического обучения. Интеллектуальный анализ данных тесно связан с различными информационными технологиями, такими как статистика, искусственный интеллект, экспертные системы и системы баз данных. Знания в области интеллектуального анализа данных обычно включают закономерности, концепции, шаблоны и правила.

На основе извлеченных ключевых поисковых данных текстовой информации и ключевой пространственной векторной информации в этом исследовании была проведена углубленная разработка для получения следующих двух аспектов знаний: (1) для уточнения металлогенического фона, металлогенического периода, генетического типа и типа минерализации, чтобы установить металлогеническую закономерность и металлогеническую модель, чтобы определить концептуальную поисковую модель исследуемого района; (2) для обучения правилам рассуждений в рамках поисково-прогнозной модели для самостоятельного изучения; оптимизация модели - искусственный интеллект, объединенный с предшествующими научными знаниями.

Инновация заключается в интеграции “накопленной базы”, записанной в виде текстов и графиков, в анализ геологических больших данных. Металлогеническая система - это процесс контроля всех геологических элементов и минерализации при формировании, изменении и сохранении рудных месторождений в определенной геологической пространственно-

временной области, а также целостности внутри серии рудных месторождений и серии аномалий минерализации, которая представляет собой естественную систему с металлогенической функцией. Определенный “металлогенический фон, металлогенический возраст, генетический тип и типы минерализации” определяют определенные “металлогенические условия, рудоконтролирующие факторы, поисковые признаки и аномалии минерализации”, которые в дальнейшем определяют определенную “поисковую модель”. Таким образом, в сочетании с серией достижений в области поисковых моделей в системе big data discovery of the earth, это исследование извлекло знания о металлогенической регулярности и металлогенических моделях, основанных на многочисленных методах интеллектуального анализа данных, таких как технология NLP, технология классификации данных и технология машинного (глубокого) обучения, и построило поисковую модель исследуемой территории, которая является еще одним ключевым фактором в реализации теорий интеллектуальных поисков.

С другой стороны, в целях интеллектуальных поисков в данном материале предлагается извлекать основные прогнозные данные исследуемой территории на основе базы векторных слоев и данных 3D геологической модели, чтобы определить характеристики геологических пространственных данных и взаимосвязь между геологическими телами. Для извлечения признаков металлогенических аномалий из пространственных векторных данных в этом исследовании были объединены традиционные методы извлечения количественных признаков, включая различные методы извлечения полезной металлогенической информации, такой как особенности структурного развития, особенности структурных зон, структурные рудоподводящие признаки и особенности массива горных пород. Между тем, на основе суперкомпьютерных мощностей и метода искусственного интеллекта были изучены автоматические металлогенические прогностические характеристики больших геологических данных, подтверждающие корреляцию между множеством геологических переменных и оруденением. Сравнивая способность различных методов прогнозирования при случайной комбинации различных рудоуправляющих факторов, в этом исследовании была проведена интеллектуальная оптимизация и оценка метода металлогенического прогнозирования, чтобы определить оптимальную модель прогнозирования и комбинацию факторов, подходящую для исследуемого района, тем самым повышая эффективность поисков.

Металлогенические особенности, которые было трудно выделить из-за асимметричной информации, были обработаны с помощью передачи геологических знаний посредством transfer learning, которая сочетает в себе знания экспертов-геологов и искусственный интеллект. Он реализует путь эволюции системы, основанный на предшествующих научных знаниях,

характеризующийся стохастическими и нелинейными взаимосвязями, и преобразует глубокое обучение, основанное на настраиваемых параметрах, в оперативное обучение, основанное на системе, формируя базу систематических алгоритмов искусственного интеллекта. Следовательно, реализуется весь процесс извлечения информации, интеграции знаний, обоснования знаний и интеллектуального прогностического анализа.

6.4. Интеллектуальная сервитизация.

Теория и метод прогнозирования на основе больших данных не только обеспечивает теоретическую и методологическую поддержку для звена сервитизации интеллектуальных поисков, но также является заключительным этапом интеллектуального прогнозирования. Поэтому в исследованиях приоритет должен отдаваться всестороннему анализу результатов извлечения металлогенических признаков и, в конечном счете, служить эффективности поисков. Это поможет достичь двух целей: (1) решить задачу интеллектуальных поисков и повысить их точность и эффективность; (2) обеспечить удобное выражение и передачу знаний. В ходе этого исследования была создана интеллектуальная платформа прогнозирования, основанная на больших геологических данных, с системой принятия решений в качестве ядра и базами данных DB, MB и KB. Кроме того, эта платформа, поддерживаемая пятью движками пространственно-временного облачного хранилища и базами пространственно-временных данных, базами знаний и графиками знаний, высокопроизводительными вычислениями, виртуальной визуализацией и разработкой приложений, а также с помощью прогнозирования поисков и обмена геологическими данными в качестве практического приложения, объединяет цифровое выражение, количественную корреляцию, виртуальное моделирование и количественное прогнозирование.

Для повышения эффективности и точности прогнозирования минеральных ресурсов предлагается не только интеллектуальный метод прогнозирования, основанный на предварительных знаниях экспертов, но и ряд методов оценки модели для обеспечения стабильности и эффективности модели прогнозирования. В процессе алгоритм автоматически усваивает логические правила и устанавливает механизм случайной комбинации факторов, контролирующих руду, для постепенного приближения к оптимальной комбинации и достижения целевых показателей по поискам. Учитывая, что количество случайных комбинаций может достигать сотен тысяч или даже миллионов, предлагается определить метод оптимального параметра на основе теории фракталов, который сокращает время обучения компьютера и ускоряет скорость сходимости модели, позволяя модели прогнозирования быстро определять оптимальное решение и достигать самооптимизации. Этот метод полезен в интеллектуальном анализе пространственных данных, а также в анализе принятия решений.

С другой стороны, для удобного выражения и передачи знаний была проведена комплексная обработка металлогенических знаний, полученных в процессе прогнозирования, и они представлены в визуальной форме, благодаря чему результаты могут быть легко восприняты. В основе лежит система принятия решений, а в качестве практического применения - прогнозирование поисков и обмен геологическими данными. Предлагается и создается интеллектуальная платформа прогнозирования на основе больших геологических данных, которая объединяет цифровое выражение, количественную корреляцию, виртуальное моделирование и количественный прогноз. Металлогенические знания, полученные с помощью текстового анализа, выражаются в виде графика знаний, который полезен для прогнозирования минеральных ресурсов и его легче понять. Для знаний, полученных в процессе интеллектуального анализа пространственных данных, визуализируется не только ядро свертки обучающей модели, но также цифровая модель и результаты прогнозирования, полученные с использованием метода Web3D. Это полезно для лучшего понимания пространственных характеристик 3D металлогенического пространства и пространственной взаимосвязи с 2D пространством, чтобы обеспечить разумную основу для эффективных поисков.

В дополнение к вышеупомянутым четырем частям, теория прототипа интеллектуальных поисков и метод геологических больших данных также обеспечивают теоретическую и методологическую основу для процесса повторного анализа “знаний”, “модели” и “данных”. Большие данные не только содержат модели прогнозирования, но и обладают динамическими характеристиками; это отражается в процессе повторной оценки, а также является ядром теории интеллектуальных поисков. Чтобы избежать образования “островов данных”, представлена модель потока данных интеллектуальной повторной обработки, в качестве основы которой используются КБ, DB и MB. Механизм автоматического сопоставления и обновления базовой базы данных может быть достигнут путем постоянного включения оперативных данных, полученных “новых знаний”, “новой модели” и “новых данных” через архитектуру облачного хранилища.

7. Вывод.

При быстром развитии науки «ручной» анализ данных не может своевременно «переварить» быстрый рост объема и сложности данных. Характеризующиеся разнородностью многих источников, различными датами, множеством характеристик и сложной семантикой, большие геологические данные, как правило, имеют сложную взаимосвязь с минерализацией. «Размерная катастрофа», возникающая в многомерной «космической» статистике, серьезно

влияет на статистическое моделирование и результаты наблюдений из нескольких источников; это может снизить надежность результатов прогнозирования. Более того, потенциальными факторами, контролирующими рудообразование, вероятно, будут пренебрегать. Поэтому в сочетании с прогностическим моделированием больших данных и «рассуждениями искусственного интеллекта» для прогнозирования минеральных ресурсов в эпоху больших данных необходимо создать эффективный, унифицированный, интеллектуальный процесс прогнозирования. В этом материале сделан обзор теории и метода интеллектуальных поисков больших геологических данных с целью создания системы прогнозирования на основе больших данных в сочетании с основанными на знаниях, моделях и данных шаблонами, чтобы сделать процесс прогнозирования интеллектуальным в эпоху больших данных.

II. СРАВНЕНИЕ МОДЕЛЕЙ МАШИННОГО ПРОГНОЗИРОВАНИЯ НА ОСНОВЕ АЛГОРИТМА ГИПЕРПАРАМЕТРИЧЕСКОЙ ОПТИМИЗАЦИИ [2].

1. Введение.

Модели, основанные на данных, обычно используются при составлении прогнозных карт, а конкретные математические модели используются для количественного описания статистики потенциальных доказательств или пространственного распределения для прогнозирования целевых зон поисков. С быстрым развитием теории и технологии машинного обучения набор инструментов, основанный на моделях, управляемых данными, все больше обогащается. Технология машинного обучения может адаптивно моделировать взаимосвязь между входными и выходными данными и эффективна при решении задач нелинейного прогнозирования. В соответствии с ожидаемыми результатами алгоритма машинное обучение можно разделить на «обучение под наблюдением», «обучение без учителя» и «обучение с подкреплением». Контролируемое обучение использует известные обучающие выборки для корректировки гиперпараметров классификации или регрессии для достижения желаемой производительности, включая искусственную нейронную сеть (ANN), дерево решений, адаптивную нейронную нечеткую систему, ограниченную машину Больцмана, машину опорных векторов (SVM), случайный лес (RF), машина экстремального обучения (ELM) и модели MaxEnt.

При использовании модели машинного обучения для прогнозирования поисков важно эффективно определять внутренние ключевые гиперпараметры моделей, такие как вес сети ANN, коэффициент штрафа за отказоустойчивость и гиперпараметр ядра SVM, а также количество растущих деревьев RF. При практическом применении модели машинного обучения очень сложно эффективно определить оптимальное значение гиперпараметров модели. В некоторых предыдущих исследованиях использовались подходы поиска по сетке или случайной оптимизации для нахождения оптимального значения в предопределенных значениях гиперпараметра, но часто не удавалось найти глобальный оптимальный гиперпараметр, который вряд ли содержался в предопределенном значении гиперпараметра. В последние годы было выдвинуто и постепенно используется множество алгоритмов оптимизации внутренних гиперпараметров модели машинного обучения, например, генетический алгоритм (GA) и оптимизация скопления частиц (PSO). Однако эти классические алгоритмы оптимизации чувствительны к начальной настройке гиперпараметров, что упрощает поиск локальных оптимальных решений и приводит к замедлению скорости сходимости алгоритма. Кроме того, алгоритм bat (BA), алгоритм firefly (FA), алгоритм ant colony (ACA) и

другие появляющиеся алгоритмы swarm intelligence также широко использовались для сбалансирования возможностей алгоритма глобального и локального поиска и повышения производительности гиперпараметрической оптимизации модели.

В этом исследовании для создания моделей прогнозирования ПИ применены многослойный персептрон (MLP), машина одноклассных опорных векторов (OCSVM) и AdaBoost. В процессе моделирования ключевые гиперпараметры различных моделей машинного обучения были оптимизированы ВА и FA. Между тем, значение площади под кривой (AUC) кривой рабочей характеристики приемника (ROC) было выбрано в качестве значения пригодности оптимальной целевой функции, оптимизируя количество скрытых нейронов MLP и коэффициент штрафа за отказоустойчивость и гиперпараметр ядра OCSVM, а также обновляя время обучения и коэффициент снижения веса слабой регрессии AdaBoost. Значение AUC кривой ROC и значение прогноз/площадь (P/A) кривой прогноз-площадь (P-A) были применены для оценки производительности модели путем анализа результатов прогнозирования методов машинного обучения до и после гиперпараметрической оптимизации. Основным вкладом является предложение о том, что производительность моделей машинного обучения в сочетании с ВА или FA может быть улучшена при составлении прогнозных карт.

2. Модели машинного обучения и гиперпараметрическая оптимизация.

2.1. MLP-модель.

MLP - это сетевая структура, обычно используемая в искусственных нейронных сетях (ANN), которые представляют собой многослойную композицию персептронов, состоящую из входного слоя, скрытого слоя и выходного слоя. Помимо входного слоя, каждый узел других слоев представляет собой нейрон с нелинейной функцией активации. Процесс обучения сети MLP состоит из построения вектора признаков и передачи его скрытому слою. Затем результат вычисляется с помощью веса и функции активации и передается следующему слою. В этом процессе непрерывное обучение и корректировка происходят только во время обучающего сигнала обратной связи. В процессе классификации ее необходимо передавать до достижения выходного уровня. Выходные шаги MLP следующие:

Взвешенная сумма входных данных была рассчитана с помощью

$$s_j = \sum_{i=1}^n (W_{ij}X_i) - \theta_j, j = 1, 2, \dots, h \quad (1)$$

где n - количество входных узлов, W_{ij} является ли вес связи от i -го узла входного слоя к j -му узлу скрытого слоя, θ_j является ли смещение j -го скрытого узла, и X_i является i -м входом.

Выходная информация каждого скрытого узла была рассчитана с помощью

$$S_j = \text{sigmoid}(s_j) = \frac{1}{(1 + \exp(-s_j))}, j = 1, 2, \dots, h \quad (2)$$

Выходная функция скрытого узла и конечная функция активации передачи выходной информации были рассчитаны соответственно по уравнениям (3) и (4):

$$o_k = \sum_{j=1}^h W_{jk} \cdot S_j - \theta_k, k = 1, 2, \dots, m$$

$$O_k = \text{sigmoid}(o_k) = \frac{1}{(1 + \exp(-o_k))}, k = 1, 2, \dots, m \quad 3, 4$$

где W_{jk} является ли вес подключения от j -го скрытого узла к k -му выходному узлу и θ_k является смещением k -го выходного узла.

Поскольку вес и смещение влияют на конечное выходное значение, необходимо найти наилучшие вес и смещение связи, чтобы обучить MLP максимально приближать фактический результат к ожидаемому.

2.2. Модель AdaBoost.

В AdaBoost веса одних и тех же выборок постоянно обновляются, а слабые регрессии с разными весами собираются вместе, чтобы сформировать окончательную сильную регрессию. Вес каждой слабой регрессии рассчитывается в соответствии с точностью прогнозирования каждой выборки в каждом обучающем наборе и общей точностью прогнозирования последнего обучающего набора. Кроме того, обновляется вес распределения каждой выборки, а результаты регрессии, полученные в результате каждого обучения, взвешиваются и суммируются как конечный результат сильной регрессии. Конкретные процессы алгоритма AdaBoost заключаются в следующем:

Слабый регрессионный алгоритм и обучающий набор были определены с помощью $\{(x_1, y_1), \dots, (x_N, y_N)\}$, и вектор веса обучающих данных был инициализирован

$$w^1 = [w_1^{(1)}, w_2^{(1)}, \dots, w_N^{(1)}]^T, w_i^{(1)} = 1/N, i = 1, 2, \dots, N \quad (5)$$

Слабая регрессия $G_m(x)$ был получен обучающий набор весовых векторов $w(m)$ распределение и ошибка обучения каждой выборки x_i в слабой регрессии были рассчитаны с помощью

$$e_{i,m} = |G_m(x_i) - y_i| \quad (6)$$

Частота ошибок прогнозирования E_m слабой регрессии $G_m(x)$ на обучающем наборе данных был рассчитан

$$E_m = p(e_{i,m} > \theta) = \sum_{i=1}^N w_{mi} I(e_{i,m} > \theta) \quad (7)$$

где E_m - сумма весов обучающих выборок, когда $e_{i,m} > \theta$, и θ было заранее заданным порогом.

Весовой коэффициент α_m из $G_m(x)$ было рассчитано с помощью

$$\alpha_m = \frac{1}{2e^{E_m}} \quad (8)$$

что продемонстрировало важность $G_m(x)$ в окончательном прогнозе.

Наконец, следующая формула была использована для обновления весового распределения каждой выборки в обучающем наборе для следующей итерации.

$$w_{m+1,i} = \begin{cases} \frac{w_{mi}}{Z_m} \times 1, e_i(t) > \theta \\ \frac{w_{mi}}{Z_m}, e_i(t) \leq \theta \end{cases} \quad (9)$$

За счет итерации вес обучающих выборок с большей ошибкой в слабой регрессии $G_m(x)$ увеличился, в то время как вес обучающих выборок с меньшей ошибкой остается неизменным. После преобразования веса модель AdaBoost сосредоточилась на обучающих выборках с низкой точностью прогнозирования.

Линейная комбинация была построена с использованием коэффициента нормализации Z_m , рассчитанный

$$Z_m = \sum_{i=1}^N w_{mi} \quad (10)$$

Окончательная сильная регрессия была получена с помощью

$$G(x) = \sum_{m=1}^M \alpha_m G_m(x) \quad (11)$$

2.3. Модель OCSVM.

Модель OCSVM, частный случай модели SVM, может обрабатывать неопознанные данные. Подмножество входного пространства оценивается как вспомогательный набор распределения вероятности высокой размерности входных данных. Выборки, не принадлежащие к вспомогательному набору и извлеченные из распределения вероятности высокой размерности, идентифицируются как многомерные аномальные выборки. OCSVM используется для определения минимальной границы области, Γ , принятие решения, $f(x)$, удовлетворяют граничным условиям. Γ содержит не менее $(1 - \nu) m$ нормальных выборок, где m - количество обучающих выборок, а ν - коэффициент пропорциональности для контроля аномальных выборок в обучающих выборках.

Алгоритм сначала нелинейно преобразует обучающую выборку из входного пространства в регенеративное гильбертово пространство, $\Phi: Rd \rightarrow H$, где размерность H бесконечна, а функция $f(x)$ легко определяется. В этом пространстве, $f(x)$ может отображать функцию внутреннего произведения и вычислять внутреннее произведение бесконечномерного пространства отображения с помощью функций ядра во входном пространстве, $K: Rd \times Rd \rightarrow R$:

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_H = \exp(-\|x - y\|^2 / 2\sigma^2) \quad (12)$$

SVM-модель находится в рабочем состоянии $\langle w, \Phi(xi) \rangle_H \geq b - \xi_i$ и $\xi_i \leq 0$, $i=1, 2, \dots, m$, максимизируя следующую формулу:

$$\left[\frac{1}{2} w_H^2 \right] + \left[\frac{1}{vm} \sum_{i=1}^m \xi_i - b \right] \quad (13)$$

где ξ_i является слабой переменной, которая может предотвратить переоснащение модели.

Это уравнение является условной оптимизацией, и решение заключается в следующем:

$$f(x) = \sum_{i=1}^m \alpha_i K(x_i, x) - b \quad (14)$$

тем временем,

$$\sum_{i=1}^m \alpha_i K(x_i, x_j), j \in [1, 2, \dots, m] \quad (15)$$

где α_i это множитель Лагранжа.

2.4. Алгоритм Bat.

ВА - это эвристический алгоритм поиска, который оптимизирует процесс поиска, имитируя использование гидролокатора для обнаружения добычи и обхода препятствий. Значение пригодности используется для выбора местоположения «летучих мышей», а итеративный процесс поиска, при котором заменяется менее приемлемое решение, моделируется на основе выживания наиболее приспособленных. После инициализации каждого гиперпараметра эвристический поиск начинается со случайной позиции, z_i в пространстве поиска с d размерностью. «Летучие мыши» ищут добычу с фиксированной частотой, на разных длинах волн и звуках. Во время поиска добычи летучие мыши автоматически регулируют длину волны в зависимости от расстояния до добычи. На каждой итерации, $(0 < t < T)$. Глобальный поиск проводится для обновления скорости полета и пространственного положения каждой летучей мыши. Пространственные позиции каждого bat используются для вычисления значения пригодности целевой функции, и в качестве текущего

оптимального положения выбирается позиция, соответствующая максимальному значению пригодности. Формула обновления скорости и положения в пространстве выглядит следующим образом:

$$f_i = f_{\min} + (f_{\max} - f_{\min}) \cdot \beta, v_i^{t+1} = v_i^t + (z_i^t - z^*) \cdot f_i, z_i^{t+1} = z_i^t + v_i^{t+1} \quad (16)$$

где v_i^t, v_i^{t+1} представьте скорость полета летучей мыши i при t и $t + 1$ соответственно; z_i^t, z_i^{t+1} укажите местоположение bat i при t и $t + 1$ соответственно; z^* представляет глобальное оптимальное положение; f_i является ли частота импульсов bat i ; β случайное число между $[0,1]$; $(f_{\min}, f_{\max})$ диапазон частоты импульсов.

После каждой итерации интенсивность и частота звука обновляются в соответствии с коэффициентом ослабления громкости импульсов и увеличенным коэффициентом частоты импульсов.

2.5. Алгоритм Firefly.

ФА - это новый алгоритм интеллектуальной оптимизации бионического роя, предложенный Кришнанандом и Гхосом в 2005 году. N «светлячков» с разными начальными значениями яркости распределены случайным образом. Яркость связана со значением функции текущего положения. Чем лучше положение, тем выше яркость. Каждый светлячок ищет другие, которые ярче в пределах прямой видимости (известные как область динамического принятия решений RDI), формируя коллекцию соседей. Светлячок с наибольшей относительной яркостью выбирается методом вероятности рулетки. Яркость, положение и область динамического принятия решений обновляются и повторяются для поиска следующего подходящего светлячка. Итерационный процесс алгоритма разделен на обновление яркости, обновление местоположения и динамическое обновление области принятия решений.

Обновление яркости зависит от значения пригодности, $f(x_i(t))$ соответствующей целевой функции в разных положениях:

$$l_i(t) = (1 - \rho)l_i(t-1) + \gamma f(x_i(t)) \quad (17)$$

где ρ является ли изменчивость флуоресцеина и γ коэффициент замещения флуоресцеина.

В пределах динамической области принятия решений, $\gamma id(t)$ firefly i выбирает других firefly с более высокой яркостью для формирования набора доменов, $N_i(t)$ и переходит к firefly j в области с вероятностью $p_{ij}(t) = l_j(t) - l_i(t) / \sum_{k \in N_i(t)} l_k(t) - l_i(t)$.

Обновления позиции следующие:

$$x_i(t+1) = x_i(t) + \alpha \left(\frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|} \right) \quad (18)$$

где, α это мобильный шаг и $\|x_j(t) - x_i(t)\|$ это евклидово расстояние между светлячками.

Область динамического принятия решений эквивалентна полю зрения светлячков, которое уменьшится, если будет слишком много компаньонов. Формула равна:

$$r_d^i(t+1) = \min \{r_{st}, \max \{0, r_d^i(t) + \beta(n_t - |N_i(t)|)\}\} \quad (19)$$

где β является ли обновленная скорость динамической области принятия решений, $|N_i(t)|$ количество соседних светлячков и rs это радиус восприятия.

2.6. Построение модели перспективности.

Региональная стратиграфия, геологические образования, магматическая активность, а также геохимические данные и данные дистанционного зондирования были извлечены после изучения основных геологических данных исследуемой территории и металлогенического состояния известных месторождений полезных ископаемых, создания пространственной базы данных металлогенической информации из нескольких источников. Были применены методы MLP, AdaBoost и OCSVM в сочетании с BA и FA для построения моделей прогнозирования минеральных ресурсов. Кривые ROC и P-A использовались для оценки эффекта металлогенического прогнозирования различных комбинированных моделей и анализа их металлогенического потенциала прогнозирования. Конкретный процесс показан на рисунке 1.

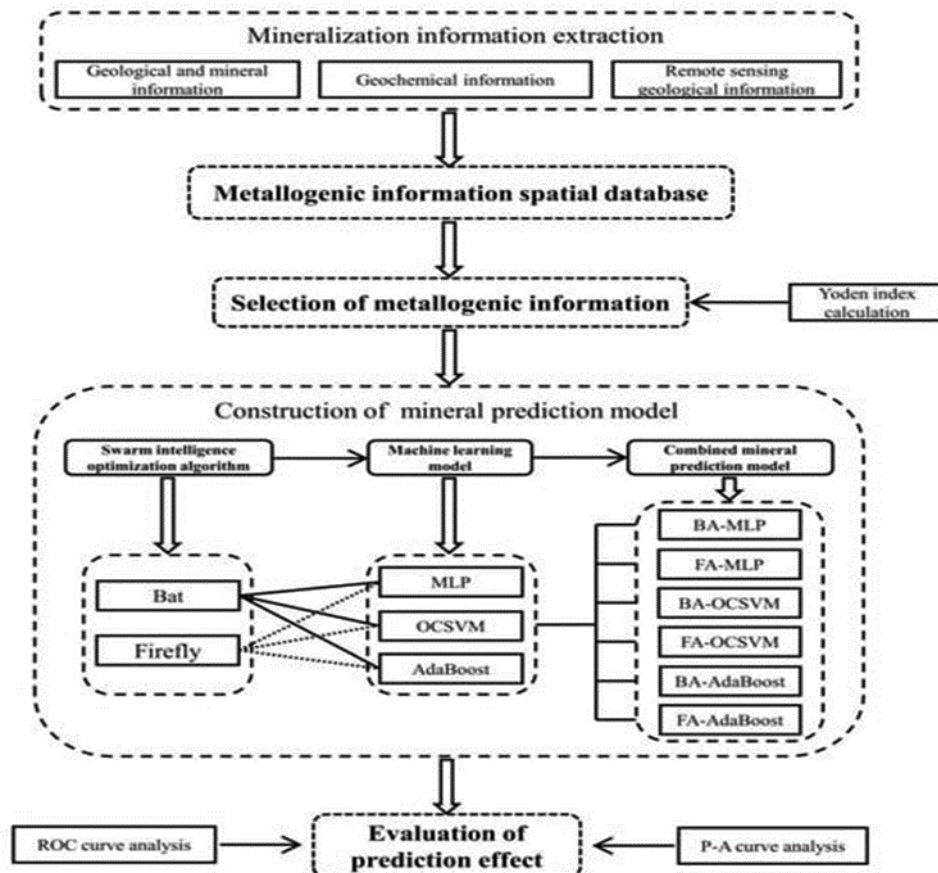


Рис. 1. Процесс расчета минерального потенциала.

3. Область исследования и извлечение информации.

3.1. Геологическая обстановка.

Район исследования был расположен на западе провинции Цинхай в Восточно-Куньлунском орогенном поясе, который является частью полиметаллического металлогенического пояса в горах Восточный Куньлунь. Он характеризуется наличием вулканогенно-осадочной формации позднего палеозоя-мезозоя на севере, морской вулканогенно-осадочной серия раннепалеозойской группы Кимантаге в середине и древней метаморфической серией палеопротерозойской группы Цзиншуйкоу на юге (рис. 2).

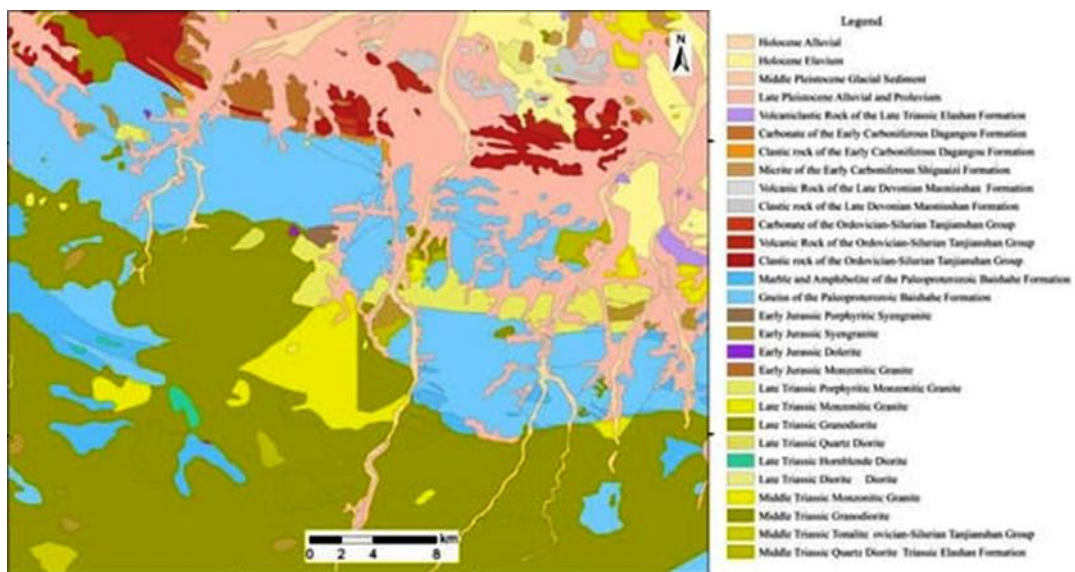


Рис. 2. Упрощенная геологическая карта исследуемого района.

Региональные геологические образования, интрузии и полиметаллическая минерализация контролируются глубинными разломами северо-западного направления, которые образуют зону Кунбэй. В этой зоне в течение силура и поздней перми наблюдалась интенсивная тектоно-магматическая активность рифтогенных и коллизионных процессов. Магматическая активность ранней перми, позднего триаса и раннего мела в основном представлена интрузивными породами средней кислотности. Региональное оруденение формируется вблизи зон контактов между различными интрузивными породами разных периодов. К настоящему времени открыто более 10 месторождений полезных ископаемых, в том числе крупное месторождение сульфидов меди и Ni, скарново-порфировое Mo-полиметаллическое месторождение и скарново-Fe-Cu.

Факторами, формирующими полиметаллическое оруденение считаются: (1) глубинные разломы северо-западного и восточного направлений; (2) метасоматические зоны скарнов, роговиков и др.; (3) магматические интрузивные комплексы.

3.2. Извлечение металлогенической информации.

Данные металлогенической информации были разделены на три категории, а именно: металлогенический геологический фон, геологические данные дистанционного зондирования и информация об аномалиях геохимических элементов. С помощью программной платформы ГИС территория исследования была разделена на 18 600 ячеек сетки, каждая размером $0,3039 \times 0,3041$ км², с учетом масштаба геологических фоновых данных, плотности отбора геохимических проб масштаба 1:50 000 и площади ячеек аномалий дистанционного зондирования, при этом гарантируя, что номера сетки в горизонтальном и вертикальном направлениях являются целыми числами. Вся информация была преобразована в единую проекционную систему координат базы больших металлогенических данных.

Интрузии средней кислотности, связанные с минерализацией по возрасту предварительно разделили на средний триас, поздний триас и ранний мел; по типам на: монцогранит, гранодиорит, тоналит, кварцевый диорит, сиеногранит и мойит.

Геохимические данные.

Для обработки использовался метод средних итераций. Анализ среднего арифметического, дисперсии и коэффициента вариации восьми геохимических элементов, включая Ag, Cr, Cu, Mo, Ni, Pb, W и Zn, показал, что как исходные, так и логарифмические значения каждого элемента, вероятно, не подчинялись нормальному распределению, поскольку каждый элемент имел экстремальное значение, а стандартное отклонение и коэффициент вариации также были большими. Таким образом, итерационный метод был применен для обработки экстремальных выбросов данных элементов следующим образом: среднее значение X и соответствующего стандартного отклонения, σ были рассчитаны значения каждого элемента; экстремальные значения были устранены с помощью $X \pm 3\sigma$ в качестве пределов были пересчитаны среднее и стандартное отклонение данных. Расчет производился, когда не было экстремального значения, и все данные были распределены между $[X - 3\sigma, X + 3\sigma]$. Из таблицы 1 видно, что после итеративного вычисления исключения распределение исходных данных было значительно улучшено.

Табл. 1.

Собственные значения каждого элемента после итерации.

Element	Ag	Cr	Cu	Mo	Ni	Pb	W	Zn
Mean	52.07	34.97	8.76	0.52	3.09	28.42	1.11	45.53
Standard deviation	1.28	1.77	1.85	1.02	2.03	1.38	1.18	1.54
Variation coefficient	0.02	0.05	0.21	1.96	0.66	0.05	1.06	0.03
Anomaly threshold	54.63	38.51	12.46	2.56	7.15	31.18	3.47	48.61

Note: $\omega(\text{Ag, Au})/10^{-9}$ and other elements: 10^{-6} .

Гистограммы логарифмической частоты каждого элемента после итерации показаны на рисунке 3; видно, что, хотя стандартное отклонение и коэффициент вариации отдельных

элементов все еще очень велики, почти все элементы подчиняются нормальному распределению.

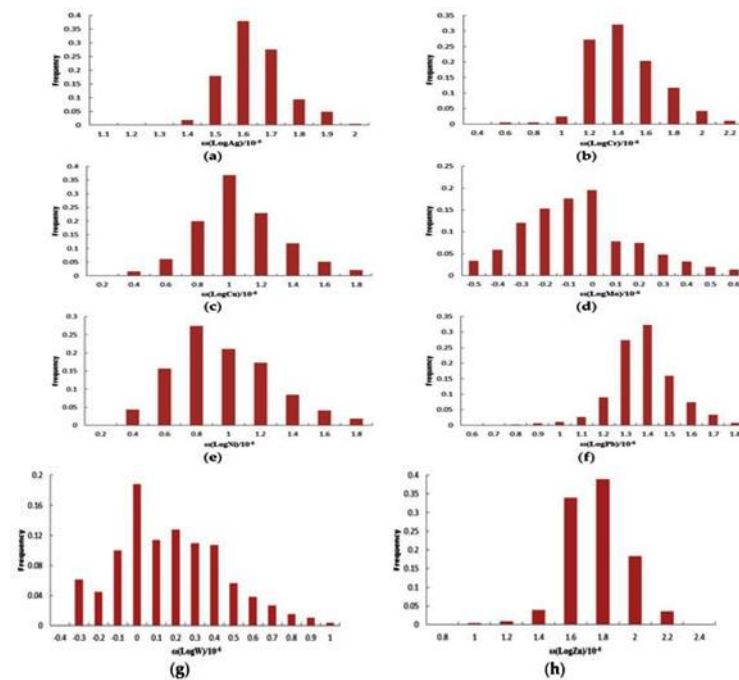


Рис. 3. Частотные гистограммы геохимических элементов.
(a) Ag; (b) Cr; (c) Cu; (d) Mo; (e) Ni; (f) Pb; (g) W; (h) Zn.

Наблюдаемые значения водных отложений по каждому геохимическому индексу были привязаны к сетке в программном обеспечении Surfer с использованием метода обратной интерполяции расстояния, и точность пространства сетки соответствовала сетке расчетных ячеек. Кроме того, среднее и логарифмическое значения стандартного отклонения каждого элемента были преобразованы в истинные значения, а сумма среднего и двукратного стандартного отклонения ($X \pm 2\sigma$) использовались в качестве нижнего предела для извлечения информации о геохимических аномалиях. Карты распределения аномалий каждого элемента наложены и сравнены с известными месторождениями полезных ископаемых (рис. 4).

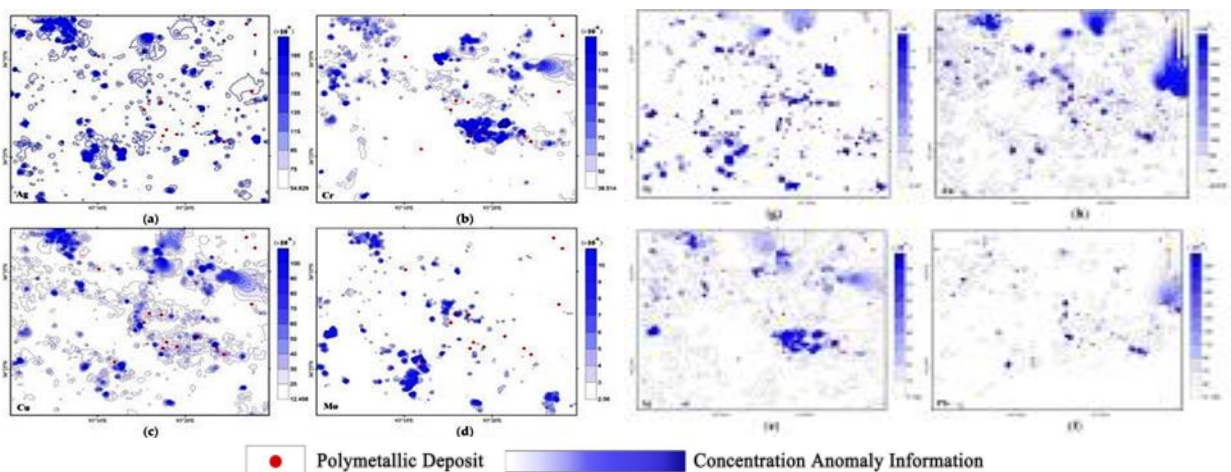


Рисунок 4. Геохимические аномалии Ag, Cr, Cu, Mo, Ni, Pb, W и Zn.
(a) Ag; (b) Cr; (c) Cu; (d) Mo; (e) Ni; (f) Pb; (g) W; (h) Zn.

Можно видеть, что аномалия каждого элемента была тесно связана с известными металлогеническими ячейками, а очерченная область фокусировки аномалии соответствовала пространственному расположению известных рудных залежей, указывая на то, что расчет нижнего предела аномалии был действительным и надежным. Векторная информация была сохранена в пространственной базе данных программного обеспечения MapGis в качестве геохимической информации о оруденении в исследуемой области.

Данные дистанционного зондирования (ДДЗ).

В этом исследовании использовалась ручная визуальная интерпретация для извлечения структурной информации о разломах и зонах контактов на основе особенностей текстуры и формы изображений ДДЗ. Были выбраны данные ETM + и “ZY-1” 02С, где ETM + были данными спутникового изображения Landsat 7 с разрешением 30 м. Линейные структуры среднего масштаба в регионе были выделены с помощью изображений ETM +, а данные “ZY-1” 02С были применены для выделения вторичных линейных структур в ключевой металлогенической области (рис. 5).

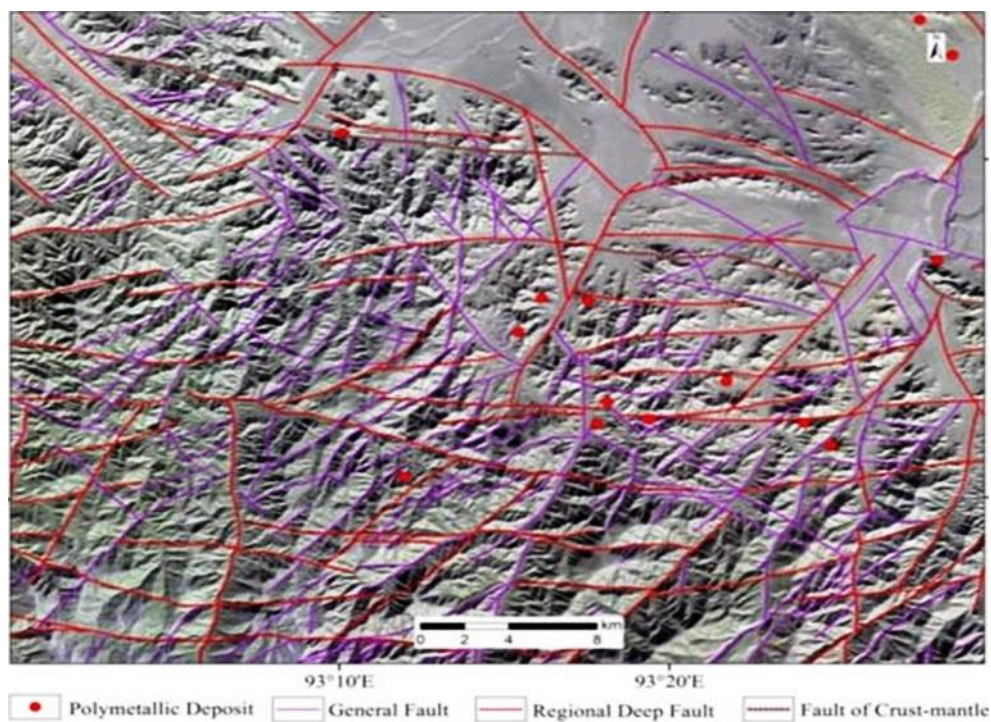


Рис. 5. Карта линеаментов по ДДЗ.

При прогнозировании на основе данных целевой переменной обычно является точечный объект (известное месторождение), а прогнозирующей переменной - объект поверхности. Когда в прогнозируемых переменных присутствуют точечные объекты, необходим буферный анализ в ГИС для преобразования их в объекты поверхности. Линейные структуры были

буферизованы на расстоянии 200 м, а результаты были выражены в виде многоугольных фигур (рис. 6).

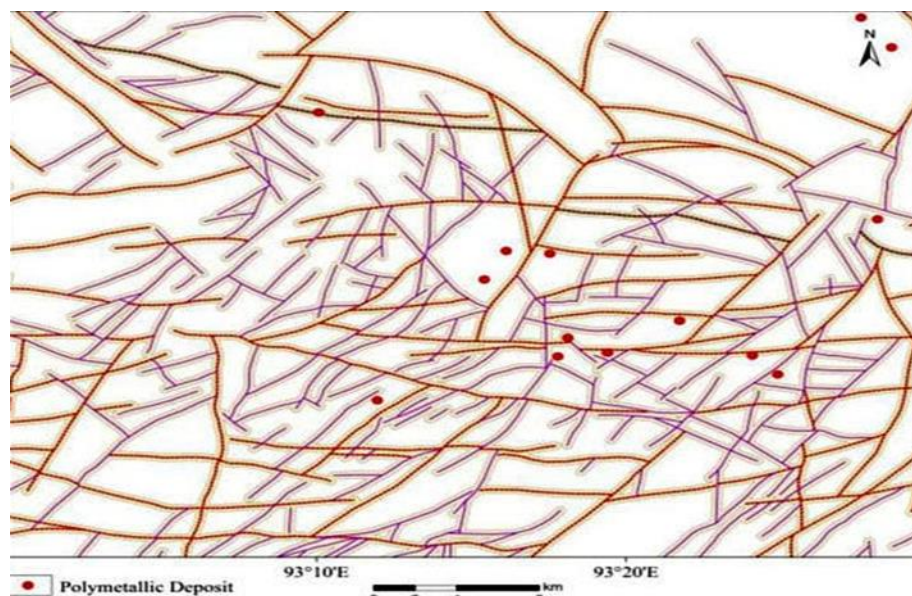


Рис. 6. Буфер линейной конструкции.

Поскольку средняя плотность сбора геохимических проб в точках была не менее $7,1/\text{км}^2$. В сочетании со спектральными характеристиками информация об изменении содержания железа и гидроксила, вызванных характеристиками ферритизации и ярпситовой минерализации, была извлечена из изображений дистанционного зондирования ASTER с анализом основных компонентов. Поскольку характеристики минерализации при хлоризации вызывают аномалию изменения гидроксила, линейные структуры и зоны изменения, которые содержали Fe^+ и OH^- , использовались в качестве фактической информации ДДЗ (рис. 7).

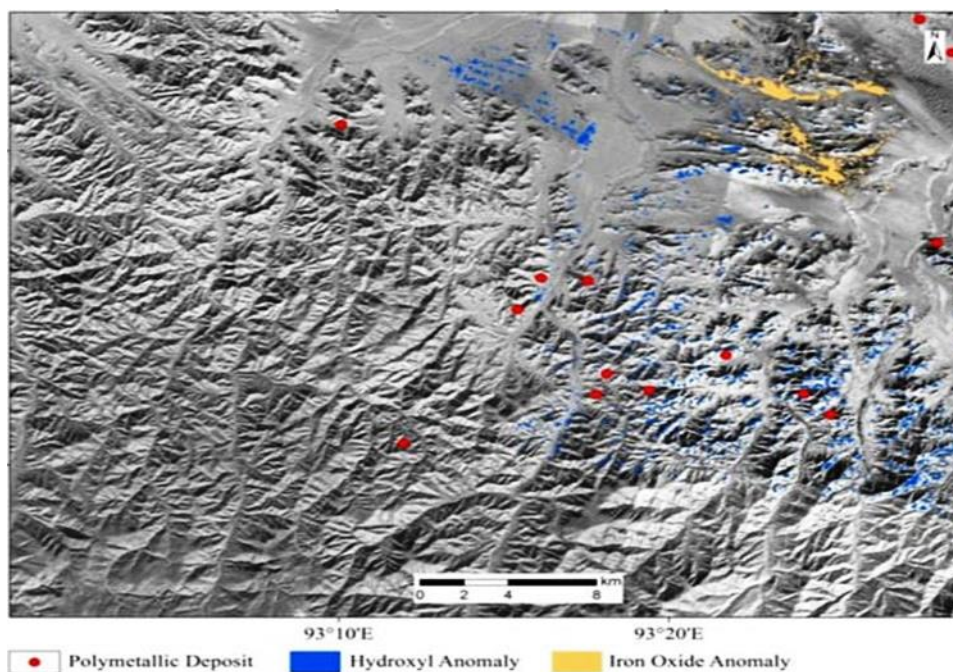


Рис. 7. Распределение информации об изменениях.

4. Картирование перспективности.

4.1. Выбор слоя карты.

При прогнозировании условий металлогенического контроля или измененные состояния условной комбинации присваивались двоичным присваиванием. Если существование объекта могло быть использовано в качестве благоприятного маркера минерализации, ячейке сетки оценки, где находился знак, присваивалось значение 1, в противном случае ему присваивалось значение 0. Ячейки сетки оценок после присвоения были равны дискретным случайным величинам в теории вероятностей. Кроме того, статистический метод был применен для количественного изучения взаимосвязи комплексных информационных переменных друг с другом, чтобы реализовать количественный отбор комплексных информационных переменных. Таким образом, были исключены информационные переменные, которые не были связаны с минерализацией или дублировали другие информационные переменные. Индекс Юдена, обычно используемый в медицинском статистическом анализе, был введен в оптимизацию слоев металлогенических доказательств. Была измерена корреляция между металлогенической информацией и положением месторождений, и комплексные переменные металлогенического прогноза были оптимизированы с использованием индекса Юдена. Индекс Юдена определяется как разница между истинно положительным показателем (TPR) и ложноположительным показателем (FPR).

$$yd = TPR - FPR \quad (20)$$

$$TPR = tp / (tp + fn) \quad (21)$$

$$FPR = fn / (tn + fp) \quad (22)$$

При решении задач в медицинской статистике один класс может быть отмечен как положительный, а другой - как отрицательный. Предполагая, что обучающие выборки состоят из P положительных и N отрицательных выборок, результаты классификации могут быть выражены в следующих четырех ситуациях: tp представляет истинные положительные результаты в прогнозируемых положительных классах; tn представляет истинные отрицательные результаты в прогнозируемых отрицательных классах; fp представляет ложноположительные результаты в прогнозируемых отрицательных классах (фактические отрицательные результаты, признанные положительными); fn представляет ложноположительные результаты в прогнозируемых положительных классах (фактические отрицательные результаты, признанные положительными). позитивы, признанные негативами). TPR представляет собой долю истинно положительных результатов в прогнозируемых положительных классах к общему фактическому положительному результату,

а FPR представляет собой долю ложноположительных результатов в прогнозируемых положительных классах к общему фактическому отрицательному результату. При прогнозировании металлогенической информации TPR представляет собой отношение количества «рудных» ячеек, признанных аномальными, к общему количеству аномальных ячеек в прогнозируемых результатах, а FPR представляет собой отношение количества безрудных ячеек, признанных фоновыми, к общему количеству фоновых ячеек в прогнозируемых результатах. Индекс Юдена находился в диапазоне от -1 до +1, представляя разницу между вероятностью положительного прогноза выборки и отрицательным фактическим прогнозируемым значением. Кроме того, это означает процент выгод, превышающих затраты при прогнозировании, и это статистика, описывающая комплексные характеристики выгод и затрат при металлогеническом прогнозировании. Индекс Юдена для каждого слоя доказательств был рассчитан с помощью обучающих ячеек. Когда индекс Юдена находится в диапазоне от 0 до 1, это указывает на то, что вероятность предсказанного положительного значения, выше, чем вероятность отрицательного фактического значения, а уровень доказательств сильно коррелирует с известными месторождениями полезных ископаемых. Когда индекс Юдена находится в диапазоне от -1 до 0, это указывает на то, что вероятность того, что прогнозируемое значение выборки положительное, ниже, чем вероятность того, что фактическое прогнозируемое значение отрицательное, а именно количество ячеек, не содержащих оруденения, идентифицированных как фоновые ячейки, составляет высокую долю от общего количества прогнозируемых фоновых ячеек. Этот результат означает, что затраты на металлогенический прогноз превышают выгоду и оказывают негативное влияние на результат. Следовательно, слои с индексом Юдена выше 0 имеют пространственную корреляцию с известными месторождениями, что позволяет прогнозировать цели поисков в качестве предикторных переменных. Результаты расчетов приведены в таблице 2.

Табл. 2.

Индекс Юдена для слоев металлогенических свидетельств.

Layer	Correlation Intension	Layer	Correlation Intension
Ag	0.3884	Iron oxide alteration	0.1342
Cr	0.3194	Hydroxyl alteration	0.0801
Cu	0.3942	Linear structure	0.1405
Mo	0.2111	Monzonitic granite	-0.0094
Ni	0.3349	Granodiorite	-0.2158
Pb	0.2073	Quartz diorite	-0.0058
W	0.2114	Carbonate	0.0669
Zn	0.2359		

Из таблицы видно, что 12 слоев доказательств имели индекс Юдена выше 0, включая Ag, Cr, Cu, Mo, Ni, Pb, W, Zn, изменение оксида железа, изменение гидроксила, линейную структуру и карбонат. 12 слоев доказательств, выбранных индексом Юдена, и один целевой слой, содержащий известные месторождения полезных ископаемых, были пространственно наложены и использовались для подготовки обучающих данных в процессе моделирования машинного обучения. Используя 12 слоев доказательств и один целевой слой в качестве входных и выходных данных для моделирования, было подготовлено 18 600 ячеек сетки на исследуемой территории, где ячейкам, содержащим руду, и ее не содержащим, были присвоены значения 1 и 0 соответственно. Эти цифры показывают, что сетка содержала или не содержала известные полиметаллические месторождения (проявления) соответственно.

4.2. Машинное обучение в сочетании с гиперпараметрической оптимизацией.

ВА и FA были применены для оптимизации гиперпараметров различных моделей машинного обучения. Эти два алгоритма оптимизации были необходимы для определения значения пригодности целевой функции в качестве стандарта для расчета гиперпараметрической оптимизации. Когда значение пригодности достигло оптимального значения, гиперпараметры моделей машинного обучения получили оптимальное решение. Значение AUC кривой ROC было выбрано в качестве значения пригодности оптимальной целевой функции. Значение AUC можно рассматривать как вероятность того, что количество правильно предсказанных ячеек сетки выше, чем количество неправильно предсказанных ячеек сетки. Чем выше значение AUC, тем лучше производительность модели при прогнозировании. Значение AUC применяется для оценки эффективности классификации при сравнении нескольких классификаторов. Согласно статистике Уилкоксона Манна–Уитни, значение AUC может быть выражено как:

$$AUC = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \varphi(x_i, y_j) \quad (23)$$

$$\varphi(x_i, y_j) = \begin{cases} 1 & x_i > y_j \\ 0.5 & x_i = y_j \\ 0 & x_i < y_j \end{cases} \quad (24)$$

При использовании ВА и FA для оптимизации гиперпараметров моделей машинного обучения пространством поиска алгоритма было одномерное или двумерное пространство с гиперпараметрами модели в качестве осей координат. Искусственно установив пространственный диапазон гиперпараметров, итеративный процесс поиска начался со случайной позиции в пространстве поиска. Таким образом, гиперпараметр модели, соответствующий максимальному значению функции пригодности, полученной в процессе итеративного поиска, оказался лучшим для моделирования гиперпараметров. В этом

исследовании количество нейронов скрытого слоя в MLP, обновленное время обучения и коэффициент снижения веса слабой регрессии в AdaBoost, а также штрафной коэффициент отказоустойчивости и гиперпараметр ядра в OCSVM были оптимизированы соответственно с использованием двух алгоритмов, и значение AUC кривой ROC было выбрано в качестве функции пригодности. Процесс гиперпараметрической оптимизации показаны на рисунке 8 и 9, а программа металлогенического прогнозирования моделей машинного обучения в сочетании с гиперпараметрической оптимизацией была написана на платформе Python.

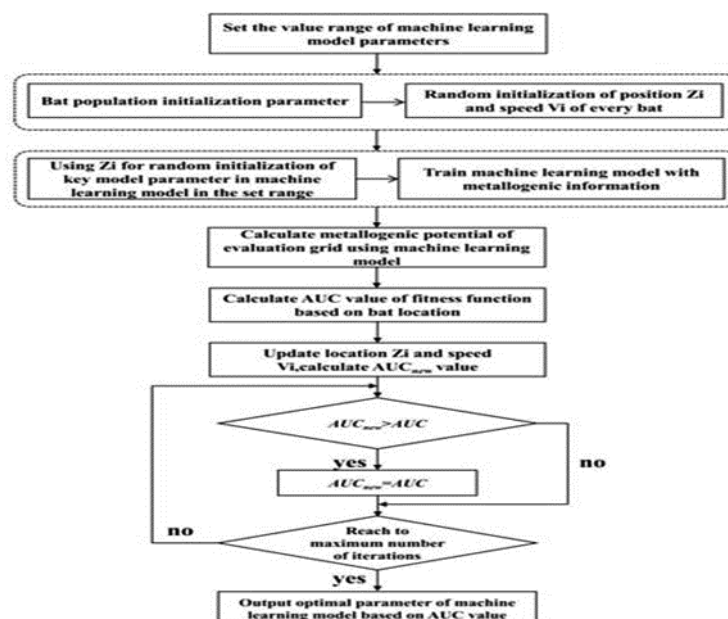


Рис. 8. Процесс оптимизации алгоритма Bat (BA).

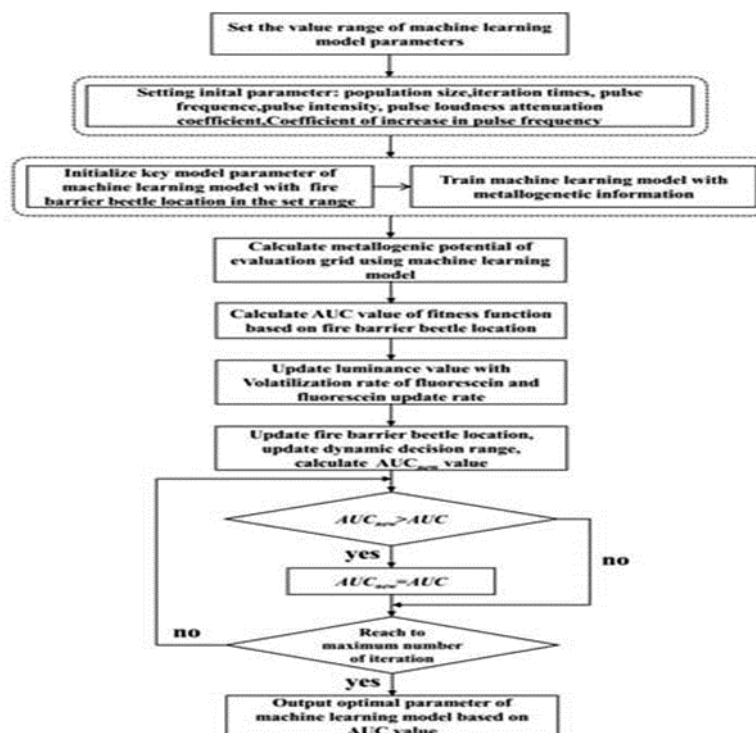


Рис. 9. Процесс оптимизации алгоритма Firefly (FA).

4.2.1. Модель MLP в сочетании с гиперпараметрической оптимизацией.

Металлогеническая вероятность каждой ячейки сетки была рассчитана с использованием MLP-модели для отображения перспективности исследуемой территории. Чем больше значение ячейки сетки, тем больше вероятность того, что ячейка сетки будет ячейкой, содержащей оруденение. В процессе регрессионного прогнозирования установка количества нейронов скрытого слоя, N_{hid} , повлияло бы на результаты прогнозирования модели MLP. Следовательно, количество нейронов скрытого слоя, N_{hid} , был важным гиперпараметром. ВА и FA использовались для оптимизации N_{hid} в модели MLP. Для определения пространства поиска популяций ВА и FA диапазон значений N_{hid} был определен как (0,300], согласно опыту.

Исходные гиперпараметры ВА включают размер популяции L , итерации T , верхняя и нижняя частоты импульсов $f_{мин}$ и $f_{макс}$, верхний и нижний пределы интенсивности импульса $A_{мин}$ и $A_{макс}$ коэффициент ослабления громкости импульса α и коэффициентом увеличения частоты импульсов γ . ВА не чувствителен к другим начальным гиперпараметрам, за исключением итераций T . Размер популяции $L = 20$, итерации $T = 50$, диапазон частот импульсов $f_{мин} = 0$, $f_{макс} = 1$, диапазон интенсивности импульса $A_{мин} = 0$, $A_{макс} = 1$, коэффициент ослабления громкости импульса $\alpha = 0,9$, коэффициент увеличения частоты импульсов $\gamma = 0,9$, и ВА был инициализирован как гиперпараметры по умолчанию.

Исходные гиперпараметры FA были установлены следующим образом: количество светлячков $n = 60$, итерации $T = 50$, летучесть флуоресцеина $\rho = 0,4$, частота обновления флуоресцеина $\gamma = 0,6$, начальное значение яркости $l_0 = 10$, значение шага перемещения $\alpha = 0,6$, динамическая скорость обновления области принятия решений $\beta = 0,8$, предполагаемый радиус $rs = 5$, количество светлячков по соседству $N(t) = 5$, и FA был инициализирован как гиперпараметры по умолчанию.

В соответствии с заданным диапазоном поиска N_{hid} ВА и FA выполнили оптимизационный расчет в определенном пространстве поиска, и значение функции пригодности AUC для разных итераций показано на рисунке 10. Видно, что значение AUC постепенно увеличивалось с увеличением итераций T . При итерациях $T = 27$, значение AUC достигло максимума, как показано на рисунке 10a, и соответствующее $N_{hid} = 96$. При итерациях $T = 36$ кривая достигла стабильного состояния, как показано на рисунке 10b, и соответствующего $N_{hid} = 39$.

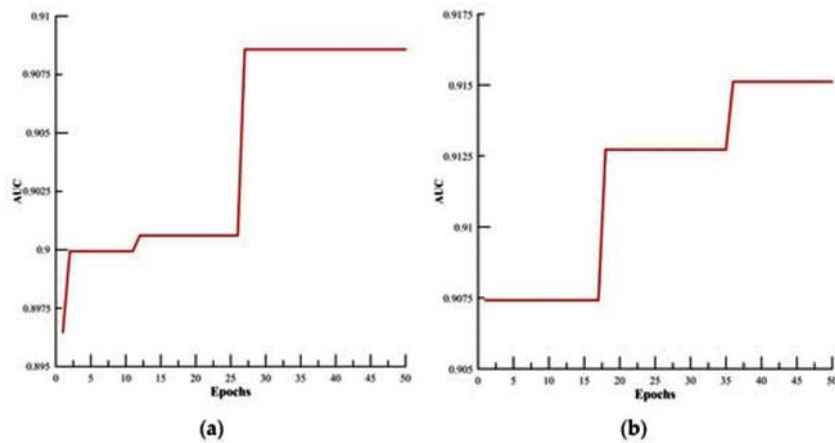


Рис. 10. Кривые величины площади под кривой (AUC):
(а) персептрон алгоритма bat (BA-MLP); (б) персептрон firefly algorithm (FA-MLP).

4.2.2. Модель AdaBoost в сочетании с гиперпараметрической оптимизацией.

Модель AdaBoost была применена к металлогеническому прогнозу района Лалинзаохуо, и дерево решений CART было выбрано в качестве слабой регрессии модели AdaBoost. Во время прогнозирования регрессии время обучения обновлению, t , и коэффициент снижения веса, v ($0 < v < 1$) слабой регрессией были два важных гиперпараметра. Небольшой t было сложно подобрать, хотя и большой t было легко перенастроить. Для того же обучающего набора меньший v это означало, что для выбора слабой регрессии требовалось больше времени. Кроме того, два гиперпараметра t и v должны быть оптимизированы и скорректированы вместе. Следовательно, ВА и ФА соответственно использовались для оптимизации и расчета t и v модели AdaBoost. Для определения пространства поиска популяций ВА и ФА диапазон значений времени обучения обновлению t и коэффициента снижения веса v были определены как $(0, 200]$ и $(0, 1]$ соответственно. В процессе расчета, $T = 20$, а остальные гиперпараметры были установлены такими же, как в разделе 4.2. В процессе оптимизации кривая изменения значения AUC была такой, как показано на рисунке 11. Как видно из рисунка 11, в модели ВА-AdaBoost, когда итерация достигла 12, значение AUC достигло максимума, а время обучения обновлению увеличилось. t и коэффициент снижения веса v составили 98 и 0,1332 соответственно. В модели ФА-AdaBoost, когда итерация достигла 16, значение AUC достигло максимума, а время обучения обновлению увеличилось. t и коэффициент снижения веса v составили 75 и 0,1335 соответственно.

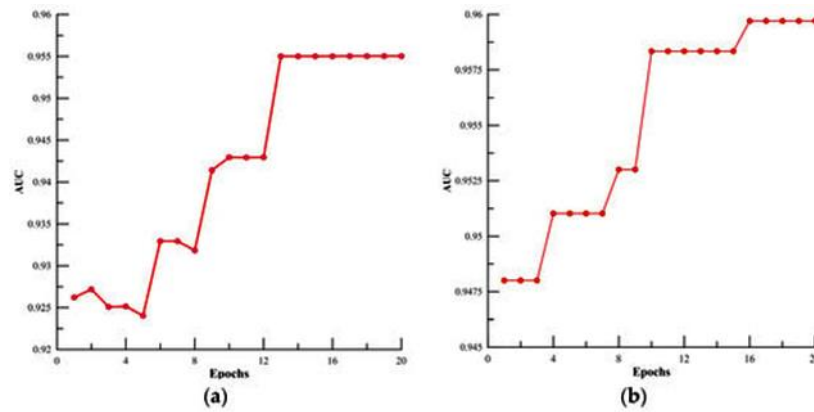


Рис. 11. Кривые значения AUC: (a) BA-AdaBoost; (b) FA-AdaBoost.

4.2.3. Модель OCSVM в сочетании с гиперпараметрической оптимизацией.

Модель OCSVM была использована для вычисления значения решающей функции $f(x)$ каждой ячейки геологической статистики. Если $f(x) > 0$, геологические статистические ячейки принадлежали к области, не содержащей оруденения; в противном случае они принадлежали к области, несущей руду (целевой район поисков). Для удобства применения было использовано отрицательное число статистической решающей функции ячейки. Чем больше она была, тем больше вероятность того, что это была ячейка, содержащая оруденение. В процессе моделирования выбор функций ядра влияет на результаты модели OCSVM. Обычно используемые функции ядра включают полиномиальную функцию ядра, радиальную базисную функцию ядра и сигмоидальную функцию ядра. В качестве основной функции была выбрана гауссова радиальная базисная базовая функция. Штрафной коэффициент отказоустойчивости γ и гиперпараметр ядра σ^2 два необходимых корректирующих гиперпараметра были оптимизированы BA и FA. Для определения пространства поиска популяций BA и FA диапазон значений штрафного коэффициента отказоустойчивости γ и гиперпараметр ядра σ^2 были определены как $(0, 10]$ и $(0, 1]$ соответственно. В процессе расчета итерации $T = 30$, а остальные исходные гиперпараметры были установлены такими же, как в разделе 4.2. В процессе оптимизации кривая изменения значения AUC была такой, как показано на рисунке 12. Как видно из рисунка 12, в BA-OCSVM, когда итерация достигла 14, значение AUC достигло максимума, а выходной коэффициент отказоустойчивости – штрафного γ и гиперпараметр ядра σ^2 составили 0,00845 и 0,88328 соответственно. В FA-OCSVM, когда итерация достигла 19, значение AUC достигло максимального значения, а выходной коэффициент отказоустойчивости – штрафного γ и значения гиперпараметра ядра σ^2 составили 6,82289 и 0,36972 соответственно.

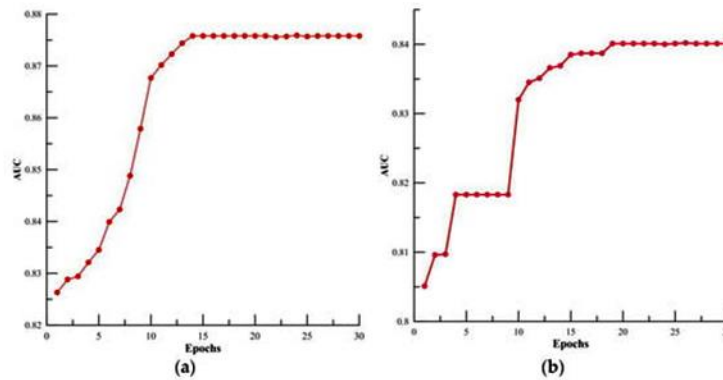


Рис. 12. Кривые значения AUC: (a) алгоритм bat - машина опорных векторов одного класса (BA-OCSVM); (b) алгоритм firefly - машина опорных векторов одного класса (FA-OCSVM).

4.3. Картирование минерального потенциала.

Модель картирования перспективности была создана с использованием оптимизированных гиперпараметров; она была применена для расчета значения металлогенического потенциала каждой ячейки. Диапазон значений минерального потенциала каждой ячейки отличался из-за разных принципов трех моделей машинного обучения. Результаты для моделей MLP и AdaBoost представляли металлогеническую вероятность ячейки сетки оценки, в то время как результаты модели OCSVM были значениями функции принятия решений, соответствующими выбранной функции ядра. Хотя диапазоны результатов трех моделей машинного обучения были разными, это не повлияло на прогнозирование, поскольку минеральный потенциал оценивался в соответствии с прогнозным значением ячейки оценочной сетки. Область с высоким значением минерального потенциала была очерчена с помощью метода оптимального порога. Этот метод определил, что территория, ограниченная оптимальным порогом, и известные месторождения имеют максимальную пространственную корреляцию, где оптимальный порог был рассчитан по индексу Юдена. Между тем, TPR представлял собой отношение количества правильно распознанных ячеек с оруденением к общему количеству ячеек с рудой в прогнозируемом результате, а FPR представлял собой отношение количества правильно распознанных безрудных ячеек к общему их количеству в прогнозируемом результате. Максимум и минимум рассчитанного индекса металлогенических преимуществ были приняты за конечные точки непрерывного интервала, который был разделен на 1000 равных подинтервалов, каждый из которых рассматривался как пороговый. Модель использовалась для расчета TPR и FPR, соответствующих пороговым значениям, а TPR минус FPR использовался для расчета индекса Юдена. Порог, соответствующий максимальному индексу Юдена, был выбран в качестве оптимального порога для определения целей поисков, а максимальные индексы Юдена и оптимальные пороговые значения различных моделей были такими, как показано в таблице 3. С помощью программной платформы ГИС и оптимальных

пороговых значений были созданы карты минерального потенциала различных моделей машинного обучения после оптимизации (рис. 13). Чтобы сравнить эффект металлогенического прогнозирования, области с низким и высоким потенциалом полезных ископаемых были представлены зеленым и синим цветами соответственно, известные месторождения полезных ископаемых были представлены красными точками, а слой известных месторождений полезных ископаемых был наложен на карту минерального потенциала.

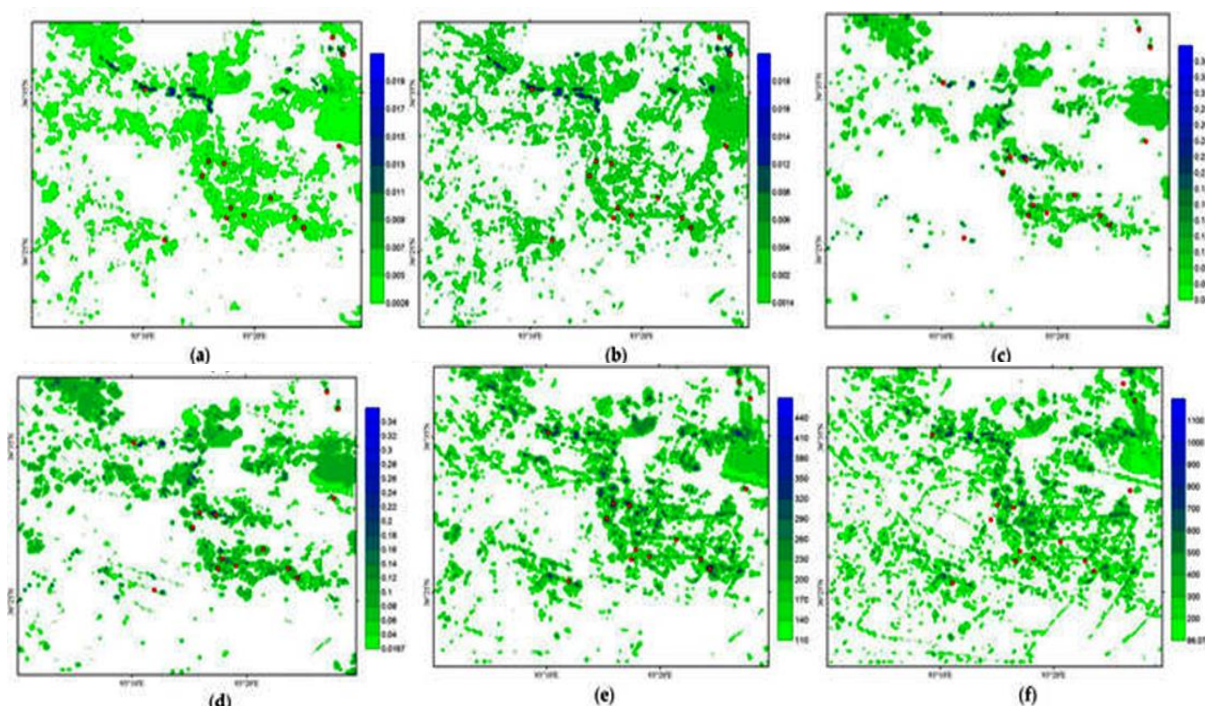


Рис. 13. Карты потенциала полезных ископаемых:
(a) BA-MLP; (b) FA-MLP; (c) BA-AdaBoost; (d) FA-AdaBoost; (e) BA-OCSVM; (f) FA-OCSVM.

Табл. 3.

Максимальные индексы Юдена и оптимальные пороговые значения для различных моделей.

Model	BA-MLP	FA-MLP	BA-AdaBoost	FA-AdaBoost	BA-OCSVM	FA-OCSVM
Maximum Youden index	0.6649	0.55489	0.807039	0.815599	0.70648	0.52872
Threshold	0.0028	0.0014	0.0452	0.0167	109.7745	96.05151

5. Результаты.

Для сравнения и анализа эффекта прогнозирования металлогении с помощью моделей машинного обучения до и после оптимизации параметров были применены оригинальные модели MLP, AdaBoost и OCSVM для изучения моделирования и прогнозирования. Установка значения гиперпараметров по умолчанию в модуле scikit-learn от Python в качестве значения гиперпараметра трех моделей, нейронов скрытого слоя, N_{hid} , для модели MLP было установлено значение 50, время обучения обновлению, t и коэффициента снижения веса, ν , модели AdaBoost были установлены равными 10 и 0,1 соответственно, а также штрафной коэффициент отказоустойчивости, γ , и значение гиперпараметра ядра, σ^2 , для модели OCSVM

были установлены как 1 и 0.1, соответственно. Результаты картирования перспективности были статистически оценены с использованием кривых ROC и P-A. Анализ ROC-кривой находит все более широкое применение в области машинного обучения. Он нечувствителен к распределению категорий и стоимости, а также обладает хорошей интуицией и глубоким пониманием. Система координат для кривой ROC формируется путем использования FPR и TPR в качестве осей X и Y соответственно. После обучения дискретные и двоичные выходные точки соответствуют точкам в системе координат, т. е. набор отдельных точек получается путем установки различных пороговых значений для одного и того же классификатора. Эти точки соединены кривой слева направо в системе координат ROC. Кривая ROC имеет монотонно убывающий наклон. Оптимальный классификатор может быть выбран из набора классификаторов в соответствии с их ROC-кривыми. Чем лучше классификатор, тем ближе его ROC-кривая к верхнему левому углу ROC-пространства. Значение AUC представляет площадь под кривой ROC, и метод его расчета показан в уравнениях (23) и (24), использующих значение AUC для оценки эффективности классификации. Диапазон значений AUC находится в пределах (0,1). Чем ближе кривая ROC к верхнему левому углу пространства ROC, тем ближе значение AUC, рассчитанное моделью, к 1,0. TPR и FPR, соответствующие разным пороговым значениям в диапазоне металлогенической вероятности различных моделей, могут быть получены путем расчета потенциала полезных ископаемых, и кривые ROC нарисованы, как показано на рисунке 14.

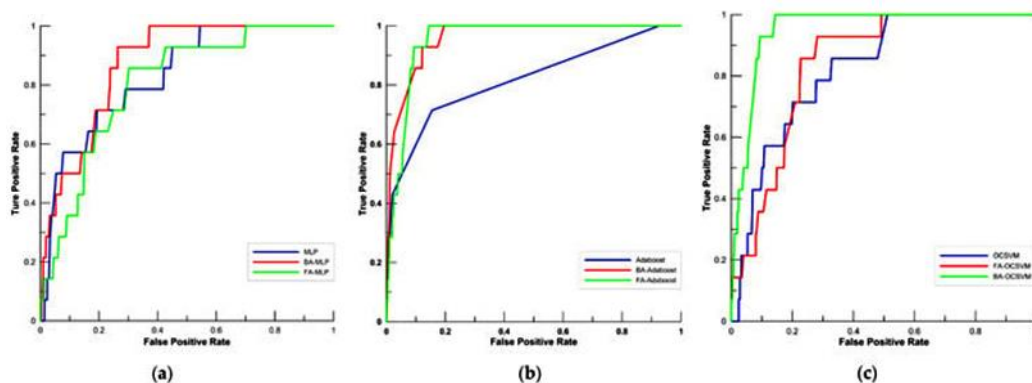


Рис. 14. Кривые рабочих характеристик (ROC) моделей прогнозирования: (a) MLP; (b) AdaBoost; (c) OCSVM.

Можно видеть, что, хотя кривые ROC пересекаются, кривые ROC (синие), соответствующие исходным моделям машинного обучения без гиперпараметрической оптимизации, все ниже, чем у моделей машинного обучения после гиперпараметрической оптимизации, что указывает на то, что оптимизированная модель машинного обучения лучше, чем соответствующая исходная модель прогнозирования.

Кроме того, на исследуемой территории точность прогнозирования модели, P , и процент прогнозируемой площади, A , могут быть рассчитаны в соответствии с картами потенциальных запасов и распределением известных рудных залежей. P представляет преимущества прогнозирования модели, а A представляет стоимость прогнозирования модели, используя значения P и A для построения кривой P - A . В соответствии со значением минерального потенциала ячейки оценочной сетки, рассчитанным с помощью трех моделей, были рассчитаны P и A соответствующих моделей, когда в качестве порогового значения был взят различный металлогенический потенциал. Кроме того, были построены кривые P - A различных моделей машинного обучения. На рисунке 15 горизонтальная ось системы координат представляет собой значение металлогенического потенциала, рассчитанного моделью.

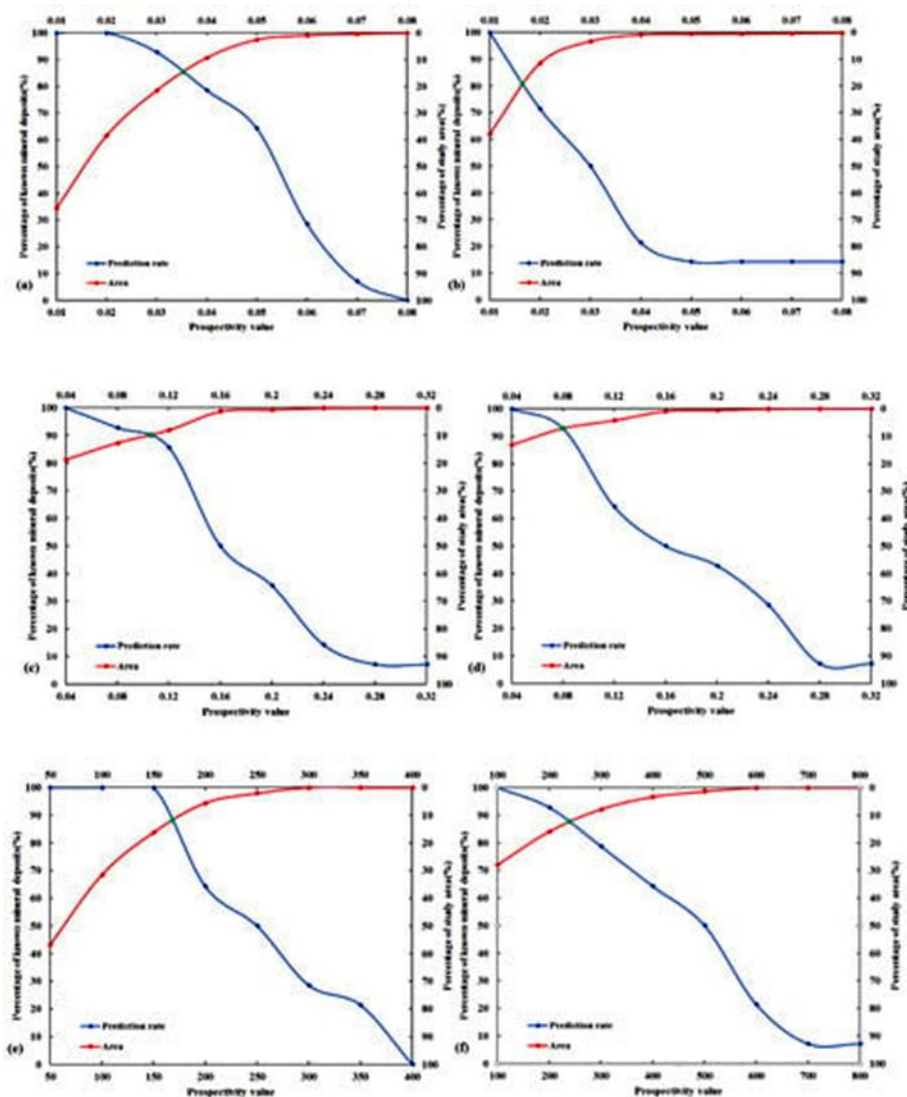


Рис. 15. Кривые площади (P - A) моделей прогнозирования: (a) BA-MLP; (b) FA-MLP; (c) BA-AdaBoost; (d) FA-AdaBoost; (e) BA-OCSVM; (f) FA-OCSVM.

Кроме того, отношение значения P к значению A может быть использовано в качестве статистики. Когда соотношение становится больше, масштаб прогнозируемых целей будет

больше; целевая площадь будет меньше; трудовые, материальные ресурсы и финансовые затраты будут меньше; эффект модели прогнозирования будет лучше. Следовательно, высота точки пересечения кривых точности прогнозирования и процентной доли прогнозируемой площади может быть использованы для оценки производительности модели. Между тем, точка пересечения двух кривых соответствует оптимальному порогу результатов прогнозирования, который может быть использован для дальнейшего определения целевой области поисков (рис. 15).

Были рассчитаны значения AUC и P/A с различными алгоритмами гиперпараметрической оптимизации и сравнены с результатами без гиперпараметрической оптимизации (табл. 4) для анализа влияния моделей машинного обучения на металлогенический прогноз.

Табл. 4.

Показатели оценки различных моделей прогнозирования.

Model	AUC	P/A
MLP	0.7828	2.449
BA-MLP	0.8712	5.533
FA-MLP	0.8017	3.745
AdaBoost	0.8173	3.156
BA-AdaBoost	0.9588	8.182
FA-AdaBoost	0.9597	10.765
OCSVM	0.8056	3.954
BA-OCSVM	0.8758	7.417
FA-OCSVM	0.8262	6.769

Анализируя кривые ROC (рис. 14), кривые P-A (рис. 15) и индексы оценки (табл. 4) моделей машинного обучения, можно сделать следующие выводы:

1. Все значения AUC, полученные моделями машинного обучения, были выше 0,7, что указывает на то, что результаты металлогенического прогнозирования согласуются с регулярностью распределения известных месторождений полезных ископаемых и точек минерализации в этом районе. Наилучший эффект показала модель AdaBoost, где значение AUC достигло 0,9579 после использования алгоритма FA, в то время как значение AUC модели BA-OCSVM составило 0,8758, а модели BA-MLP - 0,8712.

2. Алгоритмы гиперпараметрической оптимизации BA и FA имеют очевидный эффект, и точность моделей MLP, AdaBoost и OCSVM была улучшена после оптимизации. Точность модели FA-AdaBoost была улучшена наиболее значительно, с точностью на 17,42% выше после оптимизации. Для моделей MLP и OCSVM точность алгоритма BA после оптимизации была выше, чем у алгоритма FA, а точность прогнозирования BA-MLP увеличилась на 11,29%, в то время как у модели FA-MLP увеличилась только на 2,41%.

6. Обсуждение.

1. BA и FA, которые имеют одинаковые структурные характеристики, относятся к алгоритму глобальной оптимизации, основанному на случайной оптимизации популяции. В области поиска каждая летучая мышь или светлячок представляет собой решение функции оптимизации; значение пригодности оценивает их положение и находит оптимальную особь путем корректировки гиперпараметров ее собственной популяции. Из-за различий в структурах моделей машинного обучения в тесте определяется разное время итерации. Гиперпараметр оптимизации модели MLP всего один, поскольку ее структура проста, в то время как количество гиперпараметров оптимизации моделей AdaBoost и OCSVM равно двум. Позднее определенное время итерации составляет 20 и 30, а скорость вычислений низкая. Результаты показывают, что BA и FA могут достигать глобального максимума по количеству установленных итераций, и в то же время модели AdaBoost и OCSVM экономят больше времени на поиск оптимальных гиперпараметров.

2. Результаты индексов оценки показывают, что характерная регулярность кривой ROC в основном согласуется с кривой AUC, в то время как кривая P-A и индекс P/A могут обеспечить другой стандарт оценки точности модели прогнозирования с точки зрения эффективности поисков. Например, значение AUC BA-MLP равно 0,8712, в то время как значение AUC BA-OCSVM равно 0,8758. Рассчитанные значения показателей оценки кривой ROC для двух моделей очень близки, что означает, что эффект оценки модели нельзя сравнивать. Однако можно обнаружить, что пересечение кривой P-A модели OCSVM, очевидно, выше, чем у модели BA-MLP, где значение P/A BA-OCSVM равно 7,417, в то время как значение P/A BA-MLP равно 5,533, что указывает на превосходство модели BA-OCSVM с точки зрения эффективности поисков.

3. При сравнении трех видов карт потенциала полезных ископаемых (рис. 13) после гиперпараметрической оптимизации диапазон и тенденция распределения в пространстве прогнозируемого участка с высоким потенциалом имеют высокое сходство. Помимо известной рудной зоны, модели очерчивают потенциальную зону поиска на северо-востоке, северо-западе и юго-западе исследуемой территории. Минерализованная геологическая среда северо-востока относительно превосходна, где основными образованиями являются ордовик-силурийская группа Тандзяньшань и карбонатная формация раннего каменноугольного периода группы Дагангоу, а монцонитовый гранит позднего триаса является основным рудолокализирующим массивом. Скарновые месторождения, с большей вероятностью, образуются в этом регионе, что совпадает с прогнозными результатами модели, и будет ключевой областью для ГРП в будущем. Разлом Кунбэй находится на северо-западе, группа Тандзяньшань и карбонатная

формация раннего каменноугольного периода группы Дагангоу распределены линейно вдоль направления структурной линии под контролем этого разлома. Монцогранит ранней юры распространен на севере осадочных толщ, где зоны контактов с большей вероятностью образуют скарновый пояс. Если сложить металлогенический потенциал и геологическую карту, то частичная область с высокой ценностью находится в зоне контакта пластов карбонатных пород и интрузивного тела. Металлогенический потенциал на юго-западе распределен в интрузивном теле разных периодов, а зона с высокой ценностью распределена спорадически, при этом вероятность минерализации очень мала.

4. BA и FA справляются с вычислениями гиперпараметрической оптимизации; однако при моделировании расчетов обнаруживается, что все еще существуют некоторые проблемы в сочетании такого рода алгоритма оптимизации роевого интеллекта и модели машинного обучения, которые нуждаются в дальнейшем анализе. Одна из проблем заключается в том, что обоим алгоритмам необходимо задавать большое количество начальных гиперпараметров. Однако в этом исследовании проверяется только соответствующая взаимосвязь между итерациями T и значением функции пригодности, а в других используются значения по умолчанию. Другая проблема заключается в том, что модель машинного обучения, оптимизированная с помощью такого алгоритма, требует длительного времени для запуска. В модели BA-MLP время расчета модели машинного обучения с оптимизированными гиперпараметрами примерно в восемь раз больше, чем у исходной модели, что в определенной степени влияет на эффективность тестирования.

7. Выводы.

1. Модели картирования перспективности строятся в условиях, когда необходимо задать пространство поиска и автоматически выполнять поиск оптимальных гиперпараметров путем объединения алгоритмов оптимизации интеллекта bat и firefly swarm с различными моделями машинного обучения. По сравнению с традиционными алгоритмами оптимизации, BA и FA могут свободно переключаться между процессами глобальной и локальной оптимизации и имеют больше возможностей для нахождения глобальных оптимальных гиперпараметров моделей машинного обучения.

2. BA и FA оказывают различное улучшающее воздействие на модели MLP, AdaBoost и OCSVM. Точность моделей машинного обучения значительно повышается после гиперпараметрической оптимизации, что указывает на то, что гиперпараметрическая оптимизация модели является эффективной и надежной при применении методов машинного обучения.

3. Кривые ROC и P-A применяются для количественной оценки эффективности прогнозирования. Результаты оценки показывают, что значение AUC может эффективно измерять точность различных моделей, но это не единственный показатель. Значение P/A пересечения кривой рассчитывается по кривой P-A. Чем выше значение, тем точнее металлогенический прогноз. Кривая P-A представляет эффективный прогнозный параметр при ограниченных трудовых, материальных и финансовых ресурсах, а значение P/A может использоваться в качестве стандарта оценки точности для прогноза

III. ПРИМЕНЕНИЕ ПРОЦЕССА АНАЛИТИЧЕСКОЙ ИЕРАРХИИ ПРИ ПРОГНОЗИРОВАНИИ МИНЕРАЛЬНОГО ПОТЕНЦИАЛА НА ОСНОВЕ ИНТЕГРИРОВАННОЙ ГЕОЛОГО-АЭРОГЕОФИЗИКО-ГЕОХИМИЧЕСКОЙ МОДЕЛИ [7].

1. Введение.

Процесс аналитической иерархии является широко используемым методом для решения задач, связанных с комплексной оценкой множества факторов. Он сочетает качественные и количественные методы для разделения задач прогнозирования на несколько ключевых влияющих факторов, которые оценивают вероятность минерализации и формируют иерархическую структуру в соответствии с их доминирующими взаимосвязями. Целевые районы поисков могут быть определены путем попарных сравнений. Так, например, была подтверждена эффективность процесса аналитической иерархии при прогнозировании минерализации урана в бассейне Тарима, сурьмяного (золотого) оруденения в рудном районе Пуцин в Гуйчжоу и др.

В этом материале построена система индексов оценки аэрогеофизических поисков, объединив их с геологическими и геохимическими данными. Мы рассчитали относительные веса каждого показателя, чтобы сформировать стандарт оценки для геолого–аэрофизико-геохимической интегрированной модели. Также была создана карта информационной энтропии модели, обеспечивающей основу для определения целевых районов поисков.

2. Процесс аналитической иерархии (АНР).

Этот метод позволяет оцифровать личный опыт и мышление, реализуя комбинацию качественных и количественных методов, и широко применяется в различных областях. В данном исследовании АНР использовался для разделения задач прогнозирования минерализации на несколько рудолокализирующих факторов (геологические, геофизические и геохимические характеристики) и создания объективной иерархической аналитической модели для поисков. В соответствии с обобщенным опытом предыдущих исследователей по минерализации были проведены сравнения относительной важности для каждого уровня и фактора с соответствующим присвоением числовых значений. Весовые значения каждого уровня и фактора были рассчитаны с использованием математических методов, а целевые участки поисков были спрогнозированы на основе выявленных и ожидаемых аномалий. Процесс моделирования с использованием АНР можно разделить на следующие четыре этапа:

Создание модели АНР.

При использовании АНР для анализа проблем принятия решений первый шаг включает организацию и иерархизацию проблемы и построение модели иерархической структуры. В соответствии с этой моделью сложные задачи разлагаются на несколько уровней и постепенно анализируются на уровнях, которые намного проще исходного для решения сложных задач. Количество уровней в этой иерархической структуре зависит от сложности задачи и уровня детализации, необходимого для анализа. Как правило, количество уровней не ограничено. Однако рекомендуется, чтобы каждый элемент на каждом уровне доминировал не более чем над девятью другими элементами, чтобы избежать трудностей при попарном сравнении и вынесении суждений, вызванных большим количеством элементов. Надежная иерархическая структура чрезвычайно важна для решения задач. Следовательно, он должен основываться на всестороннем понимании проблемы лицом, принимающим решения. Взаимосвязи между элементами должны быть прояснены, чтобы обеспечить установление разумной иерархической структуры.

Построение матриц оценок и шкал.

Оценка уровня А с использованием матрицы суждений включает сравнение относительной важности факторов B1, B2, ... и Bn на следующем уровне В. Это сравнение обычно выражается в следующей форме (табл. 1).

Табл. 1.

Общая форма матрицы оценок.				
A	B1	B2	...	Bn
B1	B1/B1	B1/B2	...	B1/Bn
B2	B2/B1	B2/B2	...	B2/Bn
⋮	⋮	⋮	⋮	⋮
Bn	Bn/B1	Bn/B2	...	Bn/Bn

Для количественной оценки матрицы суждений используются числа от одного до девяти и их обратные величины в качестве шкалы. В таблице 2 приведены значения, связанные со значениями шкалы от одного до девяти.

Табл. 2.

Значение шкалы матрицы суждений.	
Scale	Meaning
1	Two elements are compared; both elements have equal importance.
3	Two elements are compared; the former is slightly more important.
5	Two elements are compared; the former is significantly more important.
7	Two elements are compared; the former is strongly important.
9	Two elements are compared; the former is extremely important.
2, 4, 6, 8	Represent the median values of the adjacent judgments mentioned above.

Весовые коэффициенты одноуровневого ранжирования и проверка согласованности.

На основе полученной матрицы оценок вес может быть рассчитан путем определения максимального собственного значения и собственного вектора. Затем вектор максимального веса нормализуется и выводится. Из-за сложности объективных вопросов и разнообразия человеческого познания в заключении проводится проверка на согласованность, чтобы убедиться, что суждения соответствуют здравому смыслу. Этапы проверки согласованности следующие:

1. Рассчитывается индекс согласованности C_I :

$$C_I = \frac{\lambda_{max} - n}{n - 1}$$

где λ_{max} является максимальным собственным значением матрицы оценок и n порядок матрицы суждений.

2. Определяется соответствующий средний индекс согласованности R_I :

Средний индекс случайной согласованности R_I - средний индекс согласованности, полученный путем вычисления 1000 образцов с матрицей положительных обратных значений в диапазоне от одного до девяти, как показано в таблице 3.

Табл. 3.

Средний индекс случайной согласованности.

n	1	2	3	4	5	6	7	8	9
R_I	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46

3. Рассчитывается коэффициент согласованности C_R :

$$C_R = \frac{C_I}{R_I}$$

На основе формул (1) и (2) значение C_I может быть рассчитано: когда C_I меньше 0,10, согласованность матрицы оценок считается приемлемой. В противном случае в матрицу оценок следует внести соответствующие корректировки.

Расчет весов общего ранжирования иерархии.

Если верхний уровень А содержит m факторов, а именно $A_1, A_2, \dots, A_m$ с соответствующими общими весами иерархического ранжирования $a_1, a_2, \dots, a_m$, а нижний уровень В содержит n факторов, а именно $B_1, B_2, \dots, B_n$ и их одноуровневых весовых коэффициентов ранжирования для факторов A_j обозначаем как $b_{1j}, b_{2j}, \dots, b_{nj}$ (если B_k не связано с A_j , $b_{kj} = 0$), тогда общие веса ранжирования иерархии уровня В могут быть получены в соответствии с информацией, представленной в таблице 4.

Таблица расчета веса иерархии при сортировке.

B	A	A ₁	A ₂	...	A _m	Total Hierarchy Ranking Weights of Level B
		a ₁	a ₂	...	a _m	
B ₁		b ₁₁	b ₁₂	...	b _{1m}	$\sum_{j=1}^m b_{1j}a_j$
B ₂		b ₂₁	b ₂₂	...	b _{2m}	$\sum_{j=1}^m b_{2j}a_j$
⋮		⋮	⋮	⋮	⋮	⋮
B _n		b _{n1}	b _{n2}	...	b _{nm}	$\sum_{j=1}^m b_{nj}a_j$

Относительный вес важности уровня С по отношению к общей цели также может быть получен с использованием матричного масштаба и метода расчета, упомянутых ранее.

3. Построение и вес системы оценочных индексов.

3.1. Построение системы оценочных индексов.

Целью оценки является вынесение суждений относительно оцениваемого объекта. Следовательно, создание системы оценочных индексов для оптимального выбора целевых районов поисков должно обеспечивать конкретную и оперативную систему оценочных индексов. Для создания этой системы необходимо определить общую цель, установить целевой уровень, разложить общую цель на первичную систему индексов, вторичную систему индексов и так далее. В конечном итоге это формирует таблицу общей системы оценок.

В соответствии с поисковой целью на полиметаллические руды, такие как медь, свинец, цинк и серебро – они являются объектами прогнозирования на основе аэро-геофизических аномалий. Концептуальная модель иерархического анализа прогнозирования и оценки аэрогеофизических аномалий показана на рисунке 1.

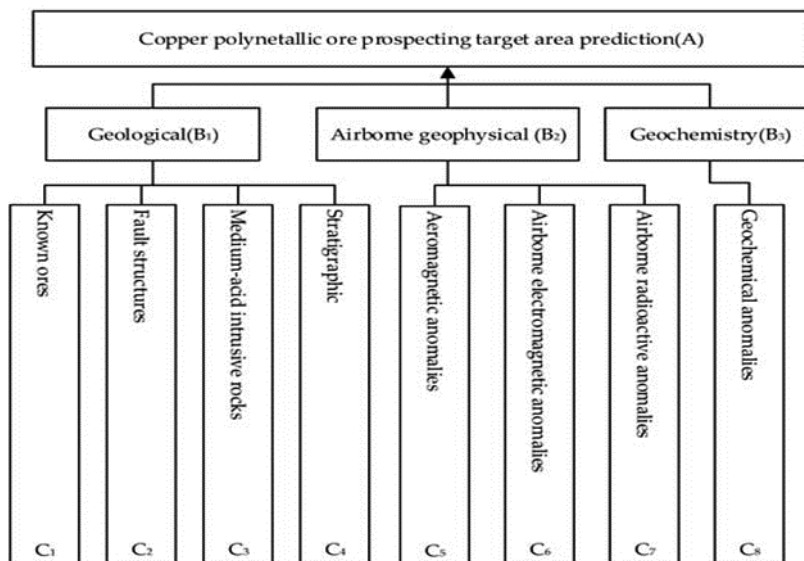


Рис. 1. Концептуальная модель иерархического анализа для прогнозирования полиметаллических руд.

Модель состоит из двух уровней. Первый уровень модели (А) по определению целевых минерализованных участков. Второй уровень - критериев (В, С), включает три основных уровня: геологическую, аэрогеофизическую и геохимическую комплексную оценку (В) и восемь конкретных оценочных показателей подуровней (С).

1. Геологические признаки (В₁)

(1) *Известные руды (С₁)*: обобщение информации об известных рудных месторождениях и точках минерализации является основным признаком для прогнозирования целевого участка.

(2) *Структуры разломов (С₂)*: согласно региональным геологическим данным, структурные признаки в районе включают разломы и пересечения двух или более разломов.

(3) *Средне-кислые интрузивные породы (С₃)*: большинство месторождений в этом районе сосредоточено в пермских и юрских интрузивных породах, которые служат первичными рудообразующими породами. Зоны контактов с вмещающими породами считаются рудолокализуемыми.

(4) *Стратиграфический (С₄)*: статистическое распределение оруденения показывает, что благоприятными стратиграфическими факторами для медно-полиметаллических руд являются верхне-юрские и нижне-пермские породы.

2. Аэрогеофизические признаки (В₂)

(1) *Аэромагнитные аномалии (С₅)*: аномалии, которые тесно связаны с известной минерализацией, аномалии, предположительно связанные с минерализацией, аномалии с неясными свойствами и аномалии, которые, как предполагается, имеют низкую поисковую значимость.

(2) *Аэро-электромагнитные аномалии (С₆)*: аномалии, предположительно являющиеся зонами минерализованных изменений или зонами минерализованных разломов; аномалии, предположительно являющиеся зонами разломных нарушений; аномалии, предположительно относящиеся к породам с низким удельным сопротивлением; или аномалии с неясными свойствами.

(3) *Аэро-радиометрические аномалии (С₇)*: два типа аномалий, а именно аномалии калия и аномалии урана и тория.

3. Геохимические аномалии (В₂)

Согласно собранным геохимическим почвенным данным в масштабе 1:200 000, основными полиметаллическими рудами, являются руды Ag, Cu, Au и Zn. Характеристики аномальных элементов, масштаб и зональность концентрации являются важными признаками для поисков (С₈).

3.2. Определение весов оценочных индексов.

В соответствии с принципом АНР были построены две матрицы попарного сравнения для первичных и вторичных показателей системы индексов оценки для оптимального выбора целевых районов поисков. Окончательная матрица дискриминации была получена путем численного преобразования. Были рассчитаны ее максимальное собственное значение и вектор веса. Затем была выполнена проверка согласованности, чтобы окончательно определить вес каждого показателя и составить таблицу модели оценки информационной энтропии на основе значений веса, как показано в таблицах 5 и 6.

Табл. 5.

Начальное взвешенное значение каждого оценочного показателя.			
Evaluation Contents	Initial Weights	Evaluation Indicators	Initial Weights
Geological	0.5	Known ores	0.20
		Stratigraphic	0.10
		Medium-acid intrusive rocks	0.10
		Fault structures	0.10
Airborne geophysical	0.3	Aeromagnetic local anomalies	0.14
		Airborne electromagnetic local anomalies	0.08
		Airborne radiometric local anomalies	0.08
Geochemical	0.2	Geochemical anomalies	0.20

Табл. 6.

Модели комплексной информационной оценки геолого-аэрогеофизической и геохимической информации.			
Indicators		Assignment Signs	Values
Geological (50 points)	Known ores (20 points)	Large, medium, and small-sized ore deposits	20
		Within 1 km of known ore deposits.	
		Mineralization points	15
	Stratigraphic (10 points)	Within 0.5 km of known mineralization points.	
		Alteration of surrounding rocks.	10
		Ore-bearing stratigraphic layer. ($P_1d^2, P_1t, P_1, P_1b, P_1zs, P_1ds, P_2l, P_1g, P_2s, P_1w, P_1sh, P_1x$)	10
		Ore-bearing stratigraphic layer. ($O_2h, O_1w, D_1, D_2t, C_2b, C_3h^4, D_3a, T_1^1, T_1^2, T_1h, S_3$)	8
		Ore-bearing stratigraphic layer. (J_2f, J_2h, J_3mn)	6
Other stratigraphic layers.	4		
Geochemical (20 points)	Multi-element combination anomalies, there are two or more ore-forming elements with anomalous content of Cu, Pb, Zn, and Ag, at least one of which is an internal zone, and there are front-end element anomalies (one or more of As, Sb, and Hg); Cu, Mo, and Au with front-end element combination anomalies, and either Cu or Mo is in the internal zone.		20
	Multi-element combination anomalies, there are one or more ore-forming elements with anomalous content of Cu, Pb, Zn, and Ag, at least one of which is in the middle zone, and there are front-end element anomalies (one or more of As, Sb, and Hg); Cu, Mo, and Au with front-end element combination anomalies, and either Cu or Mo is in the middle zone.		16
	Multi-element combination anomalies, there are two or more ore-forming elements with anomalous content of Cu, Pb, Zn, and Ag, at least one of which is in the middle zone or above, and there are tailing elements (one or more of W, Sn, Mo, and Bi) without front-end element anomalies; Cu, Mo, and Au combinations without front-end element anomalies; multi-element combination outer zone anomalies, i.e., there are ore-forming elements (one or more of Cu, Pb, Zn, and Ag) and front-end elements (one or more of As, Sb, and Hg).		12
	Multi-element combination outer zone anomalies, without front-end element (As, Sb, and Hg) anomalies; multi-element combination anomalies without ore-forming elements (Cu, Pb, Zn, and Ag); single-element anomalies in the middle zone or above.		8
	Any single-element outer zone anomaly.		4

Note: The last column, "values" is an evaluation score, which is dimensionless data.

Indicators	Assignment Signs	Values
Medium-acidic intrusive rocks. (10 points)	Rock stock or branch	Contact zone (within and outside 0.5 km).
		10
	Permian and Jurassic rocks	Within rock stocks and branches.
		8
	Base rock	Contact zone (within and outside 0.5 km).
		8
	Other rocks	Within rock bases.
		6
		Contact zone (within and outside 0.5 km).
		6
Airborne geophysical (30 points)	Rock stock or branch	Within rock stocks and branches.
		4
	Base rock	Contact zone (within and outside 0.5 km).
		4
		Within rock bases.
		2
	Fault structure (10 points)	Intersection of two or more faults.
		10
		One fault
		7
Airborne geophysical (30 points)	Aeromagnetical anomalies (14 points)	Anomalies closely related to known mineralization (Type A anomalies).
		14
		Anomalies inferred to be related to mineralization (Type B anomalies).
		10
		Anomalies with unclear properties (Type C anomalies).
		6
		Anomalies inferred to have no exploration significance based on existing data (Type D anomalies).
		4
	Airborne electromagnetic anomalies (8 points)	Anomalies inferred to be mineralized alteration zones or mineralized fault zones
		8
		Anomalies inferred to be fault fracture zones
		6
		Anomalies inferred to be low-resistivity rock or anomalies with unclear properties
		4
Airborne radiometric anomalies (8 points)		Potassic alteration zone anomalies
		8
		Uranium and thorium anomalies.
		6

3.3. Метод взвешивания моделей прогнозирования.

Используя вышеупомянутую комплексную информационную поисковую модель АНР с размером сетки $2 \text{ км} \times 2 \text{ км}$, вся территория разделена на несколько ячеек сетки одинакового размера. Информационные значения каждого поискового показателя в пределах каждой единицы сетки суммируются. Полученное комплексное информационное значение используется для построения карты, как показано на рисунке 2.

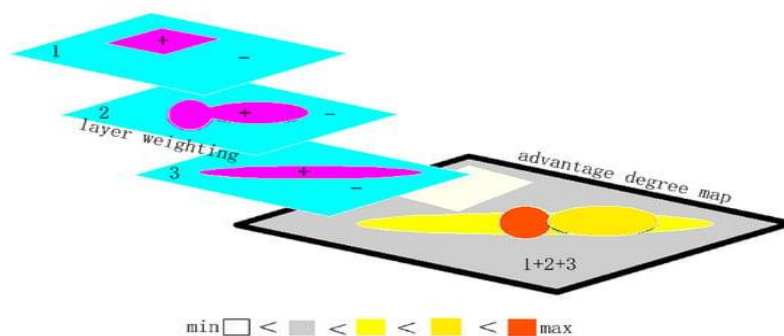


Рис. 2. Схема алгоритма информации о метках в перспективной рудной зоне.

Минимальный порог для района поисков определяется путем сравнения информационных значений известных минерализованных участков.

4. Применение.

Центрально-южная часть Великого хребта Синъянь характеризуется наложением, сочетанием и трансформацией тектоно-металлогенических области палеозоя и кайнозоя, демонстрируя типичные черты многочисленных аккреционных орогенных поясов. Рудные месторождения этой области в основном включают железо, медь, никель, золото, серебро, олово, свинцово-цинковые, медно-цинковые, медно-свинцовые, свинцово-цинково-серебряные, железо-полиметаллические и медно-полиметаллические. В соответствии с различными генетическими типами известные месторождения в регионе могут быть классифицированы на четыре категории: гидротермальный тип, скарновый, порфировый и редкоземельные месторождения щелочных гранитов.

С использованием АНР была сгенерирована карта информационной ценности, как показано на рисунке 3.

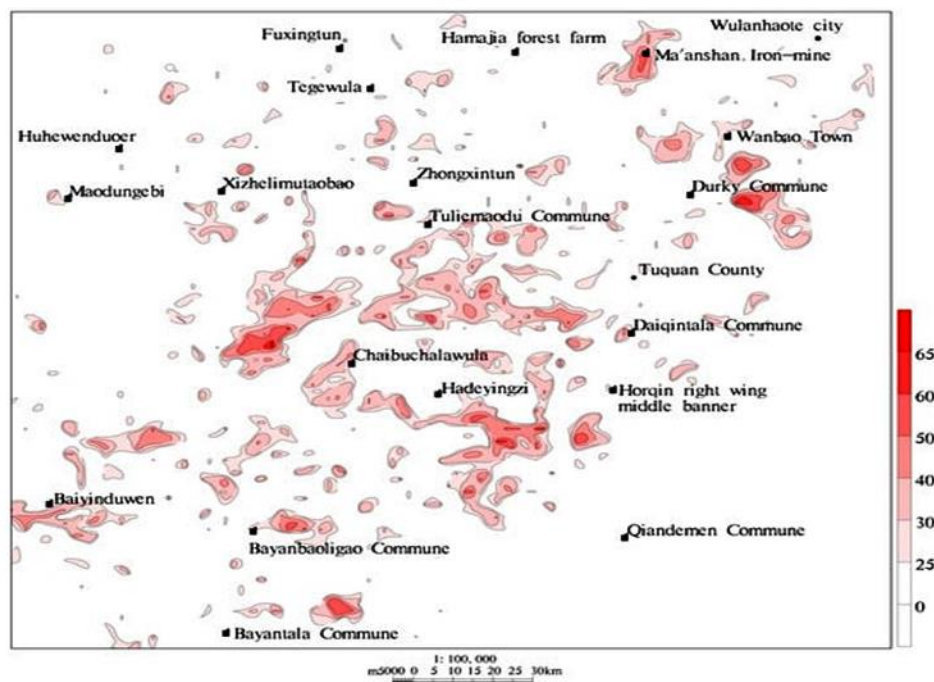


Рис. 3. Карта района работ со всеобъемлющей информацией.

Эта карта предоставляет важную информацию для определения целевого района поисков. Сравнивая информационные значения известных минерализованных участков, можно определить нижний предел комплексной информационной ценности для целевого района поиска, равный 35.

В качестве примера рассматривается оптимизация целевого участка поисков нескольких металлов на медно–железном полиметаллическом месторождении Наониушань. В районе развиты разломы, в основном северо-восточного и северо-западного направлений, аналогичные структурам месторождения меди (рис. 4).

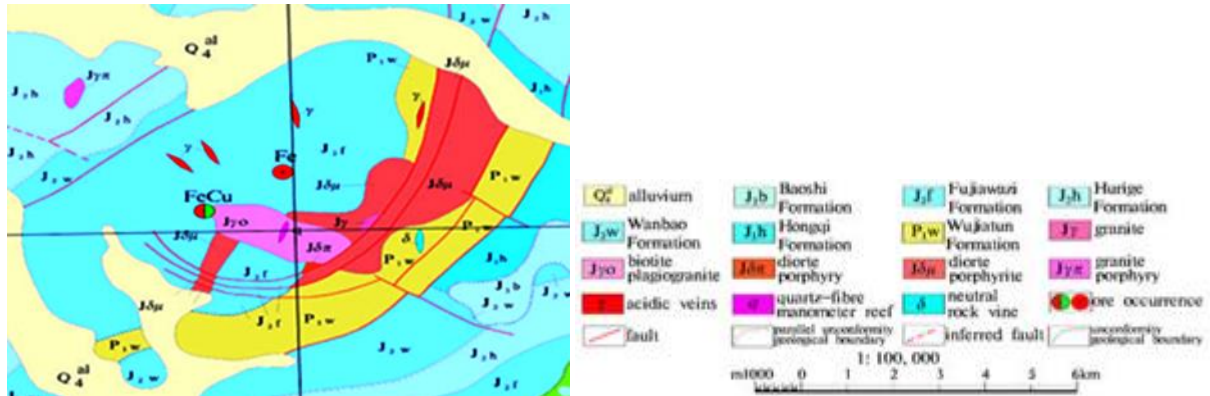


Рис. 4. Геологическая карта м-ния Наонюшань.

Между двумя поясами аэромагнитных аномалий, ориентированными с севера на северо–восток, известны два месторождения железа и меди. Пояс аэромагнитных аномалий средней интенсивности соответствует диорит-порфиритам и гранитам. Максимальная интенсивность аномалий достигает 1400 нТл (рис. 5).

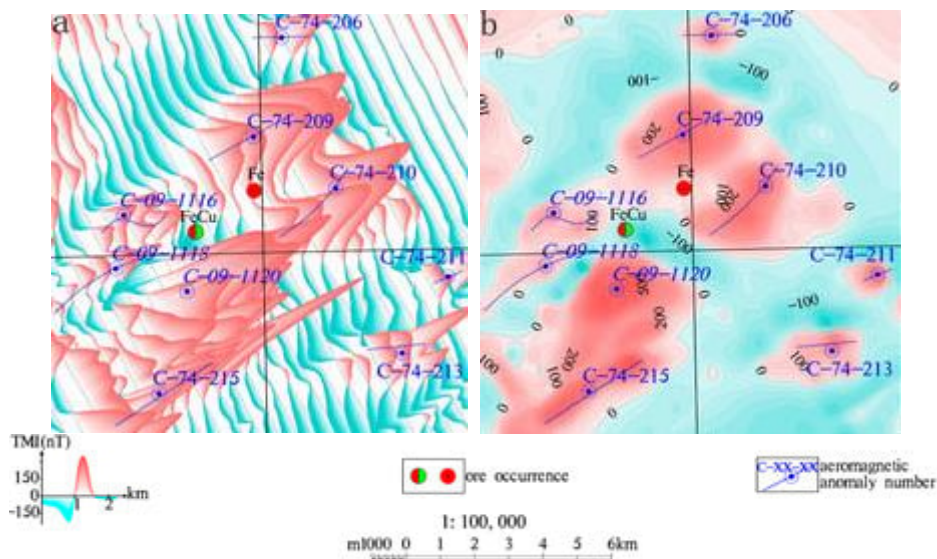


Рис. 5. Аэромагнитная карта общей магнитной напряженности (ТМН) района Наонюшань. (а) разрезы; (б) в плане.

Анализ данных о физических свойствах пород показывает, что ни одно из них не может генерировать такую сильную аэромагнитную аномалию. Предполагается, что аномалия вызвана скрытой интрузией средней кислотности, контролирующей оруденение, сходное с медным месторождением Ляньхуашань. Магнитное поле в этой области обладает характеристиками вулканической структуры: основное жерло вулкана расположено там, где

расположено гранитно-порфировое тело рядом с месторождением железа и меди, что соответствует слабой отрицательной магнитной аномалии на карте магнитного поля. Четыре магнитные аномалии средней интенсивности распределены вдоль главного вулканического жерла, большинство из которых указывают на наличие диорит-порфирита. Предполагается, что они могут быть продуктами небольших жерл, вдоль дугообразной зоны разлома вокруг главного вулканического жерла.

Карта аэро-электромагнитного поля показывает, что за исключением северо-востока и юго-запада целевого района, где наблюдаются заметные электромагнитные аномалии, в большинстве других районов преобладают ровные и низкие фоновые поля. Следовательно, все они являются телами с высоким сопротивлением и благоприятны для минерализации.

Карта аэро- гамма-спектрометрии показывает, что отражательная способность горных пород в этом районе относительно низкая, обычно она проявляется как зона с низким значением радиоактивности четвертичных и пермских образований. Напротив, юрские среднекислотные интрузивные породы соответствуют относительно высоким уровням радиоактивности. Благоприятные для минерализации аномалии не были оптимизированы в пределах целевого района (рис. 6).

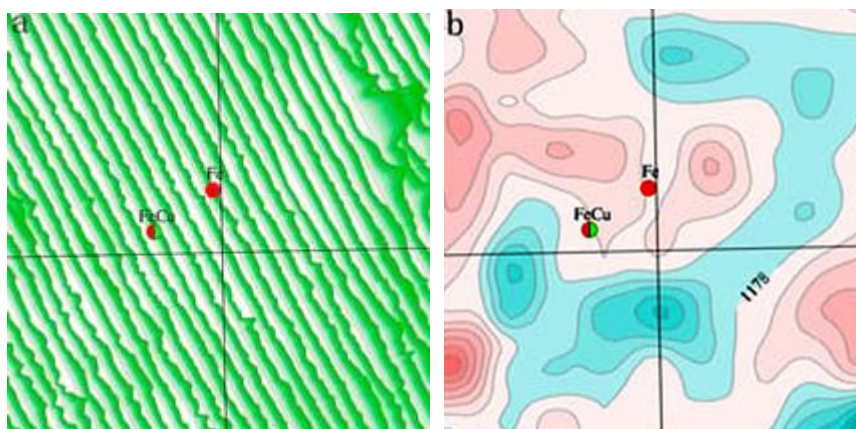


Рис. 6. Аэро-электромагнитная и радиометрическая карты.
(а) аэро-электромагнитные разрезы; (b) карта общей интенсивности.

Вблизи известных рудных месторождений имеются геохимические аномалии Cu, Pb, Au, Ag и As, причем Cu - зона высоких геохимических аномалий. Кроме того, в зоне аномалий с высоким содержанием Au в южной части целевого района также имеются низкокачественные геохимические аномалии Cu, Ag и Mo. Эти аномальные зоны служат целевыми районами поисков медно-полиметаллических месторождений и действуют как прямые признаки оруденения (рис. 7).

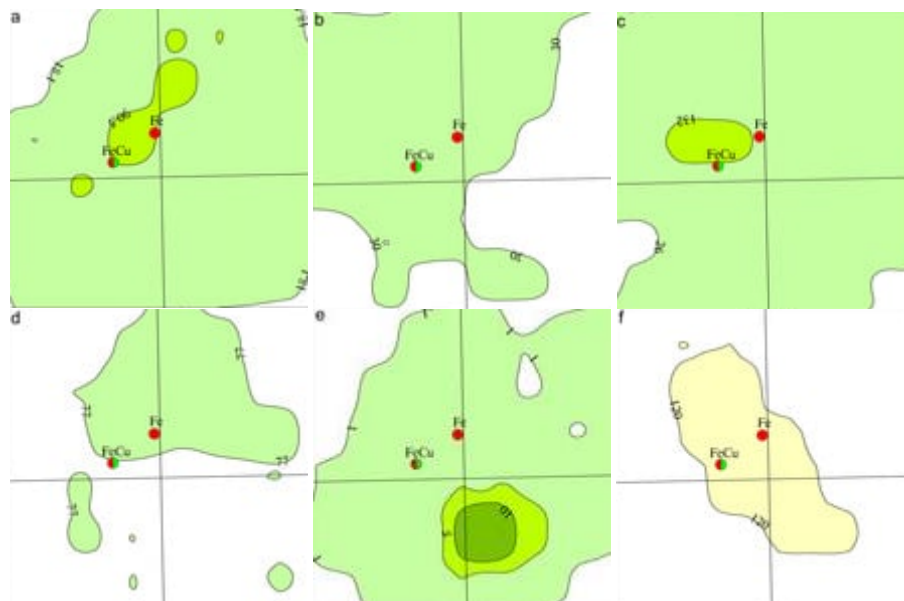


Рис. 7. Карта геохимических аномалий в районе м-ния Наонюшань.
(a) Cu; (b) Pb; (c) As; (d) Zn; (e) Au; (f) Ag.

Для вычисления суперпозиции информационной ценности каждого поискового показателя получена карта информационной (рис. 8).

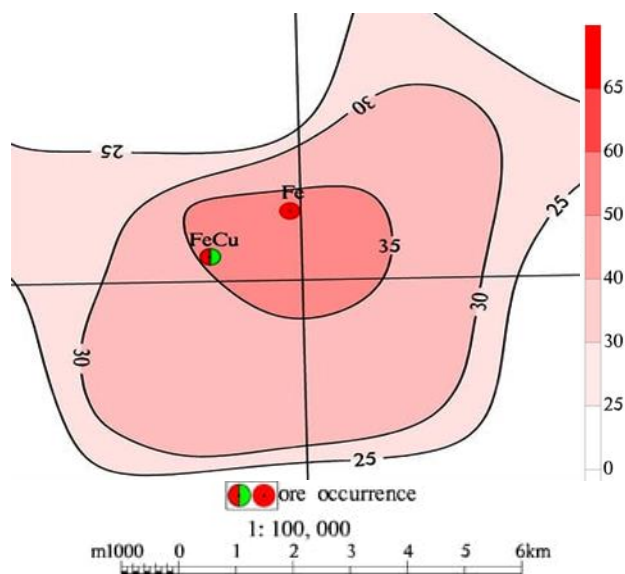


Рис. 8. Комплексный информационный показатель целевого участка поисков медно-полиметаллических руд.

Ценность информации колеблется от 25 до 50, и в соответствии с минимальным значением 35, определенным как пороговое значение для по всему региону, эта конкретная область может быть определена как целевая область поисков.

5. Выводы.

1. Построена система оценки целевых участков минерализации, основанная на принципах АНР. Были рассчитаны веса каждого разведочного показателя и сформулирована комплексная

модель оценки ценности геолого–аэрогеофизико-геохимической информации. Были разработаны различные методы взвешивания ценности информации в соответствии с характеристиками аэрогеофизических поисков. Результатом является карта, имеющая всеобъемлющую информационную ценность, обеспечивающая систематический метод оптимизации целевых зон минерализации и проведения объективных и всесторонних оценок.

2. Представлен пример применения системы оценки, построенной с использованием АНР для выбора целевых участков минерализации в рудном районе Наонюшань. Была сгенерирована карта информационного наполнения исследуемой. На основе значений информационного содержания, соответствующих известным рудным залежам в пределах района, было определено нижнее пороговое значение 35 для определения перспективных районов. Последующие полевые исследования подтвердили эффективность этого подхода, поскольку выявленные целевые участки точно соответствовали приоритетным зонам поисков. В общей сложности было выделено 19 перспективных участков для медно-полиметаллических месторождений гидротермального, порфирового, мезотермального и скарнового типов.

3. Метод АНР имеет ограничения, связанные с объемом данных и пониманием интерпретатора. Он также имеет определенные перспективы, поскольку геологические, геофизические и геохимические технологии продолжают развиваться и совершенствоваться, наряду с обогащением различной информацией и повышением точности измерений.

4. Представленный метод относится к прогнозированию перспективных районов в проектах рекогносцировки, использующих аэрогеофизические данные. При проведении ГРР на конкретные минеральные ресурсы с использованием конкретных методов и технологий иерархические модели необходимо разрабатывать отдельно, чтобы они соответствовали требованиям проекта ГРР, учитывая фактические обстоятельства.

IV. ПРОГНОЗИРОВАНИЕ Fe-Ti-P-RZM ОРУДЕНЕНИЯ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ (*рифт Осло, Норвегия*) [6]

1. Введение.

Редкоземельные элементы (РЗЭ) и фосфор (Р) включены Европейским союзом в список важнейших сырьевых материалов, что подчеркивает их важность как стратегических ресурсов на будущее. РЗЭ находятся на первом месте в списке приоритетов из-за их использования в различных технологиях зеленой энергетики, в то время как Р играет решающую роль в индустрии удобрений.

В этом контексте геологические условия Норвегии демонстрируют большой потенциал для обнаружения и последующей эксплуатации важнейших видов сырья, особенно ресурсов РЗЭ. В Норвегии были выявлены значительные ресурсы Fe-Ti-P-REE, которые связаны со слоистыми интрузиями и щелочными интрузивными комплексами, что подчеркивает потенциал открытий в различных условиях. Южный рифтовый регион Осло известен наличием множества магматических проявлений, богатых Fe-Ti-P-REE, при этом месторождение Кодал является крупнейшим и наиболее известным на сегодняшний день. Однако, поскольку большая часть этих ресурсов образуется как часть магматических систем, поиск новых открытий часто затруднен отсутствием обширных ореолов изменений или дистальных следов, как, например, наблюдается в гидротермальных системах. Таким образом, разработка альтернативных инструментов поисков имеет решающее значение для поддержки новых будущих открытий.

Машинное обучение (ML) превратилось в мощный инструмент в области картирования потенциала полезных ископаемых. Одна из основных сильных сторон ML в этом контексте заключается в ее способности эффективно обрабатывать обширные и сложные многомерные наборы данных, что является существенным улучшением по сравнению с традиционными методами. Кроме того, ML позволяет объективно интегрировать данные, уменьшая потенциальные искажения, которые могут возникнуть из-за опыта человека. Однако существуют также потенциальные ограничения, связанные с этим подходом, основанным на данных. При поисках ПИ эффективное обучение моделей может быть сложной задачей из-за ограниченных и предвзятых данных. Как правило, отрицательных случаев (неминерализованных) больше, чем положительных (минерализованных), просто из-за редкости минеральных ресурсов, и доступные данные могут быть неполными. Например, могут быть доступны только подтвержденные положительные образцы (известная минерализация), а отрицательные образцы (породы, лишенные минерализации) должны быть выбраны случайным образом и/или соответствовать конкретным эмпирическим рекомендациям из

интересующей области. Эти факторы могут повлиять на точность и надежность моделей, используемых при поисках. Модели машинного обучения часто рассматриваются как ‘черные ящики’, что означает, что может быть трудно понять, как алгоритмы получают свои прогнозы на основе многовариантных характеристик или как эти функции связаны с лежащими в основе геологическими процессами, которые контролируют минерализацию. В этом материале учитываются эти ограничения и объединяется наука о данных с опытом в геологическом прогнозировании.

Исследование посвящено использованию ML для определения потенциальных областей для ранней поисковой стадии ресурсов Fe-Ti-P-REE в районе Южного рифта Осло. Существующие пространственные наборы данных высокого разрешения, включая аэромагнитные, радиометрические и топографические карты, использовались для генерации прогнозирующих характеристик, а образцы горных пород с геохимическим анализом были обработаны в обучающие образцы. Для генерации прогностической модели были оценены два широко используемых классификатора: случайный лес (RF), основанный на разделении, и машина опорных векторов (SVM), основанная на непрерывности. Оба классификатора были реализованы с модификациями для обработки несбалансированных наборов данных (т. е. отрицательных выборок значительно больше, чем положительных) путем корректировки весов классов на основе частот классов и корректировки матрицы затрат, чтобы более жестко «наказывать» за отнесение выборок к чрезмерно представленному классу (отрицательные, неминерализованные). Модель классификации RF была выбрана из-за ее превосходной производительности, а сгенерированная прогнозная карта была сравнена с устаревшими данными и проверена в полевых условиях. Было доказано, что это ценно с точки зрения сужения областей, требующих дальнейшего изучения для поисков ПИ.

2. Геологические предпосылки.

Район исследования ($\sim 2000 \text{ км}^2$) расположен в рифте Осло в ЮВ части Норвегии, (рис. 1).

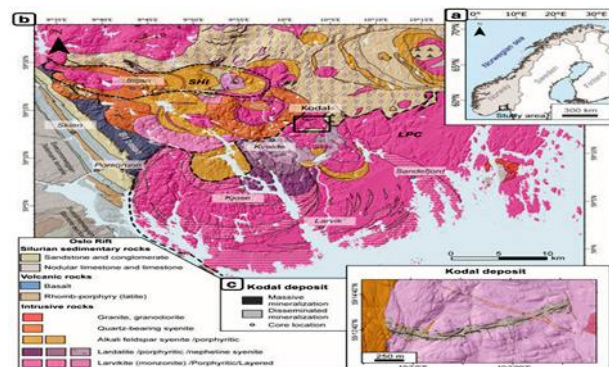


Рис. 1. Геологическая карта исследуемого района.

Магматическая активность в рифте началась с залегания базальтов B1. За этим последовало размещение объемных потоков трахитовой и трахиандезитовой лавы, в которую вторгся plutонический комплекс Ларвик (LPC) между 296 и 289 млн лет назад и более молодые интрузии от монцонитовых до кварцевых сиенитовых. Известны несколько рудных залежей, богатых Fe-Ti-P-REE (рис. 2), которые связаны с магматизмом среднего и кислого уровней, включая месторождение Fe-Ti-P-REE в Кодале (рис. 1с).

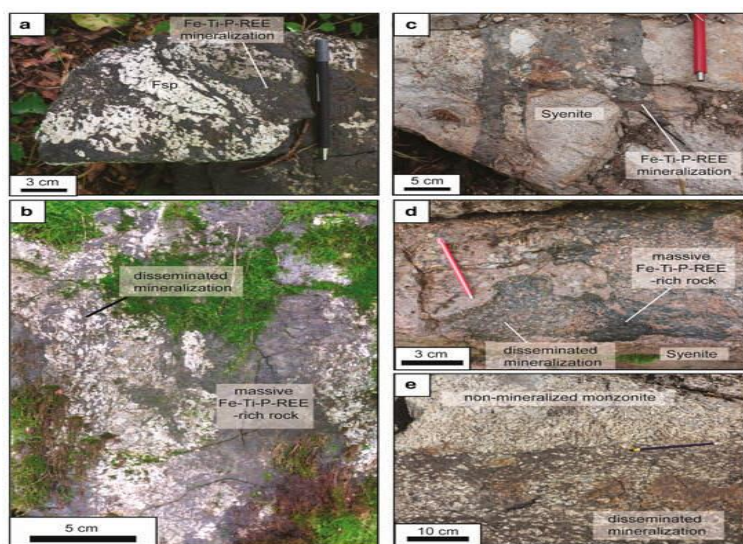


Рис. 2. Примеры минерализации обучающих образцов.
(a,b) Fe-Ti-P-REE на м-нии Кодал; (c,d) Fe-Ti-P-REE минерализация интрузий Сильян-Хвернес (SHI);
(e) Fe-Ti-P-REE минерализация в Кьосе (LPC).

Месторождение Кодал представляет собой крупнейшее известное месторождение Fe-Ti-P-REE минерализации в LPC, с общими обозначенными ресурсами 14,6 млн тонн при содержании 5,18% P_2O_5 и 24,12% Fe. Обнажена западная часть месторождения, где можно наблюдать породы, богатые титаномагнетитом, апатитом и ильменитом (основная минерализация) (рис. 2a, b). Граница между вмещающей породой и рудой определяется концентрациями Fe_2O_3 , TiO_2 и P_2O_5 .

На основе результатов предыдущих региональных исследований и известного оруденения м-ния Кодал, можно определить критические параметры, характеризующие Fe-Ti-P-REE минерализацию в регионе. Из-за обилия магнетита минерализация характеризуется положительными магнитными аномалиями (1000-3000 нТл), тогда как низкое содержание минералов, богатых калием (K), приводит к более низким концентрациям K (обычно ниже 2%) по сравнению с вмещающими породами. Более того, минерализация, по-видимому, быстрее размывается по сравнению с вмещающими породами, таким образом, присутствуя в топографических впадинах второстепенного значения. Эти контрастные петрофизические свойства руды по сравнению с вмещающей породой делают этот тип минерализации уязвимым

для обнаружения с помощью различных классических инструментов геофизической разведки, таких как радиометрическая и магнитная съемки.

3. Данные.

В этом исследовании используются геопространственные наборы данных высокого разрешения, включая аэрогеофизику и цифровую модель рельефа (DEM), для создания набора функций в качестве входных данных для модели ML. Для подготовки геопространственных наборов данных расширили охват за пределы исследуемой области, как показано на рисунке 3d.

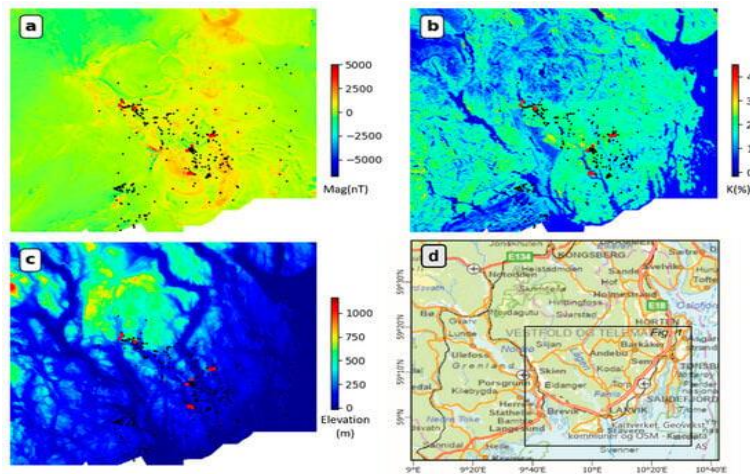


Рис. 3. Наборы данных высокого разрешения, спроецированные на однородную сетку с размером ячейки 50 м: (a) аномалия общего магнитного поля (nT); (b) концентрация калия (K) (%); (c) топография (DEM) (m); (d) географическая карта; (a–c) местоположения полевых образцов: минерализованные - красные точки, неминерализованные - черные.

Магнитные данные отображают общую аномалию магнитного поля (рис. 3a), радиометрические данные указывают на обилие калия (K), тория (Th) и урана (U). Пример обработанных K-данных в весовых процентах представлен на рисунке 3b. Данные ЦМР с пространственным разрешением 50 м – на рисунке 3c). Чтобы обеспечить пространственное выравнивание объектов, использованных при разработке прогнозной модели, все наборы данных были спроецированы на однородную сетку с размером ячейки 50 м.

Образцы горных пород были собраны из различных литологических единиц, присутствующих в регионе, включая области минерализации.

Геохимические данные, состоящие из анализа ICP-MS полевых образцов, были использованы для измерения концентрации интересующих элементов в породах. Было проведено разделение образцов на положительные и отрицательные обучающие наборы данных с использованием геохимических и минералогических индикаторов. Принимая во внимание значительное содержание апатита в минерализованных образцах, классифицировали

полевые образцы на основе концентрации P_2O_5 . Образцы с содержанием P_2O_5 , превышающим 4 мас.%, были классифицированы как минерализованные, тогда как образцы с содержанием P_2O_5 ниже 4 мас.% были классифицированы как неминерализованные. Эти классифицированные образцы были использованы в качестве основы для разработки обучающего набора данных для прогностической модели. Из собранных образцов 23 были классифицированы как минерализованные, тогда как 406 были классифицированы как неминерализованные. Распределение классифицированных образцов отмечено на рисунке 3а поверх геопространственных карт.

4. Методология и рабочий процесс.

4.1. Стратегия выборки обучающих данных.

Рисунки 3а–с иллюстрируют пространственное распределение собранных образцов, обеспечивает обширный охват и значительную долю подтвержденных отрицательных образцов. Это устраняет необходимость в создании синтетических отрицательных образцов с использованием таких методов, как SMOTE или слепой выборки. В образцах наблюдается заметный дисбаланс: соотношение минерализованных образцов (т.е. положительных) и неминерализованных образцов (т.е. отрицательных) составляет 23:406 ($\approx 1:18$). Этот дисбаланс представляет проблему при обучении модели классификации, поскольку модель, как правило, лучше обучается прогнозированию класса большинства. Однако для составления прогнозных карт более важно правильно определить класс миноритариев (т. е. минерализованные образцы). Для решения этой проблемы разработана стратегия выборки для создания обучающего набора, учитывающего несбалансированные данные.

Был выбран 50-метровый пространственный диапазон для неминерализованных образцов, которые будут отнесены к отрицательным данным обучения, в то время как больший пространственный диапазон в 100 м был выбран для минерализованных образцов, которые будут отнесены к положительным. Пример этого показан на рисунке 4.

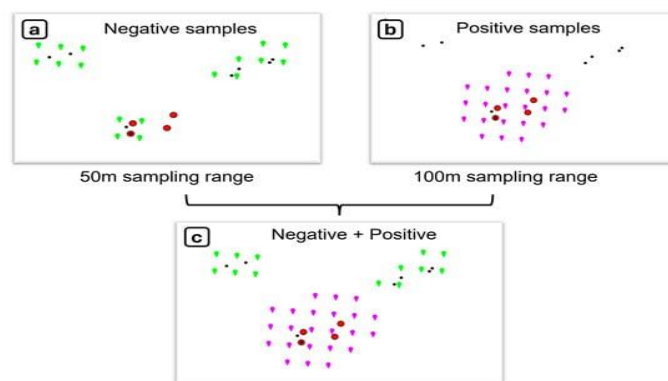


Рис. 4. Иллюстрация стратегии выборки обучающих данных.

На рисунке 4а ячейки сетки на расстоянии 50 м от неминерализованных образцов выделены зеленым цветом, на рисунке 4б ячейки сетки на расстоянии 100 м от минерализованных образцов выделены пурпурным цветом. В случаях перекрывающихся ячеек сетки положительные образцы преобладают над отрицательными образцами, как показано на рисунке 4с. Истинное местоположение неминерализованных (т.е. отрицательных) и минерализованных (т.е. положительных) образцов указано черными и красными кружками соответственно. Зеленые и пурпурные указатели соответствуют равномерной сетке (с размером ячейки 50 м), на которую проецируются пространственные наборы данных высокого разрешения. В частности, все зеленые указатели расположены на расстоянии 50 м по крайней мере от одного неминерализованного полевого образца (**а**); тогда как все пурпурные указатели расположены на расстоянии 100 м по крайней мере от одного минерализованного образца (**б**); в случаях перекрытия ячеек сетки положительные образцы преобладают над отрицательными образцами, как показано на (**с**).

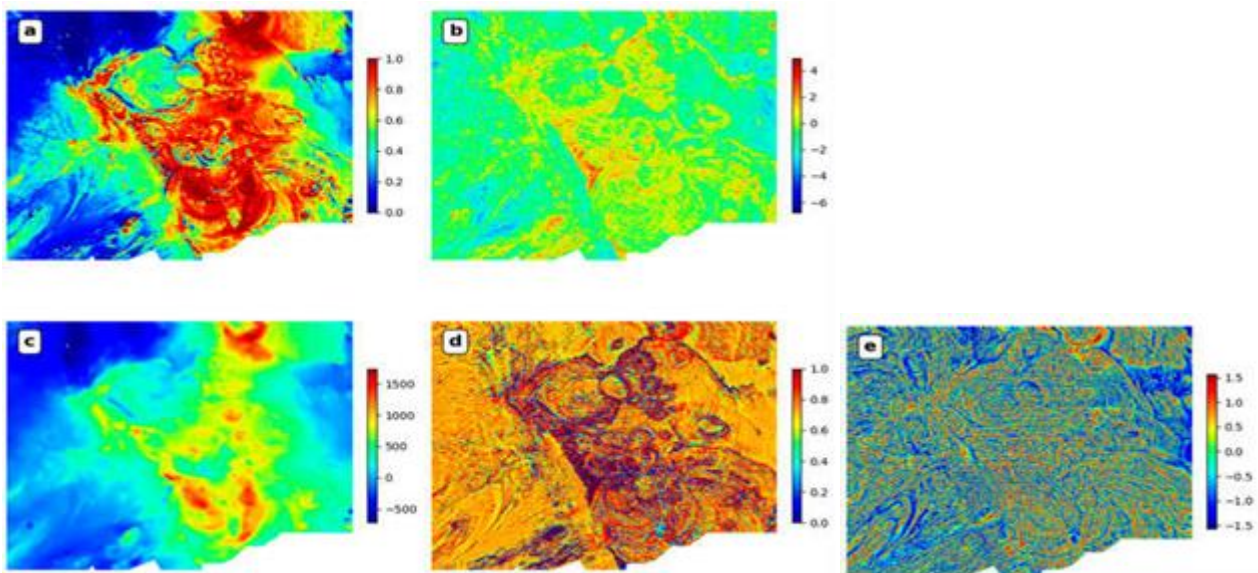
Выбор размеров пространственного диапазона довольно консервативен, поскольку он определяет относительно небольшие пространственные размеры образцов. Этот подход направлен на использование всего потенциала функций высокого разрешения для обеспечения более тонкой пространственной дифференциации между минерализованными и неминерализованными образцами. Эта схема отбора проб также разумна в контексте ГРР, потому что, если минерализованные и неминерализованные образцы будут найдены близко друг к другу в данном местоположении, местоположение все равно будет рассматриваться как потенциально способное содержать большой объем минерализованных пород (т. е. положительный показатель перспективности).

В окончательном обучающем наборе данных соотношение положительных (минерализованных) и отрицательных (неминерализованных) образцов составляет 174: 997 ($\approx 1:5,7$), что по-прежнему несбалансировано, но значительно улучшено по сравнению с первоначальным соотношением.

4.2. Разработка объектов.

Разработка объектов - важнейший процесс ML, включающий извлечение релевантных объектов из необработанных данных для повышения производительности модели. В этом исследовании были предприняты совместные усилия ученых-геологов, обладающих опытом в геофизике, геологии и минералогии, для выявления важных особенностей из трех геопространственных наборов данных с сеткой высокого разрешения (рис. 3), которые, вероятно, связаны с целевой минерализацией.

Для магнитных данных применили несколько методов фильтрации для создания объектов, которые обычно используются экспертами-людьми для облегчения интерпретации карт. Например, продолжение вверх сохраняет низкочастотную составляющую магнитных данных и подчеркивает более глубокие источники магнетизма; аналитический сигнал выделяет высокочастотные сигналы, поступающие из неглубоких источников; а различные фильтры на основе производных, такие как производные по вертикали и наклону, увеличивают границы магнитных аномалий. При необходимости выравнивание гистограммы или логарифмическое преобразование использовались для корректировки распределения данных с целью улучшения способности различать признаки. Визуальным эффектом этого является усиление глобального контраста при просмотре изображения. Например, на рисунке 3а представлены исходные магнитные данные в линейной цветовой гамме, тогда как на рисунке 5а показана версия с выравниванием по гистограмме, раскрывающая гораздо больше деталей исходного изображения. В общей сложности были сгенерированы пять объектов из исходной магнитной карты в качестве потенциальных объектов для ввода в модели машинного обучения. Эти объекты показаны на рисунке 5: (а) магнитные данные с выравниванием по гистограмме (mag_ft), (б) аналитический сигнал с логарифмическим преобразованием (mag_as_ft), (в) продолжение вверх на 500 м (mag_uc500), (г) производную по вертикали с выравниванием по гистограмме (mag_vd_ft) и (д) производную по наклону (mag_tdr) магнитных данных.



Ри. 5. Прогностические характеристики, соответствующие аэро-магнитным.
(a) mag_ft; (b) mag_as_ft; (c) mag_uc500; (d) mag_vd_ft; (e) mag_tdr.

Цветные полосы, связанные с графиками объектов, указывают относительный диапазон данных и интенсивность, но физические единицы измерения не указаны. Это связано с тем, что

в процессе проектирования объектов мог измениться масштаб или диапазон исходных данных, в результате чего физические единицы измерения потеряли смысл.

Загруженные продукты радиометрических данных представляют калий (K) в весовых процентах (%), а также уран (U) и торий (Th) в эквивалентных концентрациях исходных нуклидов, обозначаемых как эквивалентный U (eU в ppm) и эквивалентный Th (eTh в ppm) соответственно. Оценили отдельные каналы радиоактивных элементов, а именно K, eTh и eU, на предмет их абсолютных концентраций на поверхности и соотношений, т. е. eTh/K, eU/K и eU/eTh, которые указывают на их относительные концентрации. Операция ratio часто усиливает каналы с низкой скоростью подсчета по сравнению с другими каналами. В результате могут быть выделены аномальные области, в отличие от случаев, когда отдельные каналы обычно коррелируют в большинстве типов горных пород. На основе исходных радиометрических данных были сгенерированы шесть характеристик. Тремя каналами отдельных элементов, выровненными по гистограмме, были K_ft, Th_ft и U_ft, показанные на рисунке 6а–с соответственно.

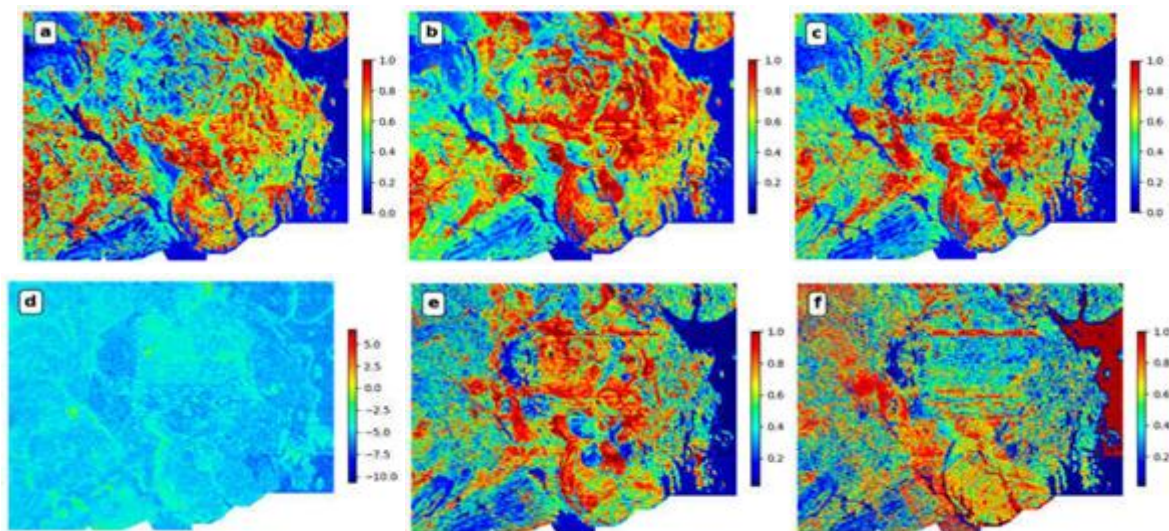


Рис. 6. Прогностические характеристики аэро-радиометрических.
(a) K_ft; (b) Th_ft; (c) U_ft; (d) Th_K_as_ft; (e) U_K_ft; (f) U_Th_ft.

Остальные три характеристики представляли собой соотношения, обозначенные Th_K_as_ft, U_K_ft и U_Th_ft, показанные на рисунке 6d–f соответственно. Для генерации функции Th_K_as_ft применили логарифмическое преобразование к аналитическому сигналу отношения eTh/K. Функции U_K_ft и U_Th_ft были созданы путем добавления небольшой константы к знаменателю, чтобы избежать деления на ноль, а затем выполнения выравнивания гистограммы. На решение использовать аналитический сигнал отношения Th/K повлияли эксперты предметной области, которые обнаружили, что он лучше объясняет обнаруженные проявления минерализации, чем само соотношение.

Цветные полосы, связанные с графиками объектов, указывают относительный диапазон данных и интенсивность, но физические единицы измерения не указаны. Это связано с тем, что в процессе проектирования объектов мог измениться масштаб или диапазон исходных данных, в результате чего физические единицы измерения потеряли смысл.

Сгенерировали пять прогностических характеристик на основе топографических данных. На рисунке 7а показана выровненная по гистограмме DEM (Топо_ft), тогда как на рисунке 7b,c показаны выровненные по гистограмме вертикальная производная (Топо_vd_ft) и наклон (Топо_slope_ft) данных DEM соответственно.

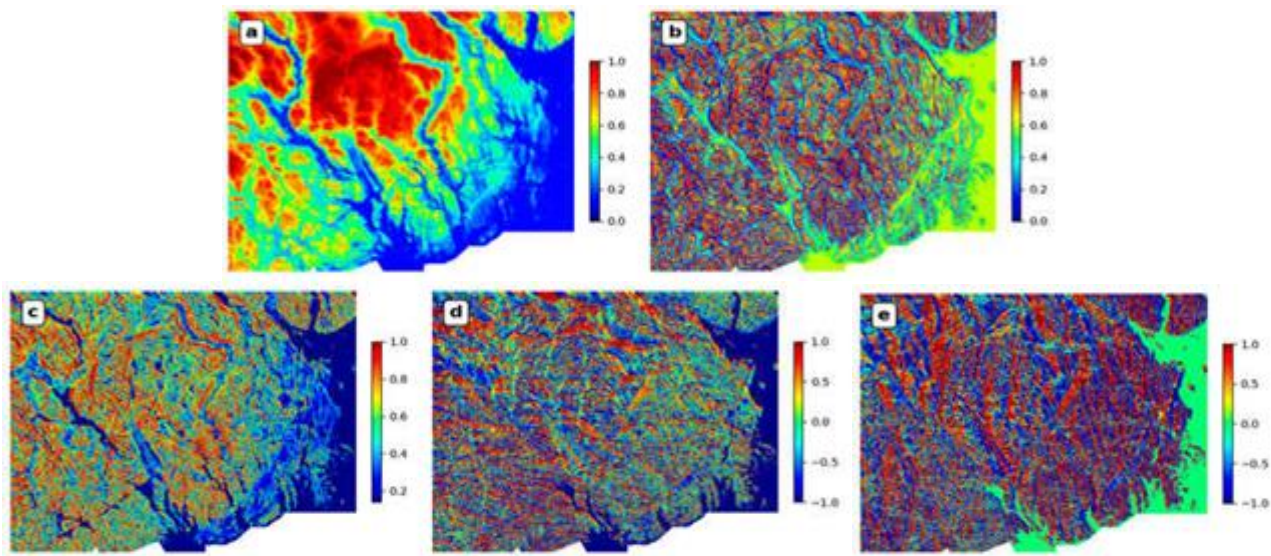


Рис. 7. Прогностические характеристики топографических данных.
(a) Топо_ft; (b) Топо_vd_ft; (c) Топо_slope_ft; (d) Топо_aspect_sine_ft; (e) Топо_aspect_cosine_ft.

Также был рассчитан топографический аспект, представляющий направление склонов. Однако значения аспекта варьируются от 0° до 360° и имеют круговой характер, что означает, что минимальный и максимальный аспекты 0° и 360° по существу одинаковы (строго на север). Чтобы получить более интерпретируемое представление направления по линейным осям, разделили аспект на две составляющие: синусоидальную и косинусоидальную, показанные на рисунке 7d, e соответственно. Синусоидальная составляющая находится в диапазоне от 1 до -1, представляя склоны, обращенные с севера на юг, где 1 указывает на уклоны, обращенные к северу, а -1 указывает на уклоны, обращенные к югу. Косинусный компонент представляет склоны, обращенные с востока на запад, и имеет ту же интерпретацию.

Цветные полосы, связанные с графиками объектов, указывают относительный диапазон данных и силу, но физические единицы измерения не указаны. Это связано с тем, что в процессе проектирования объектов мог измениться масштаб или диапазон исходных данных, в результате чего физические единицы измерения потеряли смысл.

4.3. Выбор объекта.

Была сгенерирована 2D матрица корреляции всех потенциальных объектов (всего 16), используя коэффициент корреляции Пирсона (рис. 8).

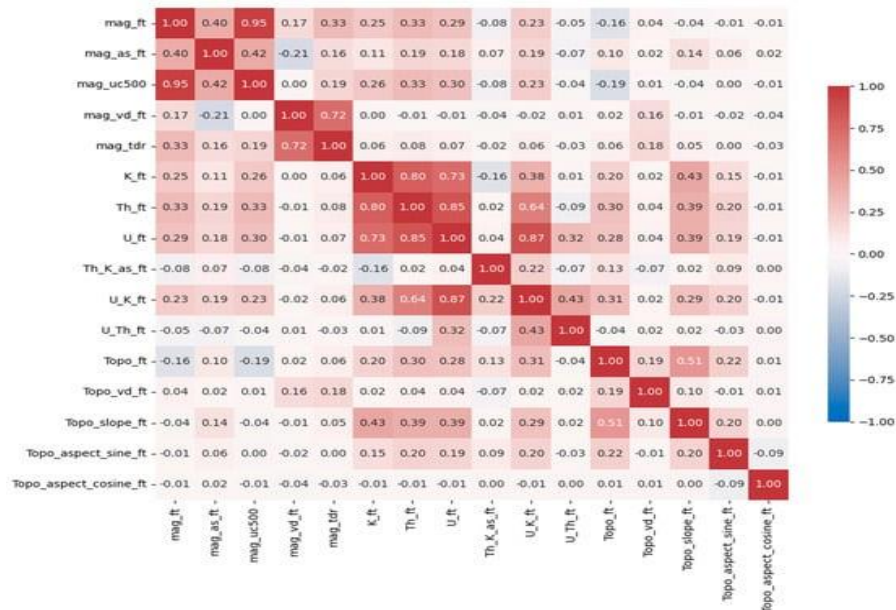


Рис. 8. Матрица корреляции между 16 потенциальными объектами, 5 магнитных объектов, 6 радиометрических и 5 топографических.

Матрица содержит значения в диапазоне от -1 (идеальная отрицательная корреляция) до +1 (идеальная положительная корреляция), при этом 0 указывает на отсутствие корреляции. Чтобы избежать чрезмерной представленности и избыточных предикторов, выбрали только независимые и некоррелированные признаки для построения моделей классификации с целью повышения производительности модели. Установили пороговое значение 0,7 для коэффициента корреляции и отбросили четыре объекта (mag_ft, mag_tdr, Th_ft и U_ft), которые демонстрировали сильную корреляцию с другими объектами, которые можно было легко идентифицировать по матрице в регионах с более темными красными цветами. 12 выбранных объектов включают 3 магнитных объекта (mag_as_ft, mag_uc500 и mag_vd_ft), 4 радиометрических (K_ft, Th_K_as_ft, U_K_ft и U_Th_ft) и 5 топографических (Topo_ft, Topo_vd_ft, Topo_slope_ft, Topo_aspect_sine_ft и Topo_aspect_cosine_ft).

Коэффициенты корреляции варьируются от -1 до +1, где -1 указывает на 100% отрицательную корреляцию, +1 указывает на 100% положительную корреляцию, а значение 0 указывает на отсутствие корреляции.

4.4. Модель классификации.

После создания обучающего набора данных, состоящего из 174 образцов, помеченных как минерализованные, и 997 образцов, помеченных как неминерализованные, каждый из которых поддерживается 12-мерным вектором признаков, была подготовлена модель классификации.

Двадцать процентов образцов (235) были отложены для тестирования, а остальные 80% (936) были использованы для обучения и валидации. Пространственно-стратифицированный отбор проб использовался для обеспечения того, чтобы образцы, находящиеся в непосредственной близости, не были разделены между обучающими и тестовыми наборами. Соотношение минерализованных и неминерализованных образцов было сохранено в обоих наборах с использованием стратифицированной случайной выборки.

В принципе, все хорошо зарекомендовавшие себя методы классификации потенциально могут точно решать задачи классификации, но они различаются своими предположениями о данных и гипотезами моделирования. Например, деревья принятия решений, используемые в случайном лесу (RF), разделяют прогностические объекты на дискретные, неперекрывающиеся группы, представленные листовыми узлами дерева, в то время как машина опорных векторов (SVM) использует гиперплоскость для разделения данных на разные классы, что позволяет гибко моделировать нелинейные взаимосвязи. Оба метода эффективны для моделирования сложных данных, и их соответствующие подходы одинаково приемлемы для исследования. Поэтому протестировали RF и SVM в задаче классификации. Использовали пакет `scikit-learn` Python (версия 1.2.2) для реализации алгоритмов RF и SVM. Оба классификатора были реализованы с модификациями для обработки несбалансированных наборов данных путем корректировки весов классов обратно пропорционально частотам классов и корректировки матрицы затрат, чтобы более строго «наказывать» за отнесение образцов к перепредставленному классу (образцы без минерализации).

Для обеих моделей провели последовательные обучающие тесты, настроив пространство параметров в диапазоне значений. Для RF изменили количество деревьев и минимальный размер конечного узла, в то время как для SVM протестировали различные функции ядра и настроили параметр регуляризации. Средняя точность и оценка ML использовались для оценки моделей в рамках 5-кратной схемы перекрестной проверки. При оптимальных настройках параметров модели RF модель достигла средней точности 0,949 и оценки ML 0,815, в то время как модель SVM дала среднюю точность 0,775 и оценку ML 0,538. Кроме того, оценили производительность модели-победителя, RF-модели, на невидимых данных, то есть на наборе тестов, который отложили, и он достиг общей точности 0,962. В частности, проверили матрицу путаницы, которая дала точность классификации 0,900 для положительных (минерализованных) образцов и 0,971 для отрицательных (неминерализованных) образцов. Также включили кривую рабочих характеристик приемника (ROC) радиочастотной модели, чтобы продемонстрировать ее дискриминационные характеристики и подчеркнуть ее способность различать положительные и отрицательные примеры в анализе. Это графическое

представление служит дополнительным наглядным пособием, подчеркивающим эффективность RF-модели в контексте данного исследования. Этот результат дал уверенность в производительности RF-модели в отношении подгонки и обработки несбалансированного набора данных. Поэтому была выбрана обученная RF-модель для прогнозирования минерального потенциала всего исследуемого региона.

5. Результаты и обсуждение.

Карта перспективности исследуемого региона на P_2O_5 была сгенерирована с использованием обученной RF-модели (рис. 9).

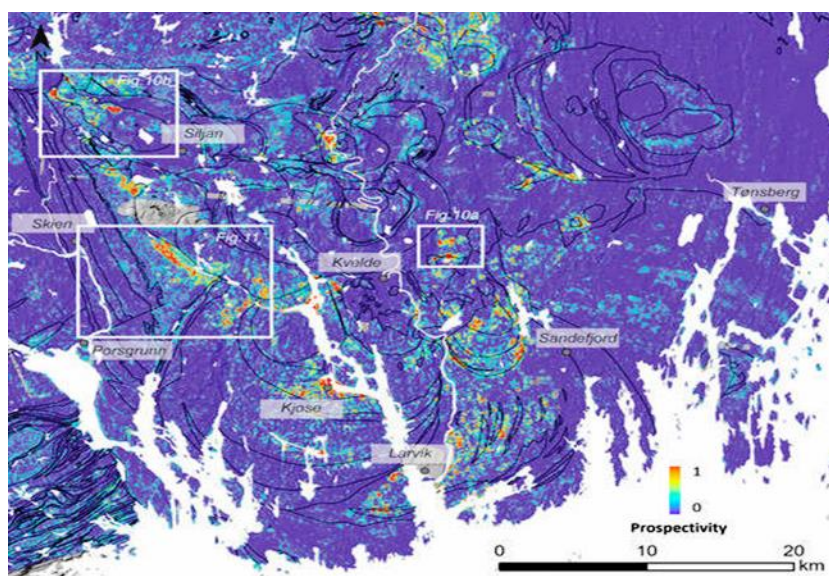


Рис. 9. Карта перспективности P_2O_5 .

Область карты содержит 2 668 931 ячейку сетки размером 50×50 м. Каждая ячейка сетки, называемая образцом, связана с 12-мерным вектором признаков и была классифицирована обученной RF-моделью. Вместо использования прогнозируемых бинарных классов (минерализованных или неминерализованных) непосредственно для каждого образца, выведены прогнозируемые вероятности принадлежности образца к минерализованному классу. Эта вероятность измеряет уверенность модели в ее прогнозах относительно принадлежности образца к минерализованному классу и, следовательно, отражает перспективность нахождения целевого минерализационного слоя.

Значения функций перестановки вычисляются для оценки важности функций в обученной RF-модели. Эти значения определяются количественно путем оценки того, насколько увеличивается ошибка прогнозирования модели, когда значения конкретного предиктора случайным образом перетасовываются или переставляются местами. На рисунке 10 показана важность признаков в порядке убывания, выделяя наиболее влиятельные предикторы. В целом,

магнитные характеристики и содержания калий (K) являются наиболее важными предикторами для модели, в то время как топографические особенности оказывают меньшее влияние. Это согласуется со знаниями предметной.

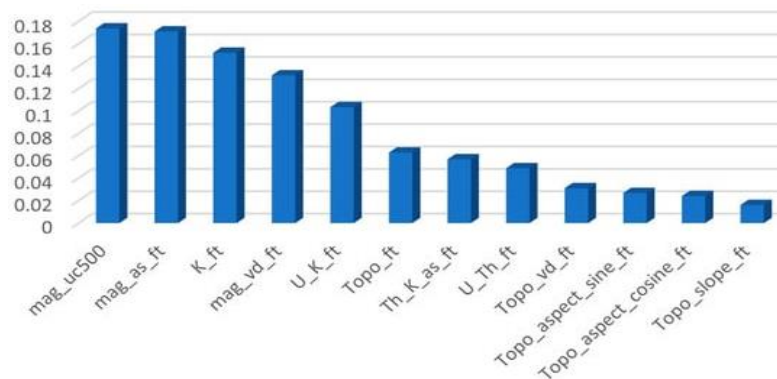


Рис. 10. Важность признаков для модели RF.

На рисунке 9 выделены три высокоперспективных района (желтым или красным) для минерализации Fe-Ti-P-REE, они составляют около 0,3% от всего региона.

5.1. Область 1: Kodal.

В дополнение к основному району, связанному с рудным месторождением Кодал, выявлен второй район непосредственно к югу и примерно параллельно известному месторождению (рис. 11).

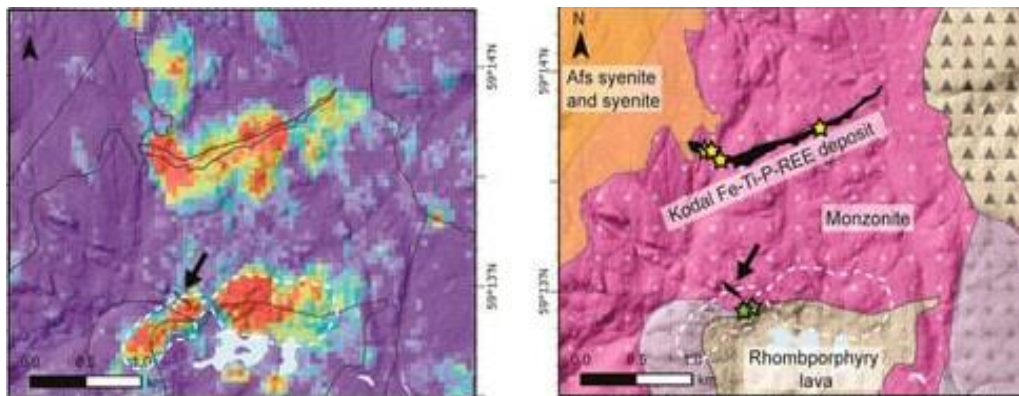


Рис. 11. Сравнение карты перспективности (слева) и геологической карты (справа) в Кодале.

Геологическое картирование, проведенное Kodal Minerals, привело к обнаружению выступа в этой области, как указано черной стрелкой на рисунке 11. В рамках этого исследования провели полевую заверку в этом высокоперспективном районе и обнаружили рудную минерализацию на поверхности. Такое соответствие между прогнозом и геологической заверкой усиливает ценность карты перспективности как надежного инструмента для сужения целевых поисковых участков. Это усиливает аргументы в пользу проведения детальных ГРП для получения существенных результатов.

Желтыми звездочками отмечены местоположения образцов, считающихся минерализованными, которые использовались в качестве положительных обучающих образцов в этом исследовании. Зелеными звездочками отмечены местоположения, где во время полевой проверки были обнаружены породы, богатые Fe-Ti-P.

5.2. Область 2: Сильян.

Карта перспективности района Сильян показывает идентичность трех основных областей, представляющих интерес, обозначенных белыми пунктирными линиями (рис. 12).

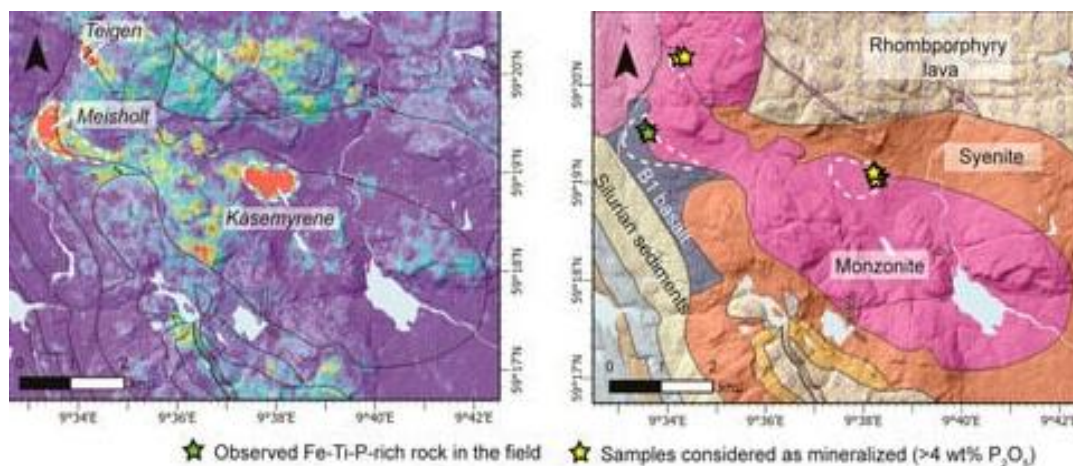


Рис. 12. Сравнение карты перспективности (слева) и геологической карты (справа) в Сильяне.

Хотя крупных известных минерализаций, непосредственно связанных с сильянской интрузией, нет, наличие нескольких участков, где выходят на поверхность породы, богатые Fe-Ti-P-REE, указывает на то, что в этом районе существуют условия для формирования промышленного оруденения.

В районе Косемирене (центральная область обведена кружком на рисунке 12b) карта перспективности предполагает, что минерализация может выходить за пределы известных в настоящее время масштабов. Это важно, поскольку потенциально может привести к открытию более крупного рудного тела поблизости.

Высокая перспективность района Teigen, связана с наличием двух минерализованных образцов в обучающем наборе данных. Однако карта не прогнозирует наличие более крупной минерализации. Эта информация имеет решающее значение для планирования ГРП, поскольку помогает сосредоточить усилия на областях с более высокой вероятностью значительных находок.

Район Мейсхольт (юго-запад на рисунке 12b) классифицируется как высокоперспективный. Согласно геологической карте, его часть залегает в базальтах. Предыдущие ГРП на контакте базальта и сиенита подтвердили наличие минерализованных зон, хотя низкое содержание в них Р, привело к исключению их из объектов ГРП. В этом районе

необходимо дополнительное картографирование для точной оценки местоположения геологических единиц и определения характера минерализации.

5.3. Область 3: контакт между В1-базальтом и монцонит/сиенитом.

На рисунке 13 показана прогнозируемая перспективная область минерализации Fe-Ti-P-REE, обведенная черным пунктиром.

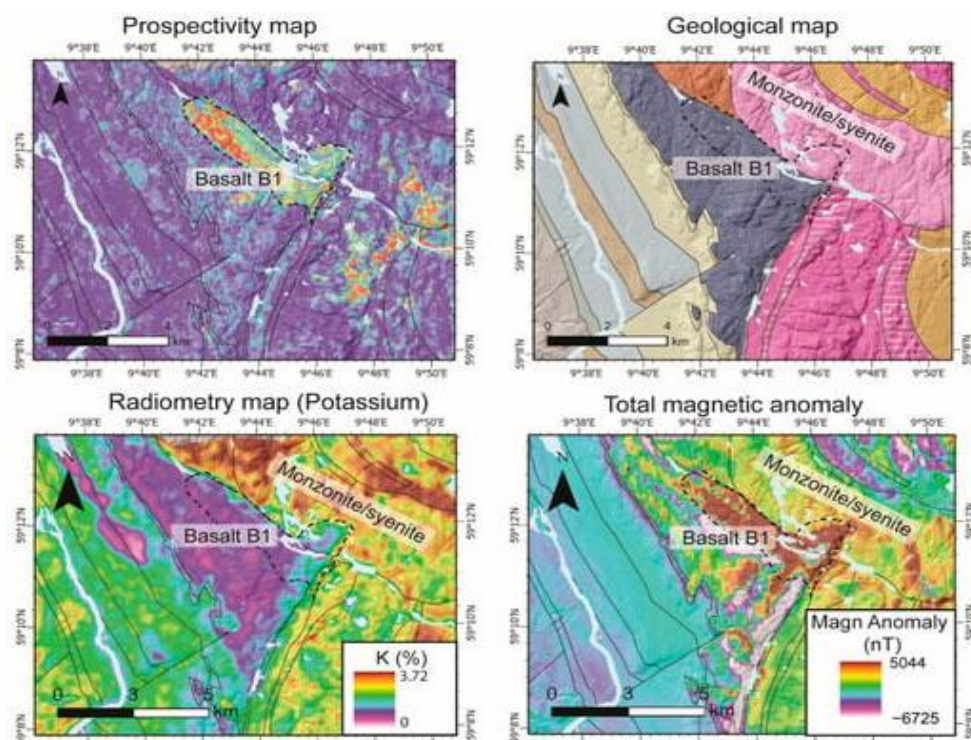


Рис. 13. Сравнение карты перспективности В1-базальта с геологической картой, картой калия (К) и общей картой магнитных аномалий.

В этом районе преобладают базальты, а не монцониты или сиениты. Базальт был сформирован во время пермского рифтогенеза. Fe-Ti-P-REE минерализация не связана с этой литологией. Есть два возможных объяснения того, что этот район обозначен как перспективный. Первое заключается в том, что местоположение границы между базальтом и монцонитом/сиенитом на геологической карте неточно. Хотя геологические границы иногда могут быть неточными из-за разного качества картографирования, маловероятно, что проблема заключается исключительно в качестве картографирования в данном случае. Базальт - это тип породы, который обеднен калием (К) по сравнению с окружающими типами пород, и особенно монцонитом и сиенитами, содержащими целевую минерализацию. Региональная радиометрическая съемка калия показывает, что базальтовый полигон на геологической карте хорошо согласуется с радиометрическими данными, за исключением небольшого участка, где в сиените, по-видимому, присутствуют низкие значения К. Вторая возможность заключается в том, что модель ошибочно принимает базальт за минерализацию Fe-Ti-P-REE. Поскольку

базальт обладает некоторыми петрофизическими (высокомагнитными) и геохимическими (с низким содержанием K) свойствами, аналогичными минерализации, эти параметры играют важную роль при построении карты перспективности (анализ важности признаков на рис. 10), часть базальтовых участков была классифицирована как имеющие потенциальную минерализацию. Это подчеркивает важность использования доступной геологической информации для поддержки интерпретации прогнозных карт.

Область с более низкой перспективностью, обозначенная белым пунктиром, тем не менее интересна. Она расположена среди пород сиенитового и монцонитового типов. Результаты радиометрии показывают, что, хотя калий не имеет четкой характеристики базальта, в этом районе содержится не так много калия, как в окружающих монцоните и сиените. Сильная магнитная аномалия связана с пегматитовыми сиенитами, содержащими крупные кристаллы магнетита. Эти пегматиты содержат мало фосфора, но имеют относительно высокую концентрацию REEs, что указывает на то, что карта перспективности также содержит информацию, которая может выделить минерализации, которые не являются точно намеченными, но обладают схожими свойствами. Это сопоставимо со случаем Мейсхольта в Сильяне, где концентрация P_2O_5 не достигает 4 мас. %.

Методы машинного обучения в значительной степени зависят от измеренных данных, особенно поверхностных сигнатур, и их связи с наблюдаемой минерализацией. Однако минерализация является результатом сложных процессов, требующих углубленного изучения для получения детального понимания. Тем не менее, эффективное применение методов машинного обучения к доступным картам с высоким пространственным разрешением дает ценную информацию для руководства дальнейшими исследованиями. В частности, поиски в исследуемом районе является сложной задачей из-за обширного растительного покрова. Геофизические и геохимические данные экономически эффективны с точки зрения сбора крупномасштабных пространственных характеристик.

Для успешного составления карт перспективности с использованием машинного обучения необходимо знать о природе методов, основанных на данных, и подчеркивать глубокое геологическое понимание. Тщательная оценка соответствующих свойств в доступных наборах данных и квалифицированная интерпретация прогнозных карт имеют жизненно важное значение. Не менее важно умелое обращение с данными для создания информативных характеристик, которые существенно влияют на точность прогнозов. Комбинируя эти факторы, можно принимать более обоснованные решения и получать ценную информацию для эффективного направления поисков.

6. Вывод.

Была успешно сгенерирована карту перспективности минерализации Fe-Ti-P-REE путем обучения модели машинного обучения на геофизических и топографических данных высокого разрешения. Для создания набора данных были интегрированы геохимические анализы почвы и коренных пород. Данный подход способствует развитию поисков, представляя практическую основу для ГРП на ранней стадии. Прогностическая модель эффективно сужает районы для дальнейших поисков в регионе рифта Осло.

V. МАШИННОЕ ОБУЧЕНИЕ ДЛЯ ОПРЕДЕЛЕНИЯ ГРАНИЦ ГЕОХИМИЧЕСКИХ АНОМАЛИЙ (*район Юаньбонанг, провинция Ганьсу, Китай*) [5]

1. Введение.

Проверка аномальных концентраций металлов является основой геохимических поисков месторождений полезных ископаемых, а геохимические аномалии обычно являются результатом образования месторождений в виде первичных или вторичных ореолов рассеяния. Для оценки концентраций металлов в исследуемой области предлагаются четыре общих критерия: (1) величина геохимических значений и фон; (2) размер и форма аномальных областей; (3) геологические условия; и (4) степень, в которой местная окружающая среда могла повлиять на содержание металлов и характер аномалий. Другим критерием является зональность элементов в аномалиях. Для изучения этих критериев существуют различные методы, начиная от традиционных стандартных статистических методов и заканчивая методами машинного обучения (ML). Например, среднее значение $\pm 2 \times$ стандартное отклонение, графики вероятностей и гистограммы частотных распределений являются обычными для различия аномально высоких, а также аномально низких уровней концентрации металлов. Статистические анализы часто выявляют линейные корреляции. Учитывая пространственную структуру геохимических данных, скользящее среднее, пространственный факторный анализ и кригинг являются одними из популярных методов пространственной статистики для моделирования распределения геохимических данных. Что касается геохимического фона, то в данных часто встречаются локальные пространственные вариации, которые могут привести к неправильной классификации слабых геохимических аномалий как фоновых. Для решения проблемы слабых геохимических аномалий, которые имеют лишь немного более высокие значения концентрации по сравнению с фоновыми, но могут быть связаны с месторождениями, семейство фрактальных/мультифрактальных моделей широко использовалось для моделирования изменений геохимического фона. Другие популярные методы для анализа изменений геохимического фона включают анализ поверхности тренда и географически взвешенную регрессию. Хотя эти методы доказали свою эффективность при выявлении слабых геохимических аномалий, их предположения о распределении данных (например, регулярно и плотно распределенные точки отбора проб) не позволяют с точностью отобразить сложность изменения геохимического фона.

Алгоритмы машинного обучения хорошо подходят для количественной оценки сложных и нелинейных геохимических данных, позволяя обойти традиционные многомерные статистические методы, которые полагаются на данные, соответствующие заранее

определенной многомерной гипотезе распределения вероятностей. Когда количество помеченных образцов ограничено, неконтролируемое машинное обучение, в частности сеть автокодирования и ее вариации, продемонстрировали эффективность в обнаружении геохимических аномалий. Наличие достаточного количества помеченных образцов, представляющих классы геохимических аномалий (т. е. связанных с известными месторождениями полезных ископаемых) и классы геохимического фона, позволяет геохимикам-поисковикам создать контролируемую систему распознавания образов с машинным обучением для обнаружения геохимических аномалий. Современные алгоритмы ML с контролем обладают потенциалом для обнаружения геохимических аномалий. Например, машина опорных векторов (SVM) и ее естественное расширение - машина опорных векторов одного класса (OCSVM) были внедрены для исследования геохимических аномалий в сложных геологических условиях.

Несмотря на машинное обучение, особенно его подмножество глубокого обучения (например, сверточные нейронные сети), демонстрирующее замечательные результаты в конкретных областях исследований, растущая сложность модели и полезность моделей "черного ящика" затрудняют понимание обработки данных в определенных алгоритмах ML. Необходимо основное внимание уделять методам, которые учитывают интерпретируемость прогнозов машинного обучения относительно геохимических аномалий, помогая геохимикам понимать их при поисках. С этой целью были применены график частичной зависимости (PDP) и график накопленного локального эффекта (ALE), чтобы помочь посредством анализа геохимических данных в рамках регрессии и классификации с помощью ML-моделирования. Более того, был принят анализ значений SHAP, чтобы помочь понять вклад известных месторождений в качестве обучающих выборок при моделировании классификации геохимических данных методом случайного леса (RF).

2. Материалы и методы.

Исследованный район Юаньбонанг находится в провинции Ганьсу, Китай. Он относится к западному сектору Циньлинского орогена с золото-сурьмяным рудным потенциалом. Основные типы горных пород (плагиоклазово-кварцевые песчаники) терригенно-осадочных формаций Дахэба и Цзянлигоу, которые различаются между собой составом, циклическим расположением и размерами зерен. В юго-восточном секторе исследуемой территории закартирована неогеновая формация, состоящая в основном из конгломератов. Четвертичные отложения на исследуемой территории, состоят из песка, гравия, глины и гумуса. Преобладающие структуры - разломы северо-восточного простирания. Выявленные рудные

объекты в исследуемом районе - 10 месторождений золота и 2 - сурьмы. Они относятся к средне- и низкотемпературному типу магма-гидротермальных месторождений (рис. 1).

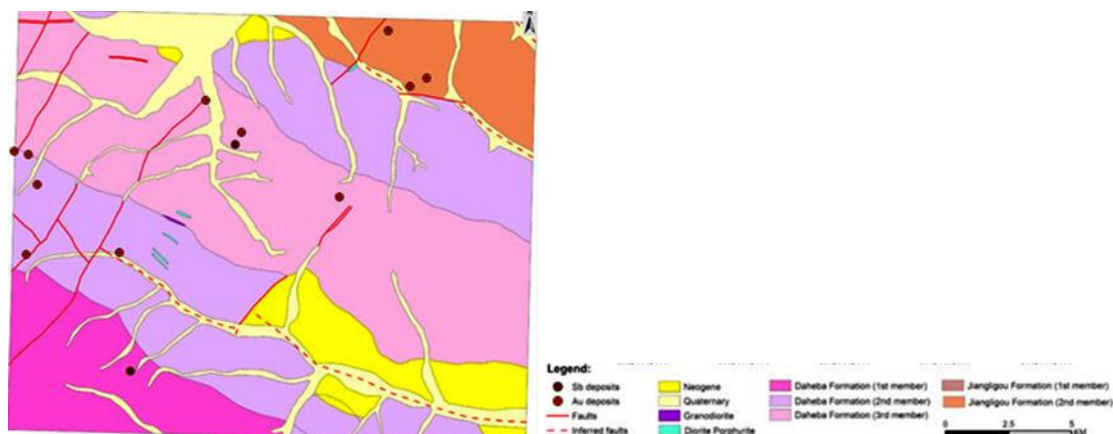


Рис. 1. Упрощенная геологическая карта исследуемого района.

Набор использованных геохимических данных включает 1859 образцов донных отложений ручьев, собранных в масштабе 1:50 000. Масс-спектрометрия с индуктивно связанной плазмой использовалась для анализа образцов на 16 элементов (Au, As, Sb, Hg, Co, Cr, Cu, Ni, Pb, Zn, Sn, Ag, Mo, Cd, W и Bi). Среди них восемь элементов (Au, As, Sb, Hg, Cu, Pb, Zn и Ag) были идентифицированы как потенциально связанные с минерализацией золота и сурьмы в районе. Эти элементы были выбраны при анализе геохимических аномалий.

2.1. Случайный лес (RF).

В качестве метода коллективного обучения RF включает набор деревьев решений (деревья классификации и регрессии - CART), что составляет основу RF-метода. С помощью обучающего набора данных можно построить ансамбль деревьев принятия решений и получить результаты прогнозирования, используя режим прогнозов (в рамках классификации) или среднее значение прогнозов (в рамках регрессии) по отдельным деревьям. В процессе построения RF-модели независимые деревья строятся с помощью уникальной начальной выборки обучающего набора данных, посредством чего функция начальной выборки случайным образом делит обучающий набор данных на однородные подмножества. Затем случайные подмножества образцов используются для выращивания и обучения каждого дерева, а оставшееся подмножество (т.е. готовые образцы - OOB) используются для валидации и оценки точности модели. Важной особенностью RF является его способность оценивать относительную важность входных переменных путем построения чередующихся деревьев решений, включающих и исключающих каждую переменную при расчете среднеквадратичной ошибки (MSE) и чистоты узла. Оценка относительной важности переменной выполняется путем оценки приращения ошибки прогнозирования, когда данные OOB для этой переменной изменяются, в то время как все остальные переменные остаются неизменными. Радиочастотное

моделирование для определения границ геохимических аномалий выполняется как в рамках регрессии, так и в рамках классификации.

2.2. График частичной зависимости (PDP).

В PDP показаны изменения средних прогнозируемых значений по мере того, как указанные признаки изменяются в пределах их предельного распределения. PDP внес большой вклад в понимание различных контролируемых приложений ML в различных дисциплинах. PDP может продемонстрировать взаимосвязь (например, линейную, монотонную или комплексную) между целевой переменной и предикторными переменными. Позвольте $S \in \{1, \dots, p\}$ и пусть C быть дополняющим набором S . Затем PDP для регрессии определяется следующим образом:

$$\hat{f}_{x_S}(x_S) = E_{x_C} [\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C) dP(x_C) \quad (1)$$

где S и C являются подмножествами предикторов; x_S обозначает объекты, для которых создаются PDP; и x_C представляет дополнительные функции, используемые в модели ML $\hat{f}$; x_S и x_C составляют общее пространство объектов.

PDP отодвигает на второй план результаты моделирования ML по сравнению с распределением объекта в наборе C . Таким образом, функция отображает взаимосвязь между объектами в S и прогнозы.

На практике оценка $\hat{f}_{x_S}$ выполняется путем вычисления средних значений в обучающих данных с помощью метода Монте-Карло; таким образом,

$$\hat{f}_{x_S}(x_S) = \frac{1}{n} \sum_{i=0}^n \hat{f}(x_S, x_C^{(i)}) \quad (2)$$

где $\{x_C^{(1)}, \dots, x_C^{(n)}\}$ представляет различные значения x_C наблюдаются в данных обучения, и n обозначает количество экземпляров в наборе данных.

Это уравнение дает среднее предельное влияние на прогноз при заданном значении x_S . Однако PDP предполагает, что особенности в C не коррелируют с особенностями в S . Нарушение этого допущения может привести к появлению очень маловероятных или невозможных точек данных в PDP. В контексте классификации, где модель ML выводит вероятности, PDP показывает вероятность для определенного класса при различных значениях для объекта в S .

2.3. График накопленного локального эффекта (ALE).

График ALE, как и PDP, описывает среднее влияние объектов на прогноз с помощью модели ML. Однако считается, что это более быстрый и непредвзятый подход, чем PDP. Используя график PDP или ALE, сложную функцию прогнозирования можно свести к функции, которая зависит только от одного (или двух) объекта путем усреднения влияния других

объектов. Однако PDP усредняет прогноз по предельному распределению. Напротив, вместо прямого усреднения прогноза, график ALE вычисляет различия в прогнозе, обусловленные особенностью в S , и оно интегрирует производную по объекту в S для оценки эффекта; таким образом,

$$\hat{f}_{x_S ALE}(x_S) = \int_{z_{0,1}}^{x_S} E_{X_C | X_S} \left[\hat{f}^S(X_S, X_C) | X_S = z_S \right] dz_S - constant = \int_{z_{0,1}}^{x_S} \int_{x_C} \hat{f}^S(z_S, x_C) \mathbb{P}(x_C | z_S) dx_C dz_S - constant \quad (3)$$

Формула для ALE указывает, что она усредняет различия в прогнозах, а не сам прогноз, поэтому это градиент. На практике различия в прогнозах заменяются различиями в прогнозах на интервале (рис. 2).

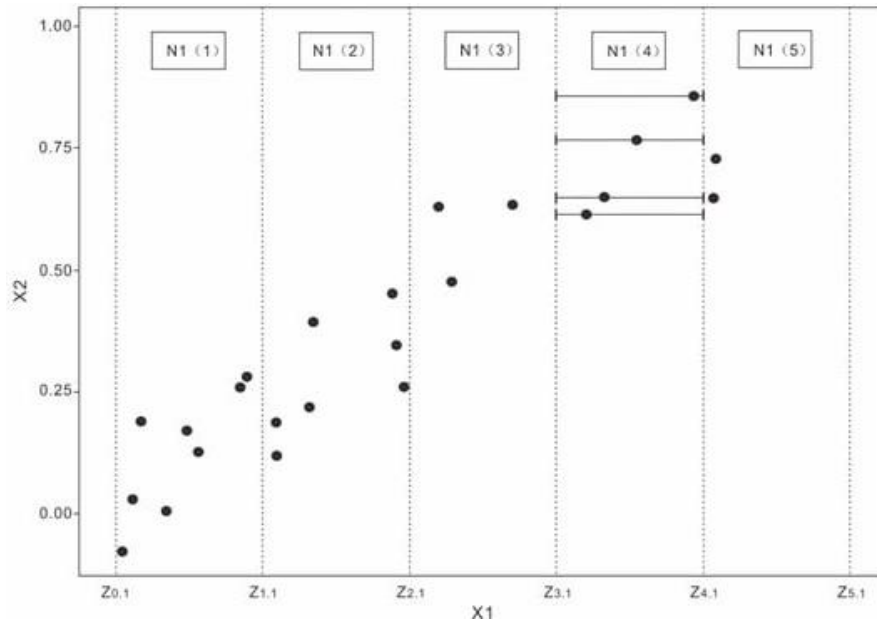


Рис. 2. Схематический пример расчета ALE для объекта X1 (коррелирует с X2).

Сначала $X1$ разделяется на интервалы (пунктирные вертикальные линии). Для точек данных внутри интервала различия в прогнозе при замене $X1$ на верхнюю и нижнюю границы интервала (сплошные горизонтальные линии) вычисляются, накапливаются и центрируются, получая кривую ALE.

2.4. Анализ SHAP.

Анализ SHAP является популярным методом для характеристики производительности модели ML, основанной на теории игр. Первоначально это было предложено как аксиоматическое описание справедливого распределения общего избытка для всех участников, которое может быть использовано в прогностических моделях, в которых каждая переменная напоминает игрока.

Давайте возьмем коалиционную игру в качестве кортежа (N, v) где $N = \{1, 2, \dots, n\}$ представляет собой конечный набор n игроки и $v: 2^N \rightarrow \mathbb{R}$ выполняет характерную функцию с

$v(\Phi)=0$, определено более S (т.е. подмножества N), который характеризует выигрыш каждого подмножества N . Цель этих определений в такой коалиционной игре - получить решение для распределения общей суммы выигрыша между игроками, т.е., $v(N)$ для всех игроков. Здесь, $\Phi_i(v)$ обозначает назначенный игроку выигрыш i в терминах характеристической функции v . Таким образом, значение Шепли представляет собой решение для распределения Φ с уникальными свойствами.

Для игры $\langle N, v \rangle$, ценность игрока по Шепли i определяется следующим образом:

$$\Phi_i(v) = \sum_{S \subseteq N - \{i\}} \frac{(|N| - |S| - 1)! |S|!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (4)$$

Таким образом, ценность игрока по Шепли i представляет собой взвешенную сумму вклада i в каждом подмножестве N . Вклад i для подмножества S определяется $v(S \cup \{i\}) - v(S)$.

Хотя значение Шепли изначально было сформулировано в контексте теории игр с простым математическим определением, недавно оно было использовано для объяснения сложных моделей, генерируемых алгоритмами ML, но смысл использования его в качестве инструмента для интерпретации модели еще не был проиллюстрирован.

3. Результаты и обсуждение.

Поскольку в этом исследовании были проанализированы восемь элементов (Au, As, Sb, Hg, Cu, Pb, Zn и Ag) (рис. 3), параметры для количества деревьев (k) и количества объектов (m), отобранных случайным образом при каждом разделении, были установлены как $k = 1000$ и $m = 3$ после процесса проб и ошибок. Параметры k и m использовались для регрессии и классификации в RF-моделировании.

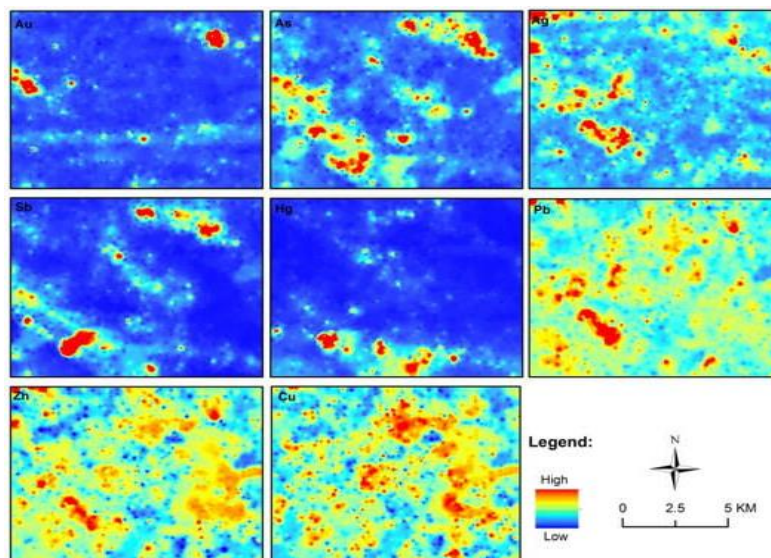


Рис. 3. Геохимические элементы, используемые в радиочастотном моделировании.

3.1. Регрессия с помощью RF-моделирования.

Регрессионный анализ в этом исследовании направлен на поиск ассоциаций между известной минерализацией и потенциально релевантными геохимическими элементами. Поскольку на исследуемой территории преобладает минерализация Au, золото служило зависимой переменной, а остальные семь элементов служили независимыми переменными, формируя входные данные для моделирования случайной регрессии лесов (RFR). На основе рангов переменной важности было установлено, что As и Sb вносят наиболее значительный вклад в RFR с точки зрения приращений MES и чистоты узлов (измеряемых индексом Джини). Это говорит о том, что As и Sb тесно связаны с распространением Au (табл. 1).

Табл. 1.

Переменная важность для регрессии RFR.

Variables	As	Sb	Hg	Cu	Pb	Zn	Ag
%IncMSE	5.13	3.61	0.94	1.54	0.37	1.91	0.02
IncNodePurity	2914	1590	1381	1269	1206	1124	1134

Дополнительную информацию могут предоставить графики ALE и PDP, поскольку эти графики могут помочь в описании влияния объектов на прогноз с помощью модели ML. Наиболее важные геохимические элементы, As и Sb, имели положительную взаимосвязь с Au, и они демонстрировали схожие траектории (рис. 4).

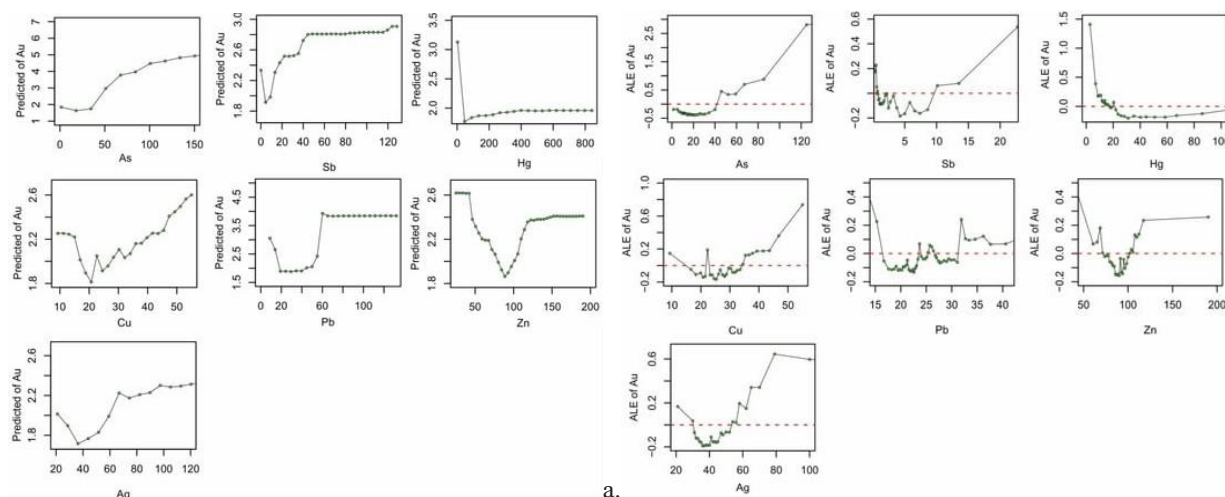


Рис. 4 Графики частичной зависимости (а) и список (б) геохимических элементов для RFR-моделирования.

Для PDP наибольшей проблемой является допущение независимости, а именно, что характеристики для расчета PDP не должны коррелировать с другими характеристиками. Однако график ALE вычисляет различия в прогнозах и изолирует влияние интересующих объектов, тем самым блокируя влияние других объектов. Таким образом, в целом траектории всех геохимических элементов в PDP соответствовали траекториям на графике ALE, что доказывает надежность PDP и графика ALE в этом тематическом исследовании.

Более того, графики ALE были центрированы так, чтобы среднее влияние на прогноз равнялось 0 путем вычитания константы. Значение ALE может быть интерпретировано как основной эффект объекта при определенном значении по сравнению со средним прогнозом данных. Оценка ALE с положительным значением (например, 0,3) означает, что прогноз Au выше на 0,3, чем средний прогноз. Таким образом, они обеспечивают потенциальные пороговые значения для определения геохимических аномалий (например, значения переменных на графике ALE при $Au = 0$). Поскольку геохимические переменные As и Sb были признаны наиболее важными при моделировании RFR, пороговые значения 42 для As и 9 для Sb (т. е. пороговые значения ALE $Au = 0$) были выбраны из соответствующих графиков ALE (рис. 4) и использованы для определения границ геохимических аномалий (рис. 5).

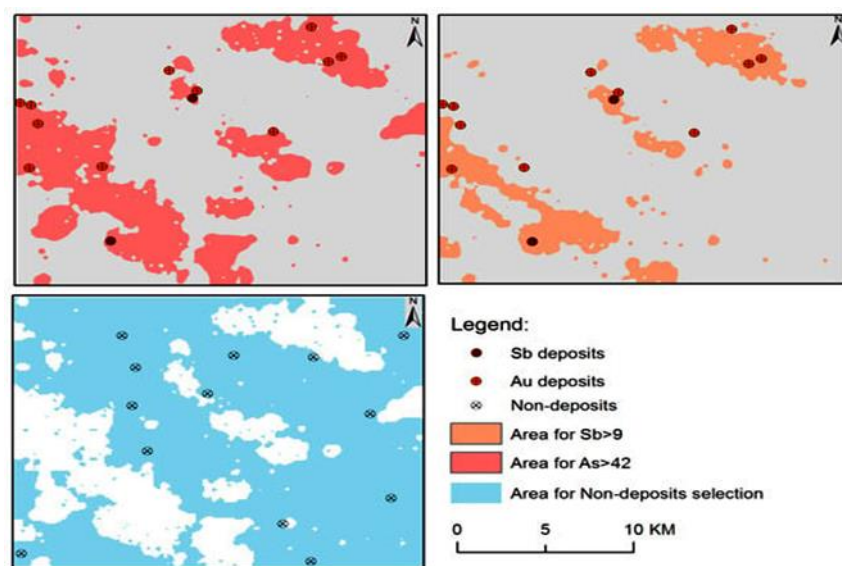


Рис. 5. Выбор площадей без месторождений на основе графиков ALE As и Sb.

Геохимические аномалии As очертили 12 из 13 известных минерализованных участков в исследуемом районе, и они показали тенденцию к северо-западу. Напротив, геохимические аномалии Sb очертили меньшее количество из 13 известных местоположений, но их последовательный северо-западный тренд подразумевает 3 похожих перспективных района для дальнейших поисков (рис. 5). Более того, в сценарии MPM с использованием контролируемого обучения отбор отрицательных проб (например, не залежей) для контролируемого MPM остается сложной задачей. Здесь области, исключая аномалии As и Sb, являются предпочтительными местами для создания отрицательных образцов. Таким образом, 13 известных минерализованных участков были использованы в качестве положительных образцов, а области, исключая аномалии As и Sb (т. е. области со значениями $Sb > 9$ и $As > 42$), были выделены и использованы в качестве областей для отбора случайно отрицательных образцов (13 не залежей) для последующего моделирования классификации RF (рис. 5).

3.2. Классификация с помощью радиочастотного моделирования.

RFR-моделирование позволило получить представление о взаимосвязи As, Sb, Hg, Cu, Pb, Zn и Ag с целевой минерализацией Au в исследуемом районе. Однако существуют сценарии, в которых минерализация отсутствует в случае ложных аномалий с высокой магнитудой. Например, участки с высокими значениями Au могут не означать благоприятности района для минерализации. Необходимо выявить корреляцию геохимических элементов с известными минерализованными участками. Классификация RF была применена для сопоставления известных областей минерализации и концентраций изученных геохимических элементов (Au, As, Sb, Hg, Cu, Pb, Zn и Ag).

При моделировании RF важность достоверных карт оценивалась по среднему снижению точности и среднему снижению индексов Джини. Первое вычисляется на основе данных OOB и указывает на снижение точности во всем радиочастотном моделировании, в то время как второе измеряет среднее повышение чистоты, достигнутое за счет использования разбиений определенной переменной.

Результаты RF-моделирования показывают, что As и Sb являются наиболее важными геохимическими элементами, что согласуется с результатами RFR-моделирования (рис. 6).

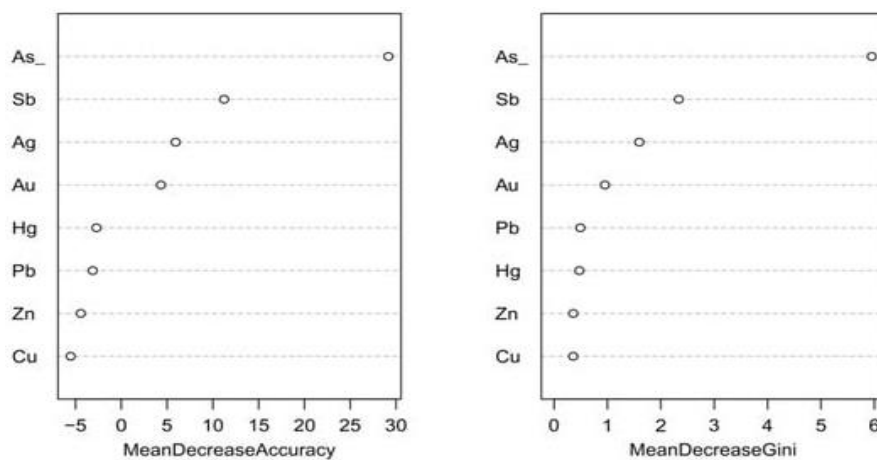


Рис. 6. Важность геохимических переменных при радиочастотном моделировании.

Au заняла четвертое место после As, Sb и Ag. В отличие от графика PDP и ALE, при RFR-моделировании Au, As, Ag и Sb не только продемонстрировали схожие траектории на графике PDP и ALE (рис. 7), но также продемонстрировали значительные различия в других геохимических переменных (Hg, Cu, Pb и Zn). Геологические/геохимические процессы можно распознать по непрерывному диапазону переменных откликов и соответствующему относительному увеличению или уменьшению концентрации элементов, а съемки с высокой плотностью данных способны обнаруживать процессы локального масштаба.

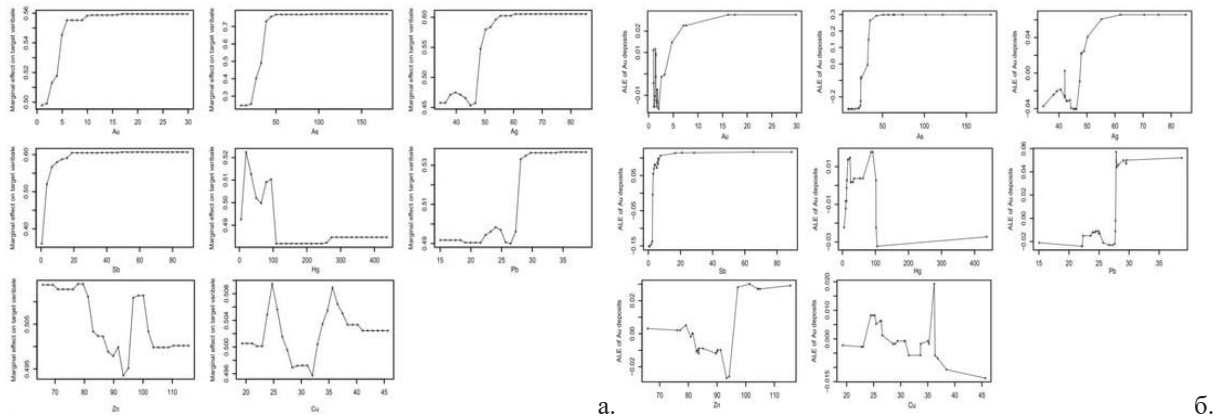


Рис. 7. Графики частичной зависимости (а) и геохимических элементов ALE (б) при радиочастотном моделировании.

Поведение геохимических переменных в радиочастотном моделировании и пространственное распределение их концентраций (рис. 3) подразумевают, что Au, As, Ag и Sb оказали положительное влияние на формирование известной минерализации. Совпадение местоположений известной минерализации с меньшими и менее интенсивными аномалиями может быть более показательным для залегания месторождений. Это побуждает к внедрению интерпретируемого ML для обнаружения геохимических аномалий.

Помимо PDP и графика ALE, которые проливают свет на взаимосвязи между представляющими интерес геохимическими элементами и известной минерализацией (т. е. концентрацией элемента Au при RFR-моделировании и месторождениями Au при RF-моделировании), были приняты значения Шепли для иллюстрации поведения обучающих выборок при прогнозировании на основе ML. В контексте RF-моделирования для MPM и с учетом текущего набора геохимических переменных значение Шепли служит приближенным показателем того, насколько конкретное значение геохимической переменной способствует расхождению между фактическим прогнозом и средним прогнозом.

Рисунок 8 показывает вклад каждого элемента экземпляра (т.е. обучающей выборки) в радиочастотном моделировании в известную минерализацию. На рисунке 8 а,б показаны две отрицательные обучающие выборки (т.е. не залежи), тогда как на рисунке 8 с,д показаны обучающие выборки, взятые из мест известной минерализации. Как отрицательные, так и положительные обучающие выборки подразумевают, что наибольший вклад внесли As, Sb и Ag.

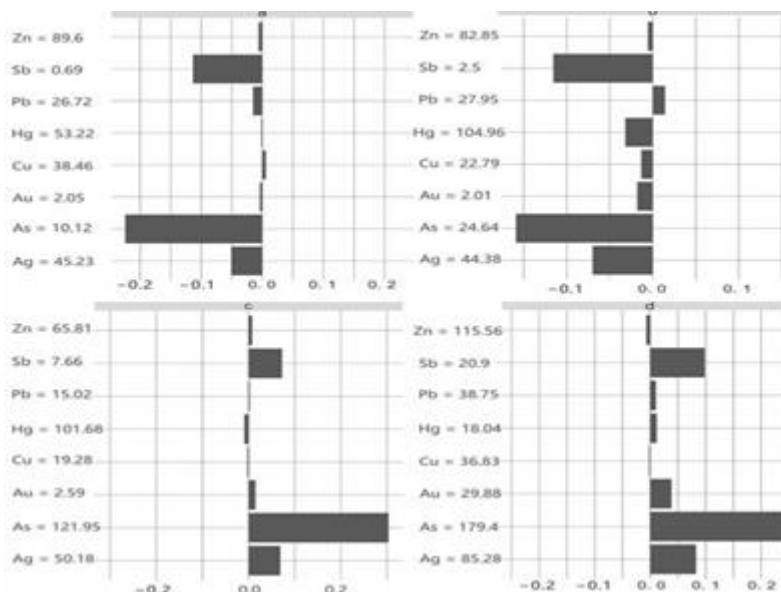


Рис. 8. Значения Шепли для обучающих выборок при радиочастотном моделировании: (a,b) отрицательные выборки; (c,d): положительные

Это согласуется с важностью переменной с точки зрения точности и индекса Джини (рис. 6). Более того, геохимические переменные отрицательных обучающих выборок показали отрицательный вклад, в отличие от положительных обучающих выборок, и значения Шепли положительных выборок различались, при этом As (помечено как 5), Sb (помечено как 12) и Ag (помечено как 4, 5, 6 и 8) имели отрицательный вклад (рис. 9).

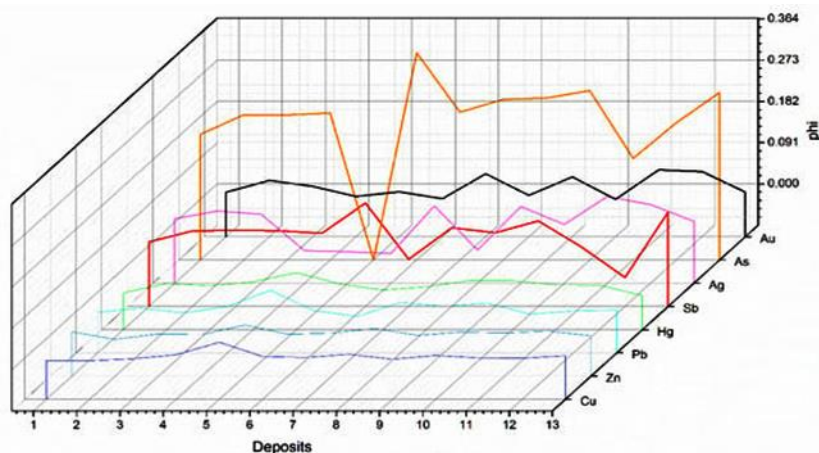


Рис. 9. Значения Шепли положительных обучающих выборок в радиочастотном моделировании.

Эти результаты свидетельствуют о том, что значение Шепли является перспективным инструментом для оценки эффективности обучающего набора данных при составлении карт перспективности полезных ископаемых на основе машинного обучения (МРМ), особенно в контексте определения мест, не являющихся месторождениями, для использования в качестве отрицательных обучающих выборок.

Окончательная карта перспективности показывает, что области с высоким потенциалом очерчены большинством известных месторождений (рис. 10).

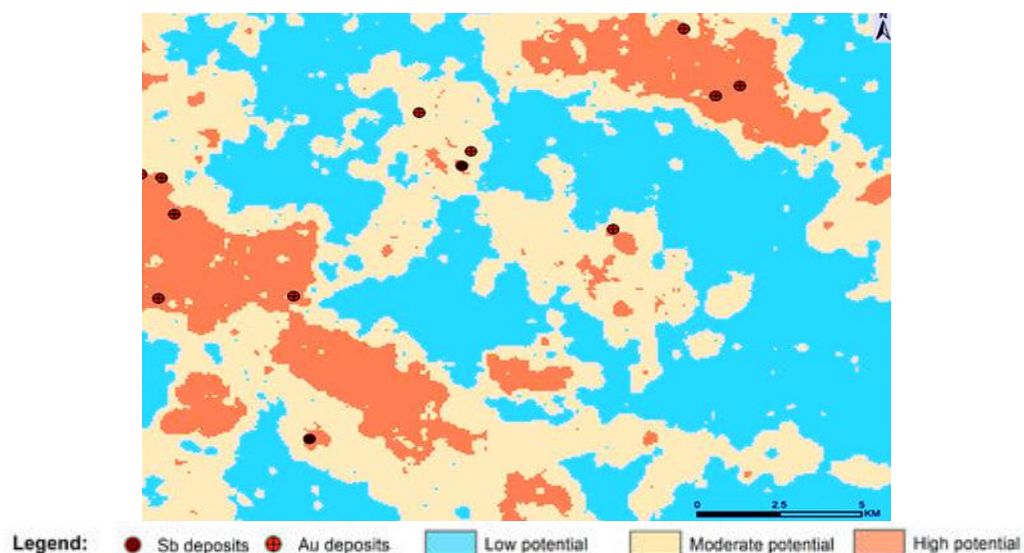


Рис. 10. Карта перспективности золота в результате радиочастотного моделирования.

Эффективность моделирования случайного леса (RF) оценивалась с использованием кривой рабочих характеристик приемника (ROC), которая иллюстрирует частоту истинно положительных результатов по сравнению с частотой ложноположительных результатов, при этом площадь под кривой (AUC) указывает на общую производительность RF-моделирования. Используя вероятность появления минерализации для вычисления AUC, получили значение 0,917 для месторождений (т.е. положительных обучающих выборок) и 0,182 для неместорождений. Эти результаты свидетельствуют о том, что радиочастотное моделирование эффективно проводило различие между известными рудными залежами и выбранными местами без залежей (рис. 11а,б). Взаимосвязь между полученным распределением значений перспективности и известными месторождениями полезных ископаемых оценивалась с использованием графиков прогнозных площадей (P-A) и фрактального анализа площадей концентраций (C-A). График P-A иллюстрирует отношение площади, покрытой классами перспективности (относительно общей площади исследования), к соотношению известных залежей (относительно общего числа известных залежей), предсказанных этими соответствующими классами перспективности. Результаты показывают, что радиочастотное моделирование выявило 9,7% исследуемой территории, точно отразив 84,6% известных месторождений. Напротив, фрактальный анализ C-A выявил три категории вероятности: высокий потенциал (вероятность $>0,80$), умеренный потенциал ($0,22 < \text{вероятность} < 0,80$) и низкий потенциал (вероятность $< 0,22$) (рис. 11в).

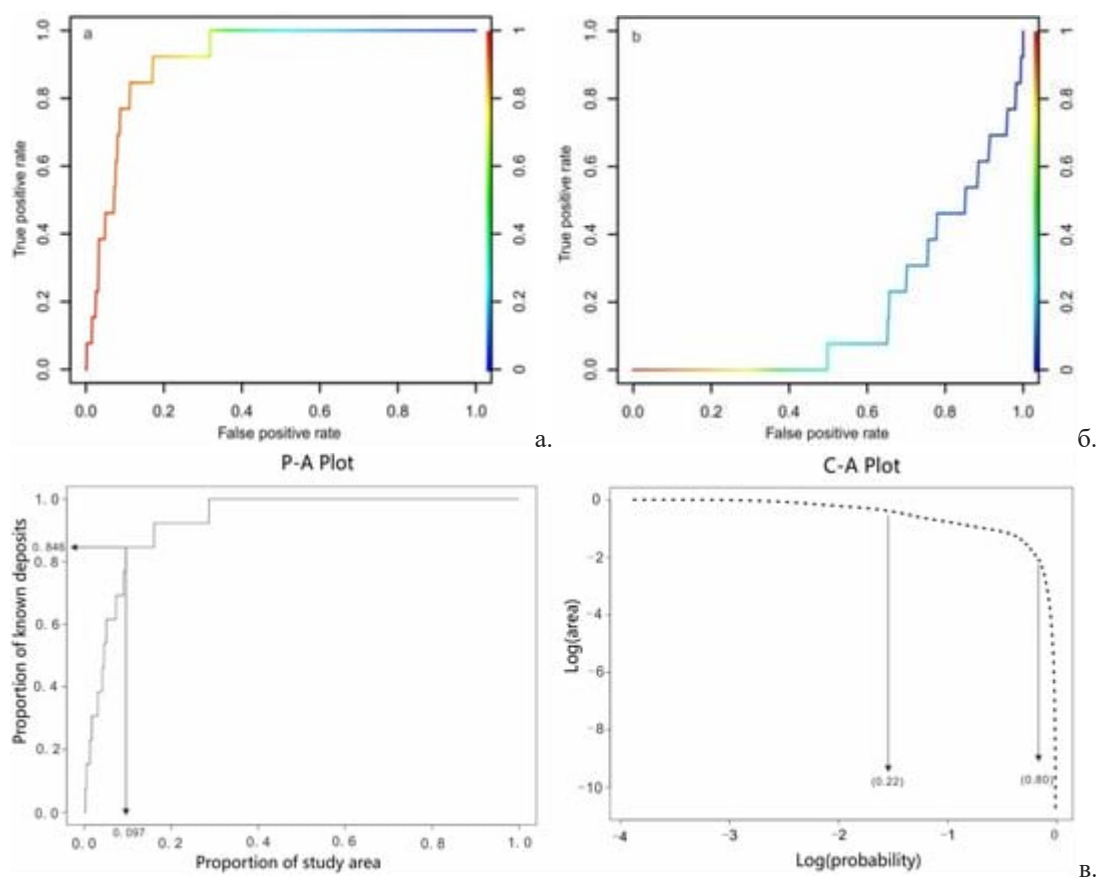


Рис. 11. Кривые ROC для (а) месторождений ($AUC = 0,917$) и (б) не-месторождений ($AUC = 0,182$).
Графики P-A и C-A для карты перспективности золота (в).

4. Выводы.

1. Алгоритмы машинного обучения продемонстрировали свою эффективность при составлении карт перспективности (MPM). Тем не менее, природа большинства алгоритмов ML в виде черного ящика приводит к ограниченному пониманию взаимосвязи между входными переменными (т. е. картами фактических данных) и целевой переменной (т. е. рудными залежами). Были использованы анализы PDP и ALE plot для визуализации влияния геохимических переменных на минерализацию в контексте моделей регрессии и классификации. Был использован SHAP - анализ для выяснения эффективности отдельных обучающих выборок при моделировании машинного обучения. Использование этих инструментов интерпретации позволило получить количественную информацию о концентрациях геохимических элементов для определения границ геохимических аномалий. Эти инструменты повысили достоверность полученной информации о геохимических аномалиях, предполагая, что исследования интерпретируемых прогнозов на основе машинного обучения могут еще больше продвинуть практику составления карт перспективности (MPM).

2. Проведенный анализ включал сочетание математических концепций с идентификацией геохимических аномалий, то есть преобразование математических параметров в геохимические

параметры для улучшения распознавания и оценки геохимических аномалий. Тем не менее, в исследуемом районе интерпретация геохимических аномалий с использованием предлагаемой методологии, вероятно, может быть обогащена дополнительными анализами, например, комбинированным петрографическим и микроскопическим анализом руд, поскольку это обеспечит более целостное понимание происходящих металлогенических процессов и позволит получить более глубокое представление о минералогическом составе и микроструктурных особенностях руд исследуемой территории. Это рекомендация для дальнейшего изучения.

VI. ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОРФИРОВОГО ОРУДЕНЕНИЯ НА ОСНОВЕ МИКРОЭЛЕМЕНТОВ ЦИРКОНА. (металлогеническая провинция Тетис) [1].

1. Введение.

Порфировые месторождения дают примерно 75% мировой добычи меди, 50% добычи молибдена и 20% добычи золота, а также значительную долю серебра, цинка, олова, свинца и других металлов. Эти месторождения образуются флюидами, связанными с водными и окисленными магмами, которые отличаются повышенным соотношением Eu/Eu^* , V/Sc и Sr/Y . Эти геохимические признаки становятся всё более ценными при поисках порфировых месторождений. Однако выветривание, гидротермальные изменения и метаморфизм могут значительно изменить эти признаки, что приводит к неоднозначным выводам при их использовании в качестве прогнозных индикаторов.

Циркон, стабильный и повсеместно распространённый акцессорный минерал, встречающийся в среднекислых породах, стал эффективным индикаторным минералом для порфировых месторождений. Он отражает составные части и физико-химические параметры исходной магмы, такие как температура, степень окисления и содержание воды. Для отслеживания плодородных магм часто используются различные индикаторы, в том числе соотношения Eu/Eu^* и Ce/Ce^* в цирконах. Однако диаграммы бинарной классификации микроэлементов циркона, используемые в разных регионах, неприменимы в других областях, где границы классификации неясны.

Алгоритмы машинного обучения (ML) обладают значительными преимуществами при работе с большими наборами данных и многомерными характеристиками благодаря своей интерпретируемости и прозрачности. Эти методы открывают значительные перспективы для поисков порфирового оруденения, поскольку размер наборов данных продолжает расти. Предыдущие исследования были сосредоточены на отдельных месторождениях, и исследования для рудных районов и провинций не проводились. Рудные тела, сформировавшиеся в определённых условиях, часто имеют значительные различия в геохимических характеристиках.

В рамках этого исследования были собраны в общей сложности 3751 наборов данных по цирконам из рудных районов провинции Тетис. Основные результаты этой работы можно описать следующим образом: на данных по цирконам использовались различные алгоритмы ML и была создана и визуализирована модель классификации порфировых месторождений различных типов на основе различия в характеристиках микроэлементов.

2. Геологические условия.

Тетийская металлогеническая провинция простирается примерно на 10 000 км с востока на запад. Она подверглась расширению, осадочному накоплению и поднятию Нео- и Палео Тетиса, а также двум крупномасштабным коллизиям с субдукцией. Динамические геологические процессы привели к образованию всемирно известных металлогенических поясов с крупнейшими порфировыми медно-золото-молибденовыми и др. полиметаллическими месторождениями. Металлогенические пояса и месторождения показаны на рисунке 1 и в таблице 1.

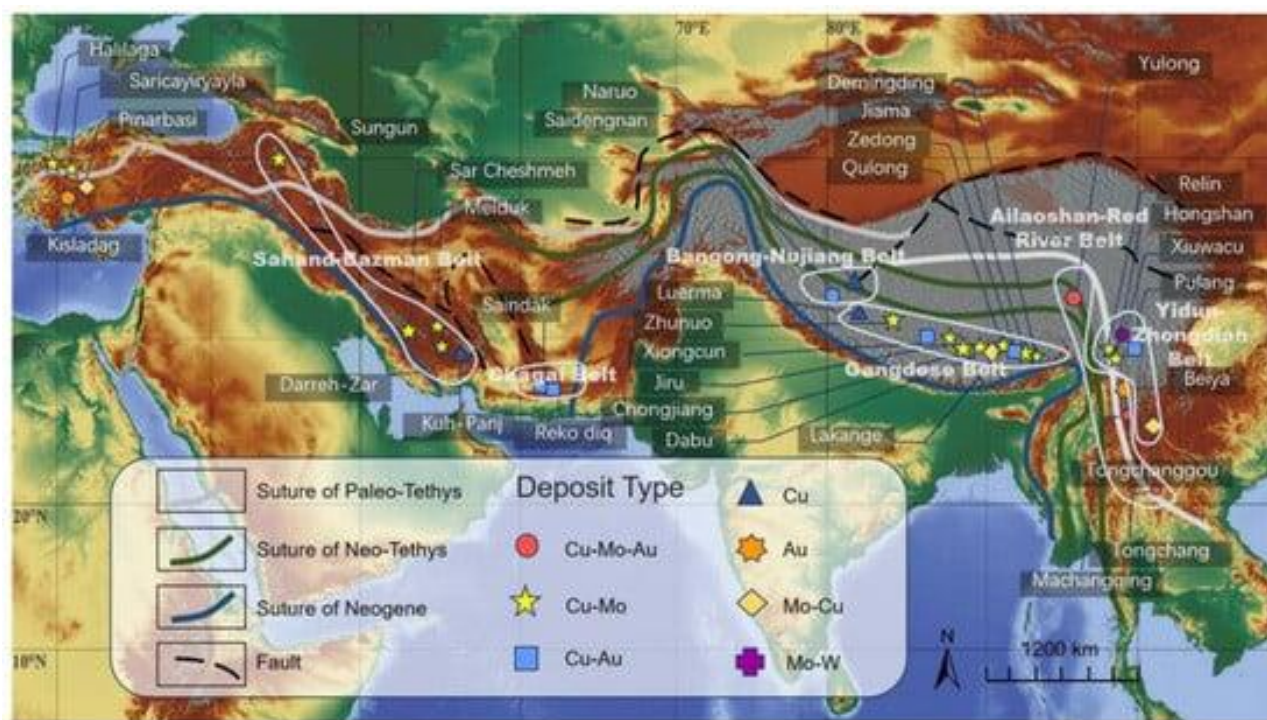


Рис. 1. Тектоническая обстановка и расположение металлогенических поясов в Тетисе

Табл. 1.

Собранные данные по цирконам на месторождениях Тетийских металлогенических поясов

Deposit	Type	Tonnage(Mt) and Grade	Zircon Count	Fertility
GangdeseBelt				
Xiongcu	Cu-Au	874, Cu: 2.96 @0.34%, Au: 218.4t @0.25g/t	283	Fertile
Qulong	Cu-Mo	2200, Cu: 8 @0.3–0.6%, Mo: 0.03–0.06%	97	Fertile
Jiama	Cu-Mo	1055, Cu: >5 @1.16%, Mo: 0.024%	77	Fertile
Jiru	Cu-Mo	42, Cu: 0.18 @0.43%	32	Fertile
Lakang'e	Mo-Cu	Cu: >0.1, Cu: 0.13–0.49%	37	Fertile
Zhunuo	Cu-Mo	403, Cu: 2.3 @0.57%	54	Fertile
Demingding	Mo-Cu	Cu: >0.5 @0.26 %, Mo: >0.5 @0.14 %	15	Fertile
Chongjiang	Cu-Mo	45, Cu: 0.18 @0.4%, Mo: 0.01 @0.02%	28	Fertile
Dabu	Cu-Mo	250, Cu: 0.5 @0.2%, Mo: 0.05 @0.02%	15	Fertile
Luerma	Cu	Cu: 0.26% ~ 0.73%, Au: 1.2 ~ 37.7 g/t	69	Fertile
\	\	\	485	Barren
Bangong–Nujiang Belt				
Naruo	Cu	578, Cu: 2.37 @0.41%, Mo: 104 @0.18%	100	Fertile
Guaguazi	Cu	\	35	Fertile
Saidengnan	Cu-Au	\	11	Fertile
\	\	\	214	Barren

Ailaoshan–Red River Belt					
Yulong	Cu-Mo-Au	1006, Cu: 6.24 @0.62%, Mo: 0.42 @0.04%, Au: 50.3t @0.05g/t	118	Fertile	
Machangqing	Cu-Mo-Au	Cu: 0.25 @0.44%, Mo: 0.14%, Au: 0.06g/t	65	Fertile	
Beiya	Au	Au: 323t @2.47g/t, Cu: 56.68 @0.52%	170	Fertile	
Tongchang	Cu-Mo-Au	Cu: 37,958t @1.48%, Mo: 30,370t @0.218%, Au: 0.13–0.25g/t	47	Fertile	
\	\	\	302	Barren	
Yidun–Zhongdian Belt					
Pulang	Cu-Au	804, Cu: 446.8 @0.52%, Au: 0.18 g/t	64	Fertile	
Relin	Cu-Mo	6.6, Mo: 0.049%	22	Fertile	
Xiuwacu	Mo-W	6.6, Mo: 13,627t @0.34%, WO ₃ : 8431t @0.28%	74	Fertile	
Hongshan	Cu-Mo	65.1, Cu: 0.64 @1.23 %, Mo: 5769t @0.03 %	122	Fertile	
Tongchanggou	Mo-Cu	142.5, Mo: 0.3 @0.3 %, Cu: 34 kt @0.8%	45	Fertile	
\	\	\	146	Barren	
Sahand–Bazman Belt					
Sungun	Cu-Mo	850, Cu: 0.76 %, Mo: 0.01%	45	Fertile	
Sar Cheshmeh	Cu-Mo	1700, Cu: 0.65%, Au: 0.06 g/t, Mo: 0.03%	103	Fertile	
Kuh-Panj	Cu	100, Cu: 100 @0.21%	18	Fertile	
Darreh-Zar	Cu-Mo	283, Cu: 1.1 @0.38%	1	Fertile	
Meiduk	Cu-Mo	500, Cu: 0.86%, Mo: 0.007%	2	Fertile	
\	\	\	104	Barren	
Chagai Belt					
Saindak	Cu-Au	Cu: 27Mt	64	Fertile	
Reko diq	Cu-Au	Cu: 1.23Mt@0.41 %, Au: 59.3 t@0.22 g/t	88	Fertile	
\	\	\	30	Barren	
Other Districts					
\	\	\	83	Fertile	
\	\	\	486	Barren	

3. Методы.

3.1. Данные и предварительная обработка.

Данные о магматических цирконах в основном были взяты из литературы, посвящённой порфировым месторождениям в Тетийской области, а также включали новые данные о цирконах, собранные в ходе этого исследования. Цирконы с трещинами или включениями были исключены из анализа, так как их согласованность составляла менее 90%. Наследуемые цирконы были удалены, а значения ниже минимального предела обнаружения были отфильтрованы. Образцы были взяты из 29 порфировых месторождений в шести металлогенических поясах Тетийской области.

Перед обучением и классификацией данных требуется очистка, чтобы избежать влияния аномальных данных на результаты классификации. Во-первых, данные по циркону с $La > 1$ ppm и $Ti > 50$ ppm были удалены, чтобы исключить загрязнение, вызванное апатитом и оксидом Ti-Fe. Отсутствие значений в определенных случаях потенциально может привести к ложноположительной идентификации. Во избежание влияния на эффект классификации недостающие значения удаляются. Наконец, 1203 фертильных циркона и 1069 бесплодных

цирконов были использованы для построения моделей машинного обучения. Данные подверглись логарифмическому преобразованию и стандартизации по Z-баллам после очистки данных. При подготовке модели классификации плодородного и бесплодного циркона в качестве классификационных показателей были выбраны Ti, Y, La, Ce, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu и Hf. Три подзаголовка на рисунке 2, соответственно, показывают исходное распределение микроэлементов циркона, распределение после логарифмического преобразования и распределение после стандартизации по Z-шкале.

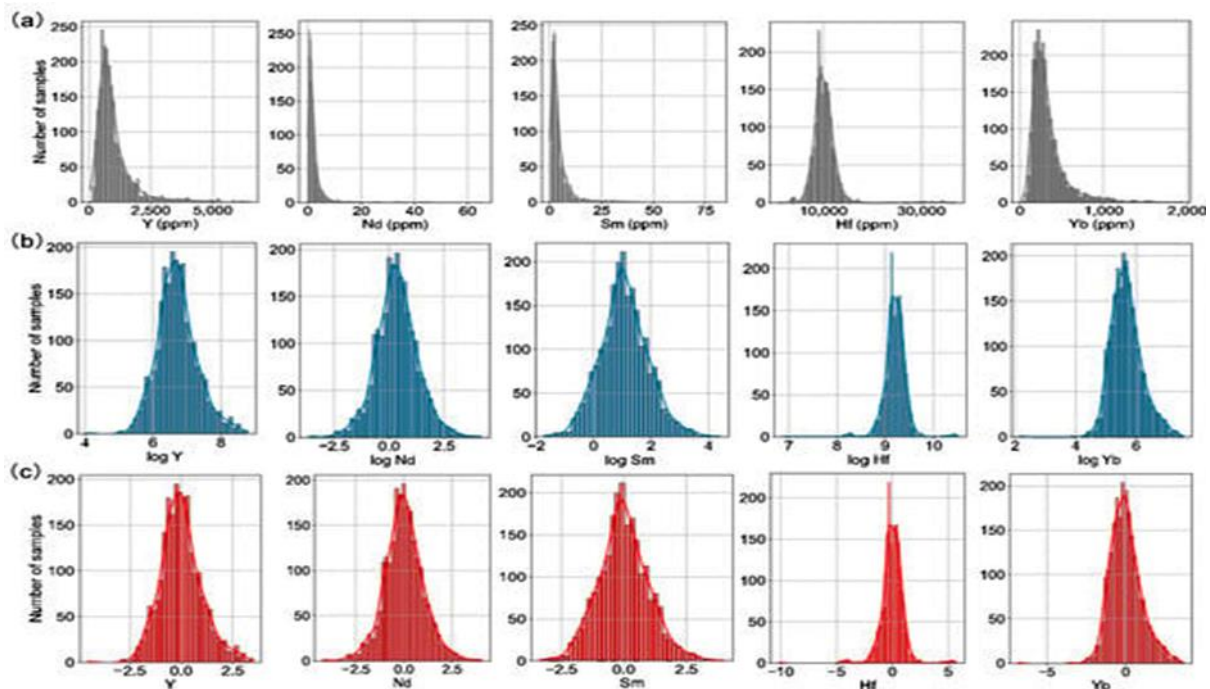


Рис. 2. (a) Исходные гистограммы микроэлементов в цирконе; (b) гистограммы микроэлементов в цирконе после логарифмического преобразования; (c) гистограммы микроэлементов в цирконе после стандартизации по Z-критерию.

Из-за большого количества показателей для построения графиков были выбраны только пять микроэлементов.

3.2. Алгоритмы машинного обучения.

«Случайный лес» (RF), предложенный Брейманом в 2001 году, — это алгоритм ансамблевого обучения, который сочетает в себе теорию ансамблевого обучения с бустингом и метод случайного подпространства. Его основной принцип показан на рисунке 3а: «случайный лес» использует метод повторной выборки с бутстрэппингом для извлечения нескольких выборок из набора данных и случайным образом выбирает некоторые характеристики каждой выборки с бутстрэппингом для создания модели дерева решений. Эти модели обучаются, прогнозируют и получают результаты независимо друг от друга. Наконец, результаты каждого дерева решений объединяются, и окончательный прогноз получается путем «голосования». Преимущества random forest заключаются в том, что он может

обрабатывать многомерные данные с большим количеством функций, обладает высокой устойчивостью к шуму и выбросам и не склонен к переоснащению.

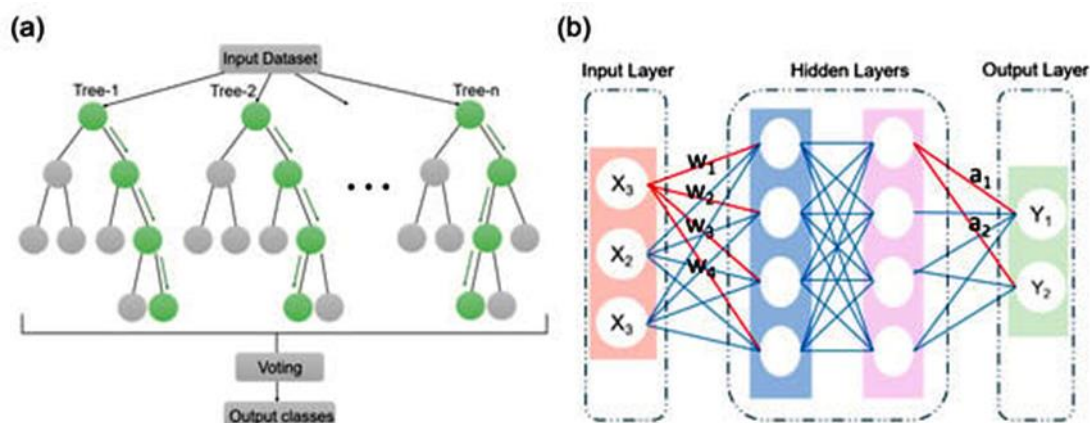


Рис. 3. (a) Принципиальная схема random forest; (b) принципиальная схема DNN.

XGBoost (eXtreme Gradient Boosting), предложенный Тяньци Ченом, представляет собой более эффективную системную реализацию градиентного бустинга на основе дерева решений с градиентным бустингом. XGBoost повышает точность прогнозирования за счет постоянного формирования новых деревьев решений для соответствия остаточным результатам прогнозирования, тем самым сокращая разрыв с фактическим значением. Целевая функция, определенная обучающей моделью XGBoost, включает функцию потерь и регулярный член. Функция потерь отражает степень соответствия модели данным, а разложение Тейлора второго порядка позволяет лучше судить о тенденции градиентного спуска, что приводит к более высокой точности и более быстрой сходимости. Обычный термин используется для управления сложностью модели. Целевая функция и модельная формула XGBoost могут быть выражены в виде формул (1) и (2):

$$\text{Obj} = + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (2)$$

Где $l(y_i, \hat{y}_i)$ представляет собой разницу между прогнозируемым значением $\hat{y}_i$ и истинное значение y_i из i -го образца, и $\Omega(f_k)$ представляет сложность k -е дерева решений. $f_k \in F$, где $F = f(x) = wq(x)$, каждый f_k соответствует независимой структуре дерева решений q и массы листьев w . w_i представляет вес i -го узла тетианского домена, и $wq(x)$ является ли оценка для образца x , т.е. модельный прогноз.

Глубокая нейронная сеть (DNN) состоит из входного слоя, скрытых слоев и выходного слоя, который передает данные на выходной уровень посредством прямого распространения. Веса и параметры смещения каждого слоя определяют передачу и преобразование данных. В процессе обратного распространения ошибка вычисляется с помощью функции потерь, и

ошибка распространяется обратно на каждый слой по цепному правилу, а функция потерь уменьшается путем обновления параметров. DNN использует нелинейное выражение скрытых слоев и функций активации для моделирования нелинейных взаимосвязей в сложных данных.

SHAP - это метод, основанный на концепциях теории игр, которые часто используются для объяснения результатов предсказаний модели DNN. Принципиальную схему DNN можно увидеть на рисунке 3б.

3.3. Оценка эффективности модели.

Точность, специфичность, полнота и коэффициент F1 — это четыре основных показателя, используемых для оценки эффективности модели. Формулы (3)–(6) содержат методики расчёта этих четырёх показателей [70], где TP, TN, FP и FN соответственно обозначают истинные положительные, истинные отрицательные, ложные положительные и ложные отрицательные случаи. TPR (доля истинных положительных результатов), также известная как чувствительность или полнота, представляет собой долю правильно предсказанных моделью положительных случаев. FPR (доля ложных положительных результатов) указывает на частоту ошибок модели при предсказании положительных случаев. Эти показатели имеют решающее значение для оценки эффективности модели, особенно при работе с несбалансированными наборами данных в рамках академических исследований.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$precision = \frac{TP}{FP + TP} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\ score = \frac{2 * precision * recall}{precision + recall} \quad (6)$$

3.4. Визуализация t-SNE.

Метод t-SNE (стохастическое приближение соседей с распределением по t-кривой), предложенный Лоренсом ван дер Маатеном и Джефффри Хинтоном, стал популярным инструментом для визуализации многомерных данных путём их преобразования в двумерные или трёхмерные данные. Хотя традиционные линейные методы (например, метод главных компонент и метод главных факторов) хорошо зарекомендовали себя в геохимических исследованиях, они основаны на линейных преобразованиях и могут не учитывать нелинейные взаимосвязи в сложных данных. Метод t-SNE основан на распределении Гаусса с

центром в каждой точке, вычислении сходства каждой логарифмической точки в многомерном пространстве, создании распределения вероятностей, а затем создании аналогичного распределения вероятностей в низкоразмерном пространстве. Алгоритм минимизирует разницу между двумя распределениями вероятностей с помощью таких алгоритмов, как градиентный спуск.

4. Результаты и обсуждение.

4.1. Обычные геохимические индикаторы.

Прежде чем применять методы машинного обучения для классификации, мы попытались использовать традиционные химические методы для выделения типов циркона. Паутинная диаграмма - это традиционный метод геохимического анализа, используемый для выявления различий в истощении и обогащении элементами между различными образцами. Были взяты случайные выборки из шести металлогенических зон и построены паутинные диаграммы на основе 16 элементов, как показано на рисунке 4, который иллюстрирует различия в содержании элементов циркона в разных террейнах и различия между рудными и безрудными цирконами. По сравнению с методами машинного обучения, паукообразные диаграммы фокусируются на истощении и обогащении микроэлементами (например, Hf, Ti, Ce и Eu). В целом можно видеть, что легких редкоземельных элементов (La, Ce, Nd, Sm и Eu) в плодородных цирконах относительно недостаточно, а тяжелых редкоземельных элементов (Gd, Tb, Dy, Ho, Y, Er, Tm, Yb и Lu) относительно много. Однако бесплодные цирконы в основном имеют одинаковые химические характеристики. Трудно отличить циркон только по степени истощения и обогащения микроэлементами. Машинное обучение в полной мере использует индексы 16 элементов для изучения сложных закономерностей и взаимосвязей, тем самым повышая точность классификации цирконов.

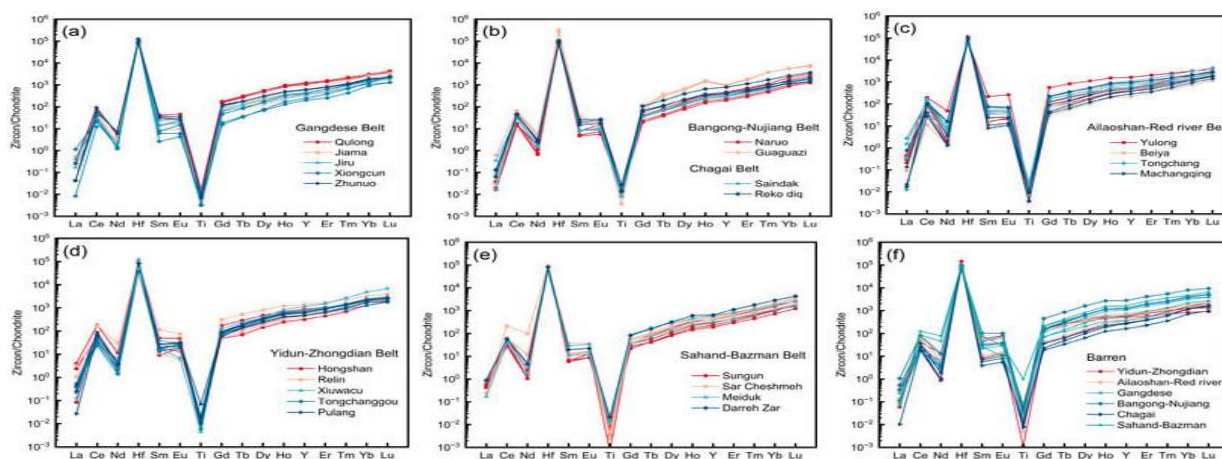


Рис. 4. Структура микроэлементов циркона с нормированным содержанием хондритов в цирконах из рудных образцов (a-e) и безрудных образцов (f) в тетианском домене.

Были предложены методы, основанные на индикаторах микроэлементов циркона, для оценки металлогенического потенциала порфировых месторождений Cu. Их исследование выявило характерные черты рудных магматических цирконов как $Eu/Eu^* > 0,3$; $10000 * (Eu/Eu^*)/Y > 1$; $(Ce/Nd)/Y > 0,01$; $Dy/Yb < 0,3$; $Ce/Ce^* > 100$. Были использованы образцы, связанные с медно-порфировым оруденением, из собранного набора данных zircon к этим методам и построили бинарные диаграммы рассеяния, как показано на рисунке 5, чтобы исследовать точность, FPR, TPR традиционных методов.

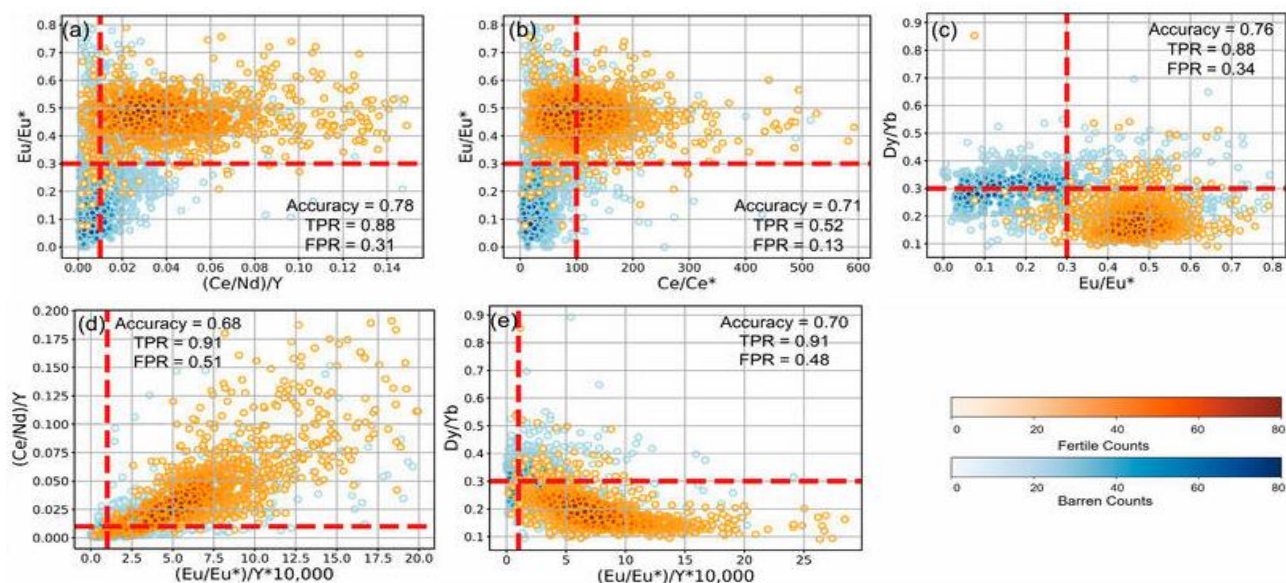
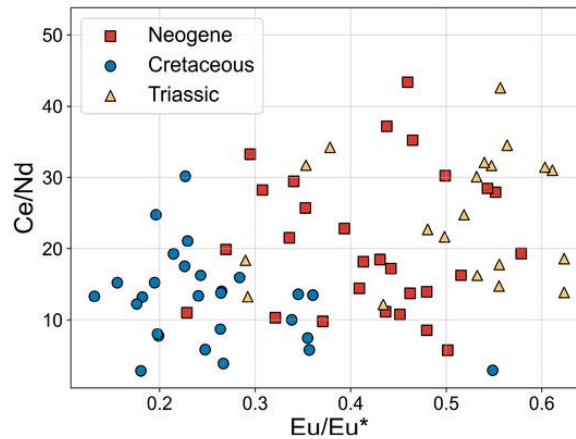


Рис. 5. Диаграммы рассеяния цирконов в порфировых Cu месторождениях, на основе традиционных методов разделения: (a) Eu/Eu^* по сравнению с $(Ce/Nd)/Y$; (b) Eu/Eu^* по сравнению с Ce/Ce^* ; (c) Dy/Yb по сравнению с Eu/Eu^* ; (d) $(Ce/Nd)/Y$ по сравнению с $(Eu/Eu^*)/Y * 10000$; (e) Dy/Yb против $(Eu/Eu^*)/Y * 10\,000$.

Рисунок 5 показывает, что классификационная модель, основанная на традиционных методах, достигла точности 0,68~0,78 и TPR 0,52~0,91. Дифференциация на основе Eu/Eu^* и $(Ce/Nd)/Y$ обеспечила наивысшую точность, достигнув 0,78. Метод дифференциации на основе Dy/Yb и $10\,000 * (Eu/Eu^*)/Y$, $(Ce/Nd)/Y$ и $10\,000 * (Eu/Eu^*)/Y$ обеспечил наивысший TPR, достигнув 0,91. Формирование и эволюция тетической металлогенической области в первую очередь обусловлены несколькими этапами субдукции и столкновения океанических плит в океане Тетис, при этом металлогенические пояса формировались в разное время. Минеральные месторождения в этом регионе в основном относятся к триасовому, меловому и неогеновому периодам. Чтобы выяснить, влияют ли цирконы из разных периодов на классификацию, были случайным образом отобраны и проанализированы с помощью традиционных методов построения графиков образцы цирконов из Cu-порфировых месторождений триасового, мелового и неогенового периодов. Рисунок 6 показывает, что классификация не зависит от возраста цирконов.



Рис/ 6. График зависимости Eu/Eu^* циркона от Ce/Nd из медно-порфировых месторождений разного возраста.

4.2. Построение моделей машинного обучения.

Для построения моделей машинного обучения разделили предварительно обработанные данные о микроэлементах циркона на обучающий набор, набор для проверки и независимый тестовый набор в соответствии с соотношением 8:1:1. Чтобы избежать значительных расхождений в размерах выборок между различными поясами минерализации, использовали стратифицированную выборку в scikit-learn на Python. Перед применением этих алгоритмов машинного обучения использовали класс Randomized Search CV в библиотеке scikit-learn на Python для выполнения метода случайного поиска для гиперпараметрической оптимизации. Затем оптимальные параметры уточняются с помощью 5-кратной перекрестной проверки, что обеспечивает более стабильную и достоверную оценку производительности модели посредством нескольких раундов обучения и тестирования. Параметры базовых классификаторов и их точность в наборе перекрестной проверки и тестовом наборе показаны в таблице 2.

Табл. 2.

Наилучшая комбинация параметров, выбранная методом случайного поиска, и точность классификатора на основе набора перекрестной проверки и набора независимых тестов.

Base Classifier	Selected Parameters	Cross-Validation Accuracy	Test Accuracy	AUC
XGBoost	learning_rate = 0.3 min_child_weight = 3 n_estimators = 150	87.0%	88.6%	0.985
Random Forest	min_samples_leaf = 1 min_samples_split = 2 n_estimators = 190	89.5%	91.0%	0.990
DNN	Activation = tanh hidden_layer_sizes = (70,10,10) Solver = adam	90.6%	90.8%	0.982

Значения площади под кривой (AUC) всех трех моделей превышают 0,95, что указывает на то, что эти классификаторы имеют прикладные значения, если они правильно устанавливают пороговые значения. Наконец, производительность моделей для

классификации положительных и отрицательных образцов оценивалась интуитивно с помощью нормализованной матрицы путаницы, как показано на рисунке 7.

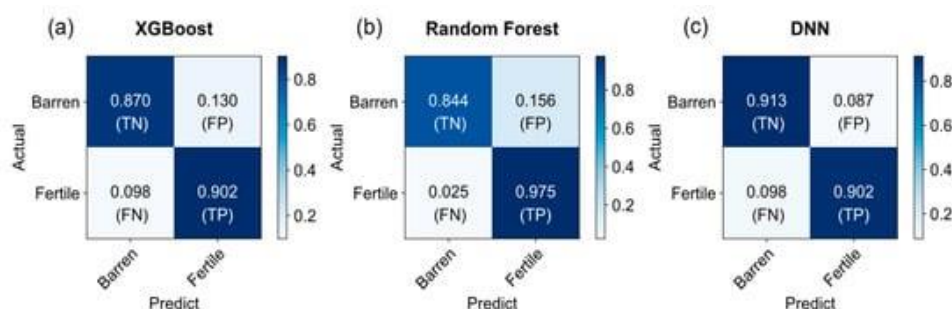


Рис. 7. Диаграммы матрицы путаницы тестового набора с использованием трех моделей машинного обучения.

Матрица путаницы показывает, что коэффициент TN для DNN составляет 0,913, что обеспечивает хорошую производительность при прогнозировании безрудных образцов. Коэффициент TP случайного леса составляет 0,975, что указывает на то, что случайный лес обладает хорошими показателями при прогнозировании плодородных образцов. С помощью моделей машинного обучения мы смогли всесторонне рассмотреть особенности всех 16 элементов. Таким образом, модель может извлекать информацию из более полных данных, выявляя тонкие различия и лежащие в основе таксономические правила. Этот метод значительно повышает надежность и точность классификации и обеспечивает более надежную поддержку для точной идентификации циркона.

Анализ признаков является важным шагом после создания модели классификации. Это помогает понять корреляцию между признаками и результатами классификации. Из рисунка 8 можно сделать вывод, что в качестве индикаторов можно отразить летучесть магматического кислорода; Ce и Eu играют важную роль в идентификации цирконов. Оба элемента являются многовалентными и могут взаимно преобразовываться при определенных окислительно-восстановительных условиях.

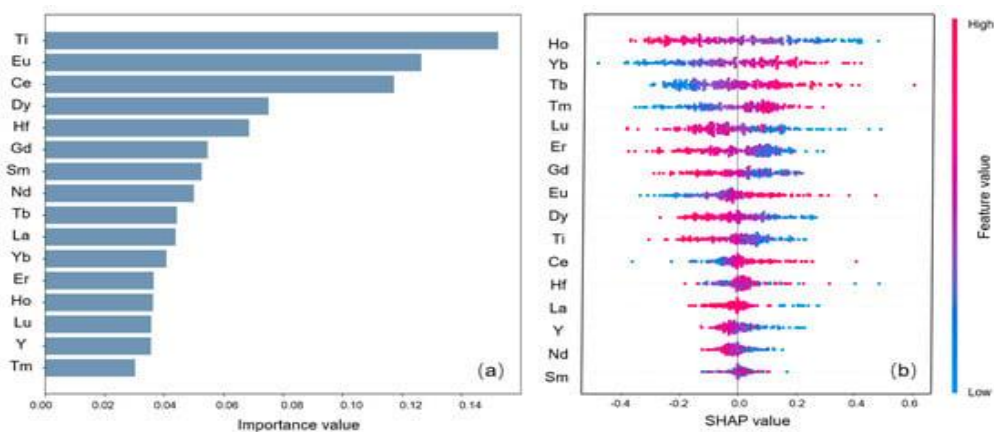


Рис. 8. Важность объекта. (a) ранжирование объектов леса случайным образом; (b) точечный график SHAP.

Образование медно-порфирового оруденения более благоприятно при более высоких степенях окисления. В водных расплавах роговая обманка кристаллизуется на ранних этапах магматической эволюции, снижая значение Y в остаточном расплаве. Значение Y, потерянное в остаточном расплаве, и значение Eu, не израсходованное при ранней кристаллизации плагиоклаза, приводят к высоким выбросам Eu и низким значениям Y в плодородных магматических свитах. В плодородных магматических свитах фракционирование роговой обманки имеет тенденцию преимущественно включать редкоземельный элемент Dy, а не тяжелый редкоземельный элемент Yb. Следовательно, в рудообразующих магматических системах значения Dy ниже, а значения Yb выше.

В предыдущих исследованиях элементы, ранее подтвержденные в качестве индикаторов рудных магматических цирконов, были подтверждены с помощью ранжирования важности признаков на основе машинного обучения, подтверждающего научную эффективность созданной модели. В этом исследовании считается, что содержание Ti, Lu, Hf и Ho в значительной степени связано с различием цирконов. Путем создания точечных диаграмм характерных элементов наблюдаются особенности распределения содержания характерных элементов в рудных цирконах, что показано на рисунке 9. Это указывает на то, что содержание Ti в плодородном магматическом цирконе в основном сосредоточено в пределах 0-5 ppm, содержание Eu сосредоточено в пределах 0-2 ppm, а содержание Ce в основном находится в диапазоне 10-60 ppm. Более того, содержание Dy в рудных магматических цирконах составляет 10-100 ppm, содержание Lu - 20-100 ppm, содержание Hf - 8000-12,500 ppm и содержание Yb - 200-500 ppm.

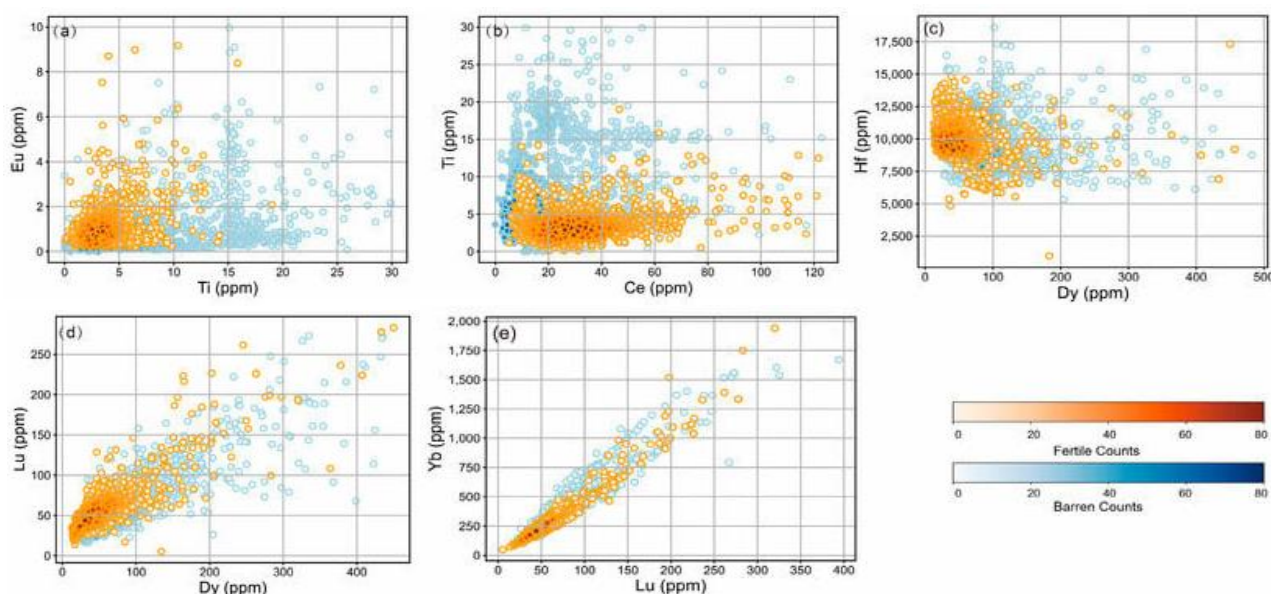


Рис. 9. Точечные графики содержания микроэлементов в цирконе:
(а) Es по сравнению с Ti; (b) Ti против Ce; (c) Hf против Dy; (d) Lu против Dy; (e) Yb против Lu.

Однако в настоящее время нет опубликованных выводов относительно роли и влияния этих элементов на механизмы минерализации, что требует дальнейших поисков и исследований в будущем.

4.3. Применение модели на порфировых месторождениях в центре Тетиса.

В центре Тетиса расположены Турецкие месторождения порфиров - три основных металлогенических пояса с севера на юг: Понтиды, Анатолиды и Пограничные складки. Среди них Тавшанлинский пояс характеризуется концентрацией порфировых систем эоценового периода, а его тектоническая среда в первую очередь связана с известково-щелочным магматизмом, возникшим в результате столкновения плит, и щелочным магматизмом, возникшим в результате постколлизийных процессов. Для исследования были использованы данные по циркону из массива горных пород, в котором в пределах пояса находится медно-молибденовое порфировое месторождение Сарычайырыяйла. На полуострове Бига в основном развита порфировая металлогеническая система, сформировавшаяся в период от эоцена до олигоцена, а образование порфировых месторождений связано с известково-щелочными вулканогенно-интрузивными породами. Тестовые образцы включают вулканические породы из района медно-молибденового месторождения Халилага, а также вулканические породы, образовавшиеся в эпоху минерализации порфирового золотого месторождения Кышладаг, и магматические породы, образовавшиеся в эпоху минерализации порфирового медно-молибденового месторождения Пынарбаши (расположение месторождений показано на рис. 1). Эти образцы получены из различных типов минерализованных порфировых образований, в том числе из рудных и безрудных. Микроэлементы циркона прошли этапы предварительной обработки, идентичные тем, которые использовались при создании модели. Затем эти предварительно обработанные данные были использованы в качестве входных данных для прогнозирования в классификаторах, и полученные результаты представлены в таблице 3.

Результаты классификации random forest и XGBoost вполне согласуются без существенных отклонений. Все три алгоритма, основанные на разных принципах, могут точно идентифицировать бесплодные образцы, например, растекшуюся лаву и мафический анклав. На основе результатов случайного леса были выбраны элементы с более высоким рейтингом Се, Еи, Тi и Ду - в соответствии с ранжированием важности признаков из классификатора случайного леса — в качестве горизонтальной и вертикальной координатных осей для построения бинарных графиков. Из рисунка 10 видно, что диапазон значений Тi и Еи для фертильных цирконов сосредоточен, в то время как значения Се и Ду разбросаны.

Источники образцов и результаты прогнозирования цирконов.

Deposit	Sample	Lithology	Prediction (0-Barren; 1-Fertile)		
			RF	XGBoost	DNN
Saricayirayla Porphyry Cu-Mo	18UD08-2	Granite	0	0	1
	18MU03	Granodiorite	0	0	1
Halilağa Porphyry Cu-Mo	18HA10-2	Andesite	1	1	1
	18HA02-1	Rhyodacite	0	0	0
	18HA03	Streatched Lava	0	0	0
	18HA06-1	Dacite	1	1	1
	18HA04-1	Feldsparphyric Rock	1	1	1
	18HA05-1	Andesite	0	0	0
Kisladağ Porphyry Au	18US12-3	Trachyandensite	0	0	1
	18US10-1	Andesite	1	1	1
	18US08-1	Andesite	1	1	1
	18US09-1	Dacite	0	0	0
	18US05-1	Dacite	0	0	1
	18US07-2	Dacite	0	0	1
	18US06-2	Dacite	1	1	1
Pınarbaşı Porphyry Mo-Cu	18GD02-1C	Granite	1	1	0
	18GD07-5	Granite	1	1	1
	18GD02-1D	Mafic Enclave	0	0	0
	18GD04-3	Monzonite	1	0	0
	18GD06-2	Granite	0	1	0

Независимо от того, является ли это рудный или безрудный циркон, распределение этих важных элементов неравномерно. Это также указывает на то, что, хотя эти элементы оказывают важное влияние на случайные решения о лесах, случайная классификация лесов не основана на отдельных элементах для классификации, а извлекается из всех признаков, что делает ее более эффективной, чем традиционные методы классификации с использованием бинарных графиков.

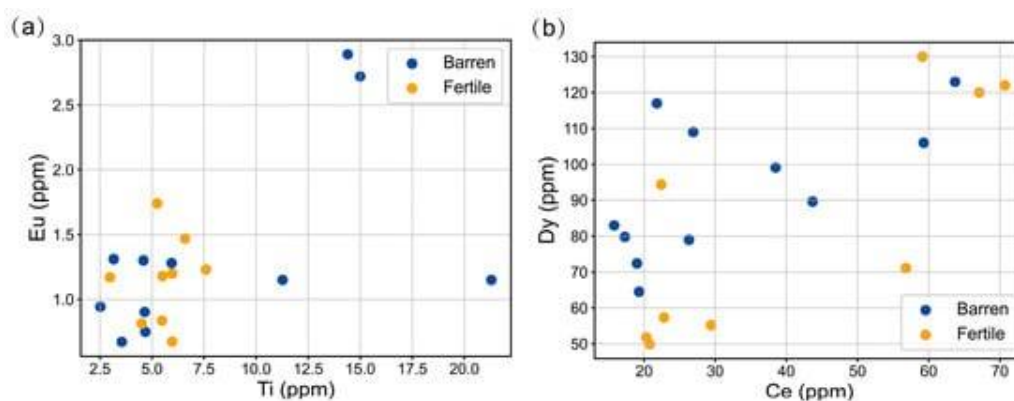


Рис. 10. Диаграммы рассеяния цирконов: (а) ЕС против Ti; (b) Dy в сравнении с Ce. (результаты классификации - на основе случайного леса).

4.4. Классификация типов порфировых месторождений.

Традиционные модели классификации, основанные исключительно на одноэлементных подходах, показали более низкую точность в практических приложениях. Основываясь на данных о большом количестве магматического циркона в исследуемом регионе, используются методы машинного обучения для исследования того, существуют ли систематические различия в микроэлементах цирконов среди различных типов порфирового оруденения. В

данном исследовании месторождения подразделяются на шесть типов: богатые Mo, Au, Cu, Cu-Mo, Cu-Au и Cu-Mo-Au. Они включают 648 точек данных по циркону. Чтобы обеспечить более четкую интерпретацию результатов классификации, включены другие соотношения элементов, имеющие металлогеническое значение, такие как соотношения Th/U, Lu/Hf и Dy/Yb. Отношение Th/U служит показателем степени метаморфизма в цирконе. Система изотопов Lu/Hf играет важную роль в установлении сроков и продолжительности металлогенических событий. Кроме того, отношение Dy/Yb используется для оценки рудного потенциала медно-порфировых месторождений.

Была выбрана модель случайного леса для классификации типов месторождений. Обучающий набор, проверочный набор и тестовый набор были разделены в соотношении 8:1:1. Из-за значительных различий в количестве цирконов из отдельных месторождений в наборе данных внедрили выборочную выборку, чтобы сбалансировать набор данных во время разделения, гарантируя, что пропорции образцов из разных типов месторождений одинаковы для обучения и тестирования. Рекурсивное устранение признаков (RFE) использовалось для предотвращения интерференции между элементами признаков. Наконец, для создания классификатора были отобраны 15 показателей: Y, Ce, Sm, Eu, Gd, Tb, Ho, Er, Tm, Lu, Hf, U, Dy/Yb, Lu/Hf и Th/U.

Отчет о классификации представлен в таблице 4, а ранжирование по важности признаков показано на рисунке 11.

Табл. 4.

Классификация типов месторождений на основе модели случайного леса.

Type	Precision	Recall	F1-Score	Support
Au-Rich	0.80	0.84	0.82	19
Cu	0.95	0.95	0.95	22
Cu-Au	1.00	0.80	0.89	10
Cu-Mo	0.80	0.67	0.73	18
Cu-Mo-Au	0.86	0.93	0.89	45
Mo-Rich	0.88	0.88	0.88	16

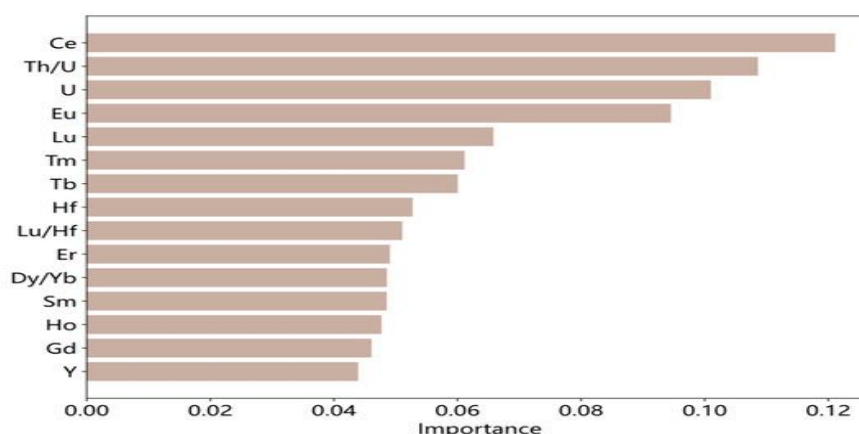


Рис. 11. Рейтинг важности признаков на основе случайного леса.

На основе метода t-SNE были сокращены и визуализированы химические данные циркона из различных типов порфировых месторождений с разной минерализацией, и окончательные результаты представлены на рисунке 12.

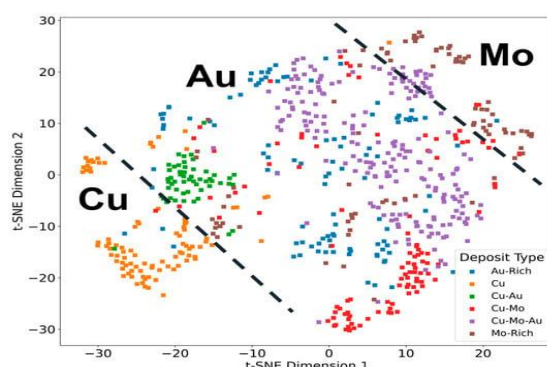


Рис. 12. Диаграммы рассеяния t-SNE для различных типов оруденения.

Можно заметить, что распределение данных по элементам постепенно переходит от порфировых месторождений с высоким содержанием меди к месторождениям с высоким содержанием золота и молибдена. Порфировые месторождения с высоким содержанием меди, молибдена и золота широко распространены в центральной части месторождений с высоким содержанием меди и молибдена. Порфировые медно-золотые месторождения расположены вблизи границы между месторождениями меди и золота.

Рисунок 11 показывает, что Ce, Th/U, U, Eu, Lu и Tm вносят значительный вклад в классификацию типов месторождений. Используя метод t-SNE, генерируются точечные диаграммы, где цветом обозначены концентрации элементов. Эта визуализация изображена на рисунке 13.

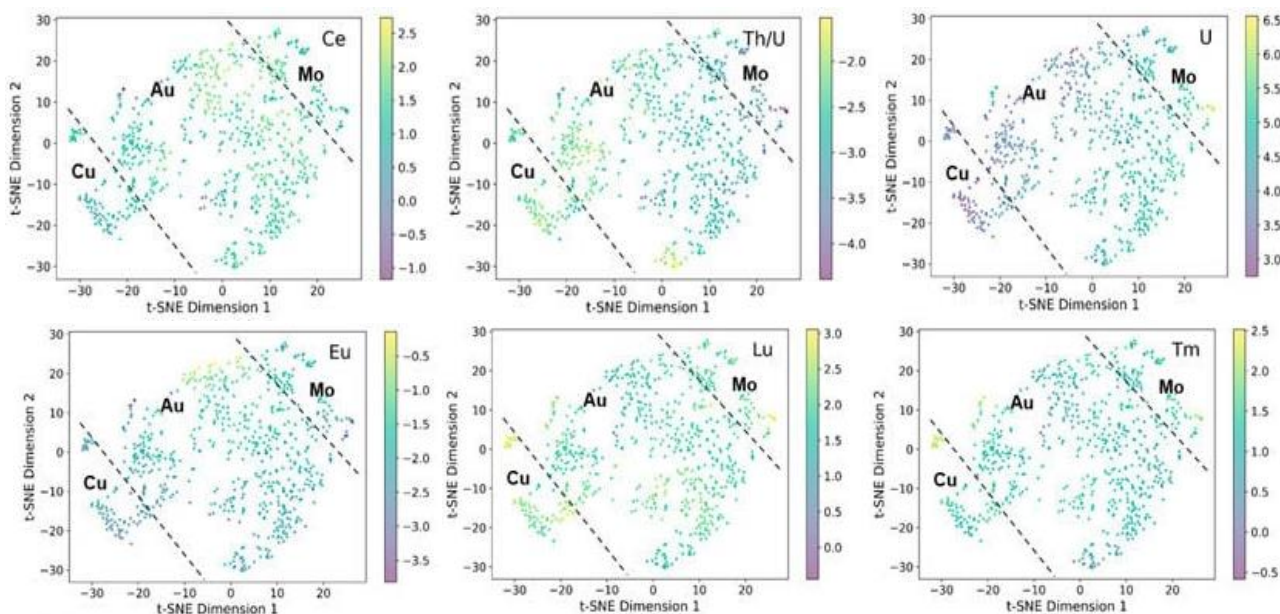


Рис. 13. Диаграмма рассеяния t-SNE для микроэлементов циркона из различных типов порфировых месторождений.

Анализируя визуальные данные о содержании элементов, можно заметить, что различные типы месторождений могут иметь разные характеристики с точки зрения содержания микроэлементов. Содержание Се в цирконах из порфировых месторождений Cu, Mo и Cu-Au заметно ниже, чем в порфировых месторождениях Au и Cu-Mo-Au. Кроме того, отношение Th/U в цирконах демонстрирует тенденции: по мере перехода от порфировидных месторождений Cu к месторождениям, богатым Au, к месторождениям, богатым Mo, отношение Th/U постепенно уменьшается, а концентрация U соответственно возрастает. Анализ концентраций Eu и Lu показывает, что богатые Au порфировые месторождения имеют относительно высокое содержание Eu и более низкое содержание Lu. По сравнению с пятью основными элементами концентрация Tm не демонстрирует значительных различий в различных типах месторождений. Эти результаты могут быть связаны с геологическими характеристиками этих месторождений и механизмами формирования.

4.5. Обсуждение.

По сравнению с традиционными химическими методами, методы машинного обучения демонстрируют более высокую эффективность при создании классификаторов циркона и изучении различных типов минерализации порфировых месторождений. Традиционные геохимические методы, как правило, основаны на лабораторных анализах, которые могут быть ограничены по времени, стоимости и объёму данных. Более того, использование 2–3 элементов для дифференциации может привести к размытым границам классификации. С другой стороны, методы машинного обучения могут обрабатывать большие объёмы сложных данных, выявляя скрытые закономерности и признаки для эффективной автоматизации классификации и прогнозирования, а также обеспечивая более высокую точность и процент истинно положительных результатов.

Однако в ходе исследования были также выявлены ограничения в применении машинного обучения. Некоторые геологические события или типы минерализации могут быть более редкими, и для них может быть сложнее получить достаточно данных, что приводит к несбалансированным наборам обучающих данных. Этот дисбаланс может повлиять на оценку моделей машинного обучения, что потенциально может привести к недостаточному прогнозированию для классов меньшинств. Кроме того, геологические данные часто демонстрируют пространственные и временные корреляции, когда данные из соседних регионов или разных периодов времени могут демонстрировать взаимозависимости и вариации. Традиционные модели машинного обучения могут испытывать трудности с непосредственной обработкой пространственных и временных структур в геологических данных, что требует дополнительных методов предварительной обработки и создания

признаков. Поэтому преодоление этих ограничений станет важнейшей задачей для геологов в будущих исследованиях.

5. Выводы.

В этом исследовании предлагаются методы машинного обучения для классификации цирконов с использованием данных о микроэлементах. В нём изучаются характеристики микроэлементов в рудных цирконах из порфировых месторождений в пределах Тетийского региона, что позволяет сделать следующие выводы:

1. Модели машинного обучения могут эффективно отличать магматические цирконы в порфировых образованиях, точность случайного леса, XGBoost и DNN составляет 91%, 88,6% и 90,8% соответственно.

2. Результаты ранжирования характеристик показывают, что Ce, Eu, Dy и Yb оказывают значительное влияние на определение рудоносности цирконов. В дополнение к элементам, которые были подтверждены в предыдущих исследованиях, в ходе этого исследования было обнаружено, что Ti, Lu, Hf, Ho и Tb также играют важную роль в определении рудоносности цирконов. Конкретные механизмы их воздействия требуют дальнейшего изучения.

3. Визуализация t-SNE представляет собой эффективный метод изучения различий в характеристиках элементов в порфировых месторождениях с разной минерализацией. Исследование показало, что в месторождениях с высоким содержанием меди, молибдена и смешанной минерализацией значения Ce и Eu относительно ниже по сравнению с месторождениями с высоким содержанием золота, а значения Lu выше. Отношения Th/U последовательно снижаются в месторождениях с высоким содержанием меди, золота и молибдена

VII. МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ РУДНЫХ МЕСТОРОЖДЕНИЙ НА ОСНОВЕ МИКРОЭЛЕМЕНТОВ В МАГНЕТИТЕ (рудный район Осса-Морена, Южная Иберия, Португалия) [3].

1. Введение.

Одной из сложных задач при разведке рудных месторождений является определение модели, которая наилучшим образом соответствует определенному типу месторождения, на основе комбинации нескольких источников данных. Точное определение таких моделей является ключевым фактором для проведения ГРП, помогая определить применимые методологии и стратегии, тем самым экономя время и деньги.

Магнетит - распространенный минерал, присутствующий в различных металлогенических средах, поэтому он является удобной мишенью для выполнения такой задачи. Его микроэлементный состав предоставляет ценную информацию для определения условий и механизмов, связанных с формированием оруденения, что в конечном итоге позволяет провести различие между безрудными и минерализованными участками. С учетом этих целей данные о микроэлементах магнетита были использованы при классификации рудных месторождений. Тем не менее, химический анализ магнетита часто заканчивается получением больших наборов данных с многочисленными переменными, что создает проблему для тщательного изучения результатов.

Было доказано, что определение состава магнетита с помощью масс-спектрометрии с лазерной абляцией и индуктивно связанной плазмой (LA-ICP-MS) способствует правильной классификации рудных месторождений с использованием обычных дискриминационных диаграмм.

Классический подход к использованию состава магнетита в качестве геохимического инструмента основан на многоэтапном анализе данных, предварительном разделении некоторых типов месторождений и применении дискриминантных диаграмм, которые обеспечивают сходство с определенным типом рудных месторождений.

Методы сокращения объема данных, такие как анализ главных компонент (PCA), дали хорошие результаты. Недавно для классификации рудных месторождений использовалось применение методов классификации под руководством машинного обучения (ML) к микроэлементному составу магнетита. Однако эти модели обычно зависят от больших наборов данных (например, ~ 17 тыс. наблюдений) для получения результатов.

Использование ML для классификации рудных месторождений на основе данных о микроэлементах магнетита представляет огромное преимущество по сравнению с классификациями, основанными на обычных точечных диаграммах. Методы ML включают

многомерные характеристики наборов данных, обеспечивая более надежную проверку и, следовательно, лучшую агрегацию различных групп, которые могут быть использованы для определения геологических моделей.

Для проверки этого утверждения применяется производительность шести моделей классификации ML, а результаты обсуждаются с использованием справочного набора данных с гораздо меньшим количеством значений (271 наблюдение) анализа микроэлементов магнетита, опубликованного разными авторами. Затем модели применяются к набору данных о составе магнетита из области, включающей несколько месторождений железа из португальского сектора зоны Осса-Морена (OMZ), и результаты обсуждаются на основе известных геологических моделей. Данные о магнетите, собранные в наборе данных тематического исследования, относятся к трем вулканогенным месторождениям (Монж, Вале-да-Арка, Серринья) из железорудного комплекса Монтемор-о-Ново (MIC), дополненным тремя железо-скарновыми месторождениями (Алвито, Азенхас и Орада).

В этом исследовании представлен новый подход к классификации рудных месторождений в Португалии путем применения шести моделей классификации ML к данным анализа микроэлементов магнетита.

Важно отметить, что, хотя размер набора данных может быть ограничен, были предприняты усилия для обеспечения его репрезентативности и разнообразия, отражающего основные характеристики магнетитов различного происхождения и эволюции. Кроме того, для оценки производительности модели были использованы строгие процедуры оценки и валидации, обеспечивающие уверенность в ее способности обобщать полученные результаты на новых образцах.

2. Классический подход.

Целью классического подхода является проведение различия между несколькими типами рудных месторождений, включая железоокисно-медно-золотые (IOCG), Кируна-апатит-магнетитовые, полосчатые железистые образования (BIF), порфировые Cu, Fe–Cu скарны, Fe–Ti, V, Cr, Ni-Cu-PGE, Cu-Zn-Pb вулканогенный массивный сульфид (VMS), а также архейские Au-Cu порфировые и Opemiska Cu жилы. Их анализ основан на использовании микроэлементов в магнетите для классификации каждого типа месторождений. Был предложен трехэтапный подход к классификации различных типов месторождений на основе микроэлементов в магнетитах.

- *Месторождения никеля:* первым шагом в подходе было идентифицировать образцы магнетита, связанные с месторождениями Ni. Эти месторождения содержат магнетит с

высокими концентрациями Ni и Cu. Путем построения графика концентраций Ni+Cr по сравнению с Si+Mg, смогли отличить месторождения Ni-Cu-PGE от других типов месторождений.

- *Вулканогенные массивные сульфидные месторождения (VMS)*: характеризуются магнетитом с высокими концентрациями Zn. Построив графики концентраций Al/(Zn+Ca) в сравнении с Cu/(Si+Ca), смогли отличить месторождения VMS от других типов.

- *Другие типы месторождений*: включая IOA, IOCG, железистые формации типа Алгома, порфировые месторождения меди и скарновые месторождения. Результаты были достигнуты путем анализа концентраций микроэлементов в магнетите, таких как Ti, V, Cr, Mn, Co, Cu, Ge, Y, Zr, Nb, Mo, Sn, Hf, Ta, W, Pb, Th и U. Построения графика зависимости Ca+Al+Mn и Ti+V elements смогли создать дискриминантные диаграммы, которые выявили отдельные кластеры или тенденции для разных типов месторождений.

Систематические различия в концентрации минорных и микроэлементов в составе гидротермального магнетита из разных типов месторождений и магматического магнетита из минерализованных и безрудных вмещающих пород. Основными факторами, определяющими изменения состава магнетита, являются температура, состав флюидов, летучесть кислорода и серы, активность силикатов и сульфидов, буферность вмещающих пород, процессы восстановления равновесия и внутренний кристаллографический контроль. Ключевые элементы-дискриминаторы, такие как Mg, Al, Ti, V, Cr, Mn, Co, Ni, Zn, Ga и Sn, играют важную роль в отличии магнетита из различных типов месторождений и вмещающих пород.

Комбинация дискриминантных графиков, таких как Ti+V и Al+Mn, а также боковые и нитевидные графики элементов магнетита позволяют эффективно отличать магнетит от бифов, жил Ag-Pb-Zn, порфировых и связанных с ними скарновых месторождений, месторождений Climax-Mo и безрудных магматических вмещающих пород. Низкие уровни Al, Ti, V, Cr, Co, Zn и Ga особенно важны для диагностики магнетита из бифов, Mg-скарновых и Ag-Pb-Zn жил. Кроме того, хотя магматический и гидротермальный магнетит из порфировых месторождений имеет схожие составы, их можно дифференцировать на основе концентраций Cr, Co, Ni и Ga, причем концентрации Ga в магматическом магнетите неизменно выше, чем в гидротермальных проявлениях.

Данное исследование поддерживает и уточняет схему классификации для выделения магнетита из различных типов рудных месторождений.

3. Машинное обучение (ML).

Был разработан классификатор случайного леса (RF) для идентификации типов рудных месторождений на основе концентраций Co, Ni, Cu, Zn, As, Mo, Ag, Sb, Te, Tl и Pb в пирите. Классификатор продемонстрировал высокую точность как для тестовых, так и для «слепых» данных тестирования, достигнув общей точности 94,5% и 93,9% соответственно, без учета неубедительных анализов. Случайные лесные классификаторы, полученные на основе химического состава отдельных минералов, обладают потенциалом полезного инструмента геохимической разведки. Тем не менее, они подчеркивают, что это следует рассматривать как предварительный положительный результат, и дополнительные разновидности пирита должны быть включены в классификатор перед широким применением при поисках полезных ископаемых. Более того, классификатор следует рассматривать как один из многих инструментов, а не как отдельный метод классификации.

RF-классификатор, используемый для классификации месторождений IOCG (оксид железа–медь–золото) и IOA (оксид железа–апатит) на основе химического состава оксидов железа, измеренного с помощью электронно-зондового микроанализатора (EPMA) и LA-ICP-MS, показал, что, несмотря на сложный и неравномерный геохимический состав оксидов железа, месторождения IOCG и IOA можно различить.

Было обнаружено, что эффективность RF-модели с использованием набора данных LA-ICP-MS превосходит классификационную модель, основанную на данных EPMA, что указывает на то, что включение большего количества геохимических переменных и данных более высокого качества выгодно для различения месторождений IOA и IOCG. Более того, были определены ключевые элементы для отличия IOCG от месторождений IOA: V, Mg и Mn в данных EPMA и Si, Mg и V в данных LA-ICP-MS.

Результаты анализа образцов магнетита с высоким содержанием Ti из различных типов месторождений, включая магматические, порфировые, IOA и IOCG, продемонстрировали, что ML при правильном применении весьма эффективно для различения высоко титанистого оруденения. Магнетит из месторождений IOA обладает элементарными характеристиками, сходными с высокотемпературным гидротермальным магнетитом из IOCG и порфировых систем, в то же время отчетливо отличаясь от магматического магнетита. Это говорит о том, что магнетит IOA, скорее всего, гидротермального происхождения. Предложена новая дискриминантная диаграмма для магнетита с высоким содержанием Ti, $\lg(\text{Al}) + \lg(\text{Ti}) + \lg(\text{V})$ в сравнении с $\lg(\text{Mn}) / [\lg(\text{Co}) + \lg(\text{Mg})]$. Метод опорных векторов (SVM) показал, что эта новая диаграмма имеет точность 97,6%. Она эффективно отражает характеристики распределения данных, наблюдаемые в методах уменьшения размерности, таких как проекции PCA и t-SNE,

тем самым предлагая ценный инструмент для различения магнетита с высоким содержанием Ti различного происхождения.

Был использован обширный набор данных (более чем 17 тыс. наблюдений из 303 различных типов месторождений, включая BIF, Fe-Ti, IOCG, IOA, Ni-Cu-PGE, порфир, VMS, скарн и V). Протестировав производительность трех моделей ML: наивного Байеса, k-ближайшего соседа и случайного леса, и пришли к выводу, что случайный лес является лучшей моделью для большинства типов месторождений и что он может быть успешно использован в поисково-разведочных кампаниях для выявления различных типов рудных месторождений.

4. Тематическое исследование рудной зоны Осса-Морена (OMZ).

Тематическое исследование включало данные LA-ICP-MS о микроэлементном составе магнетита из шести месторождений железа (рис. 1а). Выбранные месторождения относятся к металлогенической провинции Монтемор-о-Ново—Фикальо, подразделению пояса Эворы-Арасены.

Железорудный комплекс Монтемор-о-Ново.

Три из выбранных месторождений, Монж, Вале-да-Арка и Серринья, относятся к вулканогенно-экспляционному комплексу Монтемор-о-Ново. Он включает в себя несколько месторождений железа (рис. 1b), их основной металлогенический вклад связан со стадией внутриконтинентального рифтогенеза раннего—среднего кембрия (варисканский цикл). Железные руды в основном связаны с кальцитно-доломитовыми породами из формации Монфурадо, которые подверглись метасоматозу на поздних стадиях варисканского цикла. Было предложено комбинированное вулканогенно-«выдыхательное» происхождение (SEDEX-VMS) месторождений, при этом отложение руды было связано с поступлением гидротермальных флюидов на ранних стадиях океанизации Рейка.

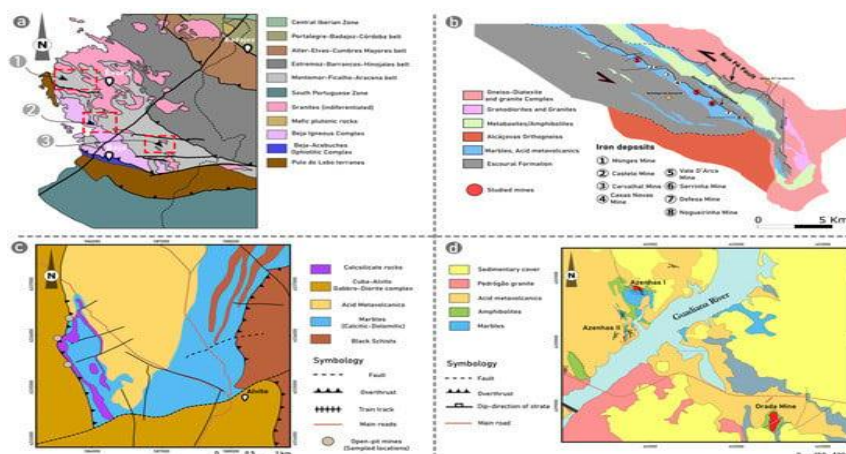


Рис. 1. Геологические схемы месторождений:

(а). зоны Осса-Морена, (б). зоны сдвига Монтемор. (с). м-ния Алвито (д). м-ний Azenhas I, II и Орада.

Месторождение Алвито (рис. 1с) представляет собой железо-скарновое месторождение, связанное с залеганием габбро-диоритового массива, входящего в магматический комплекс Бежа (рис. 1а). Размещение этого магматического тела началось в $350 \text{ млн} \pm 5 \text{ млн}$ лет назад, и в районе Алвито оно контактирует с карбонатными породами и образует экзоскарн, в котором содержится большая часть массивного магнетита. Формирование руды относится к ретроградным стадиям метасоматоза, а магнетит обнаружен в разрезах, богатых диопсидом–геденбергитом.

Месторождения Азенхас–Орада (рис. 1d) представляют собой комплекс железорудных тел, расположенных в метасоматизированных аллохтонных амфиболитах, которые тектонически вкраплены в карбонаты кембрия. Их происхождение приписывают существованию длительного температурного градиента, связанного с тектоническим сжатием. Считается, что такое тектоническое расположение способствовало отложению магнетита, связанного с метасоматизмом амфиболитов. Анализ микроэлементов магнетита показал, что генезис этих месторождений связан с высокотемпературными магматическими процессами.

5. Материалы и методы.

Для проведения анализа использовали язык программирования R и набор хорошо зарекомендовавших себя пакетов, предназначенных для реализации методов машинного обучения. Пакет "caret" служит основой для анализа, предоставляя упрощенный интерфейс для предварительной обработки данных, обучения модели и оценки производительности. Для задач классификации использовали пакеты 'random Forest', 'naive bayes', 'nnet' и 'e1071', которые позволяют реализовать предложенные классификаторы ML. Используя возможности этих пакетов, исследованы различные модели, сравнена их производительность и в конечном итоге определены наиболее подходящие подходы для достижения целей.

5.1. Данные и предварительная обработка.

Был собран справочный набор данных LA-ICP-MS анализа магнетита из месторождений полосчатой железной формации (BIF); месторождений оксидов железа, меди и золота (IOCG); месторождений порфирового типа; железо-окисно-апатитовых месторождений (IOA) типа Кируна; скарновых месторождений и вулканогенных массивных сульфидов (VMS) озера Изок VMS.

Этот справочный набор данных для процесса самооценки был разделен на две группы: (1) случайная выборка справочного набора данных, получившая название Train Dataset и содержащая 70% данных. В эту группу входят 190 измерений магнетита, которые используются для обучения различных моделей ML; (2) остальные данные, тестовый набор данных,

соответствуют оставшимся 30% эталонного набора данных, т.е. 81 наблюдение магнетита, которые используются для оценки результатов классификатора ML.

Был использован справочный набор данных, то есть 271 значение магнетита. Во всех этих наборах данных использовались девять элементов (Mg, Al, Ti, Mn, V, Co, Ni, Zn, Ga).

Чтобы обойти ограничение цензурированных данных в таких больших наборах, замена анализа ниже предела обнаружения оборудования была выполнена с использованием алгоритма логарифмической максимизации математического ожидания (EM) с помощью функции *lrEM* в пакете *zCompositions* в среде языка программирования R. Этот пакет позволяет проверять цензурированные значения с помощью функции *zPatterns*, так что переменные с более чем 40% отсутствующих значений могут быть исключены из набора данных.

5.2. Уменьшение размеров.

Анализ главных компонент (PCA) - это метод уменьшения размерности, который при применении к анализу состава магнетита может дать представление о различии рудных месторождений. PCA может помочь более эффективно визуализировать и интерпретировать сложные, многомерные геохимические данные. Элементные концентрации магнетита преобразуются в новый набор некоррелированных переменных - основных компонент (PCS). Эти ПК представляют собой линейные комбинации исходных переменных и упорядочены таким образом, что первый главный компонент (PC_1) отражает наибольшую дисперсию в наборе данных, за которым следует второй главный компонент (PC_2) и так далее. Сохраняя только несколько компьютеров, которые объясняют значительную часть различий, PCA эффективно уменьшает размерность данных, сохраняя при этом важную информацию и закономерности.

При использовании состава магнетита для выделения рудных месторождений PCA может помочь выявить геохимические тенденции и ассоциации между различными типами месторождений на основе основных PCS. Нанося оценки образцов на первые несколько компьютеров, можно визуализировать кластеризацию и разделение различных типов месторождений в уменьшенном пространстве измерений. Эта визуализация помогает понять взаимосвязи между составом магнетита и типами рудных месторождений, тем самым способствуя совершенствованию геологических моделей, применяемых к определенным геологическим объектам.

5.3. Машинное обучение.

Random forest (RF) - это метод коллективного обучения, который работает путем построения нескольких деревьев решений на этапе обучения. Результат модели случайного леса определяется путем агрегирования прогнозов по отдельным деревьям, что помогает

минимизировать переоснащение и улучшает обобщение модели. Этот метод особенно эффективен при обработке больших наборов данных с большим количеством признаков, поскольку он может эффективно обрабатывать пропущенные значения, выбросы и несбалансированные данные. Для этой классификации было определено 500 деревьев принятия решений. Метод перекрестной проверки для повторной выборки включал 10-кратное разбиение данных.

Naïve Bayes (NB) - вероятностный алгоритм машинного обучения, основанный на применении теоремы Байеса с допущением независимости между признаками. Этот метод особенно эффективен для обработки больших наборов данных. Несмотря на упрощающее допущение о независимости объектов, наивный Байес часто хорошо работает на практике, особенно когда допущение о независимости приблизительно верно или когда целью модели является ранжирование, а не прогнозирование абсолютных вероятностей.

Машины опорных векторов (SVM) - это класс алгоритмов контролируемого обучения, используемых как для задач классификации, так и для регрессии. Основная идея SVM заключается в нахождении оптимальной гиперплоскости, которая наилучшим образом разделяет точки данных на разные классы с максимально возможным разрывом между классами. В случае нелинейно разделяемых данных SVM использует функции ядра, такие как радиальная базисная функция (RBF). Эта функция преобразует данные в многомерное пространство, где становится возможным линейное разделение. SVM известна своей устойчивостью к переоснащению и способностью эффективно обрабатывать многомерные данные. Однако выбор соответствующей функции ядра и настройка гиперпараметров могут быть сложными, а алгоритм может быть дорогостоящим с точки зрения вычислений для больших наборов данных. SVM эффективно использовалась при классификации рудных месторождений с использованием данных LA-ICP-MS по магнетиту, а также при прогнозировании поисков.

Мультиномиальная логистическая регрессия (MLR), также известная как регрессия SoftMax или классификатор MaxEnt, является расширением логистической регрессии, которая позволяет прогнозировать несколько классов, а не только бинарные результаты. Этот метод особенно полезен для многоклассовых задач классификации, где целевая переменная имеет более двух возможных категорий. Мультиномиальная логистическая регрессия работает путем оценки вероятностей входного экземпляра, принадлежащего каждому классу, с использованием функции SoftMax, которая нормализует линейную комбинацию входных данных в распределение вероятностей по классам. Затем прогнозируемый класс определяется путем выбора класса с наибольшей вероятностью. Сохраняя простоту и интерпретируемость

логистической регрессии, мультиномиальная логистическая регрессия позволяет модели эффективно решать более сложные задачи классификации.

К-ближайший сосед (KNN) - это непараметрический алгоритм обучения на основе экземпляра, используемый как для задач классификации, так и для регрессии. Он работает путем сравнения нового, немаркированного экземпляра с набором помеченных экземпляров, определения k ближайших экземпляров в соответствии с метрикой расстояния, а затем прогнозирования метки на основе класса большинства или средневзвешенного значения меток соседей. KNN известен своей простотой, интерпретируемостью и способностью адаптироваться к новым данным, но на его производительность могут негативно повлиять многомерные данные или большие наборы данных из-за «проклятия» размерности и увеличения вычислительных затрат.

Искусственные нейронные сети (ANNS) представляют собой семейство моделей машинного обучения, основанных на структуре и функциях биологических нейронных сетей. ANN состоит из взаимосвязанных слоев искусственных нейронов, которые обрабатывают входные данные и корректируют их веса на основе ошибки между прогнозируемыми выходными данными и фактическими целевыми значениями. ANN обладают высокой гибкостью и могут изучать сложные нелинейные взаимосвязи, что делает их пригодными для широкого круга задач; однако они могут быть дорогостоящими с точки зрения вычислений и часто требуют большого объема данных для достижения оптимальной производительности. В данном исследовании использовался пакет "nnet", который тестировался с использованием от 1 до 15 скрытых слоев и нейронной сети прямой связи с функцией активации сигмоида. Для целей классификации была выбрана наиболее эффективная модель.

6. Результаты.

6.1. Анализ данных.

Первый подход к пониманию *поведения* данных заключается в использовании описательной статистики, а именно основных дескрипторов тенденций. Для эталонного набора данных были рассчитаны среднее значение, стандартное отклонение (sd), асимметрия (skewness), и эксцесс для шести типов рудных месторождений. Среднее значение и стандартное отклонение дают представление о значениях, используемых в каждом конкретном случае, и их вариациях, в то время как асимметрия дает представление о нормальной или асимметричной тенденции данных, а эксцесс представляет распределение данных как плоское или клиновидное.

В таблице 1 представлены эти результаты, сгруппированные по типу месторождения.

Дескрипторы справочного набора данных.

	BIF					IOA					IOCG				
	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis
Al	39	640.64	529.01	1.3	1.3	55	4308.84	2894.45	3.7	17.9	34	316.26	110.14	0.7	-0.3
Mg	39	944.89	948.55	1.0	0.3	55	5570.18	2738.01	0.5	-0.4	34	46.14	57.92	1.6	1.8
Ti	39	53.20	43.51	1.1	0.9	55	11,800.67	18,182.72	4.1	17.8	34	82.88	89.16	3.9	15.2
V	39	6.61	8.27	2.5	6.1	55	2569.07	765.08	1.6	4.0	34	654.38	548.12	0.8	-1.2
Mn	39	198.67	404.15	2.5	5.6	55	619.30	389.83	2.0	5.8	34	43.36	45.84	1.9	2.5
Co	39	2.88	2.25	0.4	-1.3	55	112.52	21.32	-1.5	2.0	34	76.73	17.70	-1.7	2.4
Ni	39	8.09	3.38	2.0	5.2	55	319.11	61.25	-0.1	1.2	34	168.30	101.21	0.6	-1.0
Zn	39	39.26	19.11	0.8	0.7	55	99.46	34.66	0.2	-0.6	34	5.45	3.23	1.2	0.7
Ga	39	2.65	1.08	0.2	-1.1	55	26.90	8.18	-0.4	-1.3	34	15.62	8.61	0.7	-1.0
	Porphyry					Skarn					VMS				
	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis
Al	68	3010.65	2072.08	0.6	-0.2	52	6121.81	2168.93	-0.3	-0.5	23	1108.97	1274.49	1.2	-0.3
Mg	68	1028.38	962.33	1.1	0.6	52	5886.31	2603.75	1.2	0.3	23	157.09	337.43	2.8	7.8
Ti	68	5094.00	6555.82	1.6	1.7	52	984.92	467.10	-0.5	-1.0	23	601.96	875.89	2.3	4.8
V	68	927.73	777.27	1.4	2.0	52	193.66	52.21	-1.1	1.5	23	916.87	1096.34	0.9	-0.7
Mn	68	2048.01	2353.73	2.5	6.5	52	2916.97	1616.78	1.5	1.7	23	801.13	587.59	0.7	-1.2
Co	68	44.60	27.33	0.4	0.0	52	62.09	27.61	2.3	5.2	23	18.58	17.57	1.0	-0.1
Ni	68	55.19	73.60	3.5	13.1	52	6.98	2.66	-1.0	0.0	23	20.78	28.52	1.5	0.8
Zn	68	647.48	513.99	1.2	0.7	52	283.40	129.73	1.1	0.4	23	151.70	219.68	2.9	8.6
Ga	68	57.64	29.11	-0.5	-0.6	52	21.17	8.58	-0.5	-1.4	23	49.99	46.83	1.5	1.4

Что касается среднего значения и стандартного отклонения, общеизвестно, что значения в основном различаются для разных элементов и для каждого типа месторождения; тем не менее, есть некоторые перекрывающиеся значения, например, Ti в порфире ($x = 2057,78$, $\sigma = 1751,09$) и скарне ($x = 211,46$, $\sigma = 483,71$) или Zn в BIF ($x = 39,26$, $\sigma = 19,11$) и виртуальных машин ($x = 49,80$, $\sigma = 17,23$).

Значения асимметрии в большинстве типов месторождений слабо положительны, что указывает на хорошее распределение по Гауссу. Значения эксцесса имеют умеренную положительную тенденцию, указывая на слегка заклинивающую структуру распределения, следовательно, с ограниченным разбросом между центральными значениями. Исключением из этого утверждения являются месторождения VMS с высокой асимметрией и эксцессом для элементов Al, Mg и Ni. Вторая группа выдающихся значений приходится на месторождения IOA, Al и Ti в IOCG, все они демонстрируют значения эксцесса выше 10.

Результаты анализа разведочных данных для набора данных тематического исследования представлены в таблице 2.

Учитывая набор данных тематического исследования, как и ожидалось от признанных типов месторождений, месторождения Алвито и Азенхас группируются по большинству элементов и значительно отличаются от других месторождений; это более заметно по более высоким значениям Al, Mg, Ti, Co, Ni, Zn и Ga. Напротив, содержание Mn в этих месторождениях ниже, чем в других. Месторождение Орада отличается характерными

тенденциями; в некоторых случаях значения состава сходны с месторождениями Алвито и Азенхас, а в других случаях средние значения больше похожи на месторождения других типов.

Асимметрия в большинстве тематических исследований слабо положительная, за исключением месторождения Орада, где она слабо отрицательная. Выделяется из этой тенденции месторождение Монжес, где Al, Mg и Ni имеют сильные положительные значения асимметрии и эксцесса.

Табл. 2.

Дескрипторы набора данных тематического исследования.

	Alvito					Azenhas					Monges				
	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis
Al	159	9851.18	2502.80	-0.1	0.5	22	2002.19	1302.13	0.7	-0.7	98	914.55	1134.41	4.7	29.8
Mg	159	3730.98	1998.24	1.1	5.0	22	8092.75	3016.74	1.0	0.4	98	2945.82	2913.96	5.5	41.4
Ti	159	2057.78	1751.09	1.0	-0.6	22	2111.46	483.71	1.2	0.3	98	569.98	212.23	0.8	-0.3
V	159	70.43	28.93	-0.3	-0.9	22	129.94	29.47	-0.3	-1.1	98	37.74	17.63	0.9	-0.2
Mn	159	2149.33	568.79	0.7	0.8	22	4827.21	681.53	0.2	-1.1	98	6927.25	4389.85	0.5	-1.1
Co	159	85.93	24.92	0.2	0.1	22	52.19	13.28	0.1	-0.7	98	0.37	0.36	0.8	-0.8
Ni	159	40.44	29.47	0.1	-1.3	22	17.01	4.56	0.3	-1.4	98	1.35	1.33	5.2	34.2
Zn	159	523.72	204.85	0.2	-0.5	22	658.76	244.29	-0.2	-1.0	98	49.80	17.23	-0.1	-1.0
Ga	159	39.62	12.29	0.7	1.5	22	12.78	8.07	0.2	-1.9	98	6.21	2.45	0.4	-0.8
	Orada					Serrinha					Vale da Arca				
	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis	n	mean	sd	skew	kurtosis
Al	35	2772.54	685.00	-0.2	-0.8	45	715.37	153.65	1.0	-0.1	12	67.14	49.04	0.4	-1.4
Mg	35	2597.43	589.30	1.9	4.6	45	4427.55	1453.70	2.0	4.5	12	2254.57	512.76	-0.2	-0.9
Ti	35	144.14	10.34	0.0	0.1	45	69.27	46.96	0.6	-1.3	12	142.46	42.48	-1.1	0.9
V	35	44.18	3.74	-0.8	3.2	45	8.95	10.25	1.2	-0.1	12	132.05	38.55	-1.8	2.7
Mn	35	497.61	54.71	-0.1	0.9	45	11,267.92	1178.06	0.2	0.0	12	7250.09	687.79	0.0	-0.9
Co	35	28.97	2.33	-1.3	4.3	45	0.53	0.06	1.1	2.7	12	0.03	0.01	2.6	5.5
Ni	35	24.40	2.42	-0.7	1.3	45	1.14	0.49	2.1	6.9	12	15.84	14.88	1.0	-0.7
Zn	35	68.82	12.62	0.3	-0.6	45	46.59	8.25	0.6	-0.4	12	109.27	15.32	-0.2	-0.9
Ga	35	8.38	0.85	-0.7	1.9	45	0.62	0.17	0.7	-0.7	12	3.47	0.52	-0.1	-1.2

6.2. Классический подход.

Классификация месторождений полезных ископаемых, основанная на дискриминантных диаграммах для эталонного набора данных, приведена на рисунке 2.

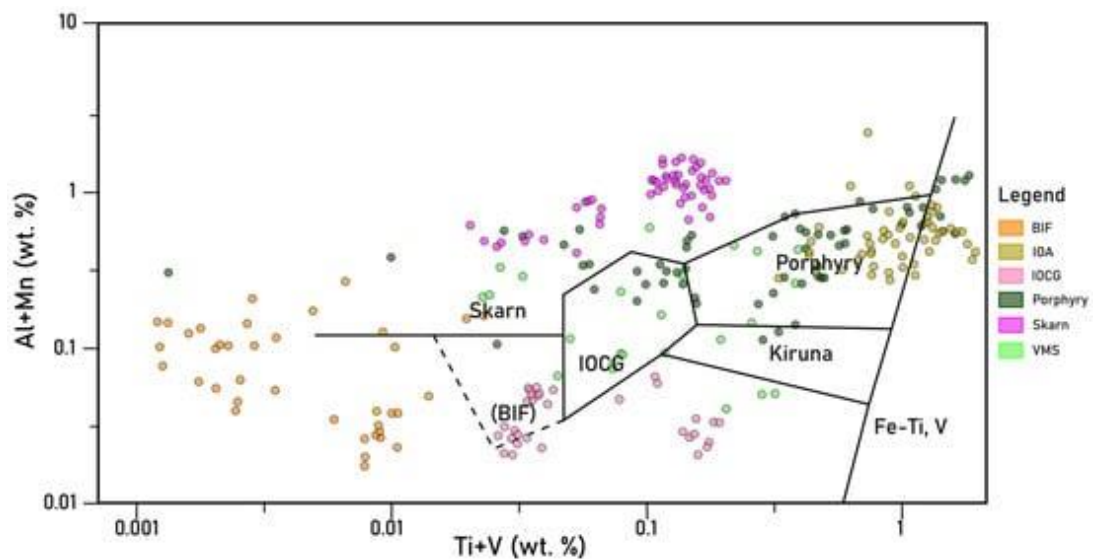


Рис. 2. Базовый набор данных, нанесенный в пространстве диаграммы.

Образцы VMS не предназначены для отображения на этой диаграмме и, следовательно, пересекают несколько определенных полей. Данные по порфирам частично совпадают с IOA. Данные IOCG из справочного набора данных выходят за пределы определенного поля, представляя значения Al+Mn ниже ожидаемых и каким-то образом находятся внутри поля BIF. С другой стороны, данные BIF находятся за пределами выявленных месторождений, попадая в области с более низким, чем ожидалось, содержанием Ti+V.

Это демонстрирует сложность использования этих диаграмм. Несмотря на эту сложность, также ясно, что на этих диаграммах можно легко отделить магнетиты от некоторых основных типов месторождений, но другие, такие как порфир, VMS или IOA, имеют вышележащие области на этих диаграммах.

Данные тематического исследования, нанесенные на диаграмму, представлены на рисунке 3. Согласно этой диаграмме, все данные нанесены на график и, таким образом, классифицированы в поле типа скарна, по существу, с различным содержанием компонента Ti+V.

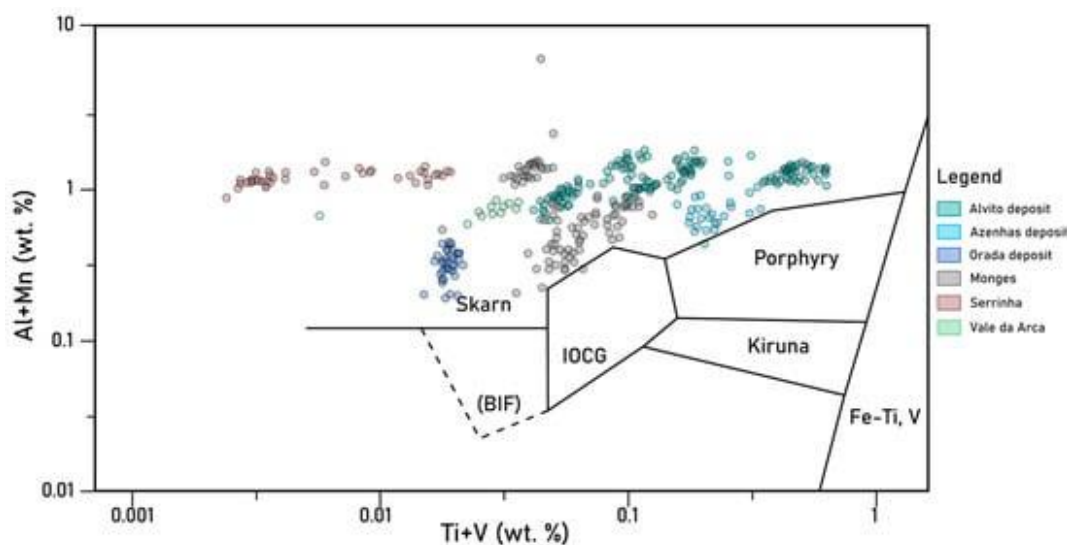


Рис. 3. Набор данных тематического исследования на диаграмме.

Легко понять, что два типа месторождений из тематического исследования португальской OMZ нелегко отличить с помощью этой диаграммы.

6.3. Уменьшение размеров (PCA).

Анализ PCA, примененный к эталонному набору данных, показывает, что PC₁ объясняет 32,48% изменчивости и в нем доминирует поведение Co, V и Ni. Компонент PC₂ объясняет 26,26% вариабельности и в нем преобладают элементы Zn, Mn и Ga.

Уравнение, выведенное из трех основных элементов, влияющих на PC₁, является:

$$0.49 \times \text{Co} + 0.46 \times \text{Ni} + 0.46 \times \text{V} \quad (1)$$

и для PC_2 уравнение имеет вид:

$$-0.57 \times Zn - 0.53 \times Mn - 0.37 \times Ga \quad (2)$$

Эти результаты усиливают роль комбинированных элементов Co, Ni и V для объяснения наиболее значительной части изменчивости (PC_1) и комбинированных Zn, Mn и Ga для объяснения другой части (PC_2). График PC_1 – PC_2 с эллипсами, соответствующими 75% данных для каждой группы, представлен на рисунке 4.

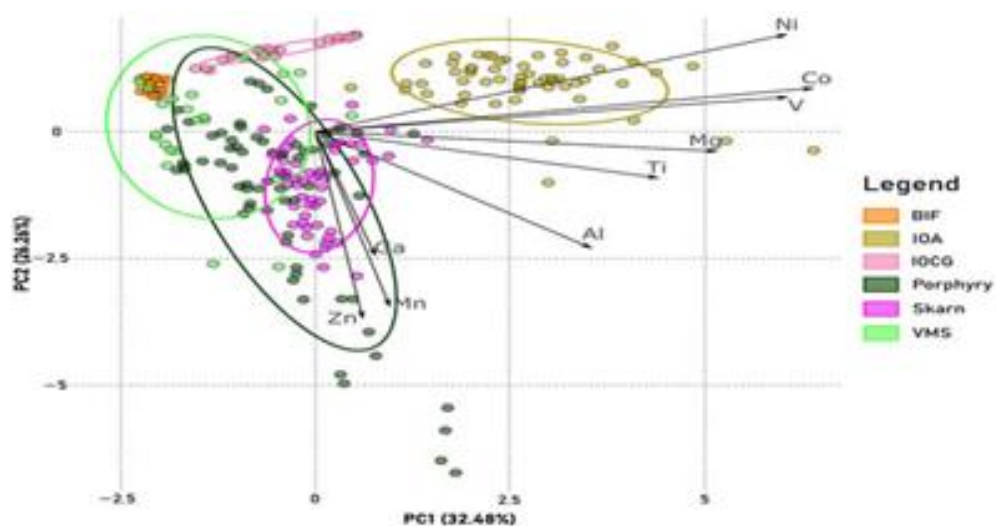


Рис. 4. Базовый набор данных в пространстве диаграммы.

Хорошее различие месторождений BIF, IOCG, скарнов и IOA возможно с использованием диаграммы PC_1 – PC_2 , тогда как порфировые и VMS показывают значительное совпадение между ними и другими месторождениями.

Анализ PCA, примененный к набору данных тематического исследования, показывает хорошую дискриминацию между всеми изученными месторождениями (рис. 5). PC_1 объясняет большую часть дисперсии (53,64%), тогда как PC_2 объясняет меньшую вариативность (17,44%). Исключением является месторождение Монжес, данные по которому в некоторой степени совпадают с месторождениями Серринья и Вале-да-Арка.

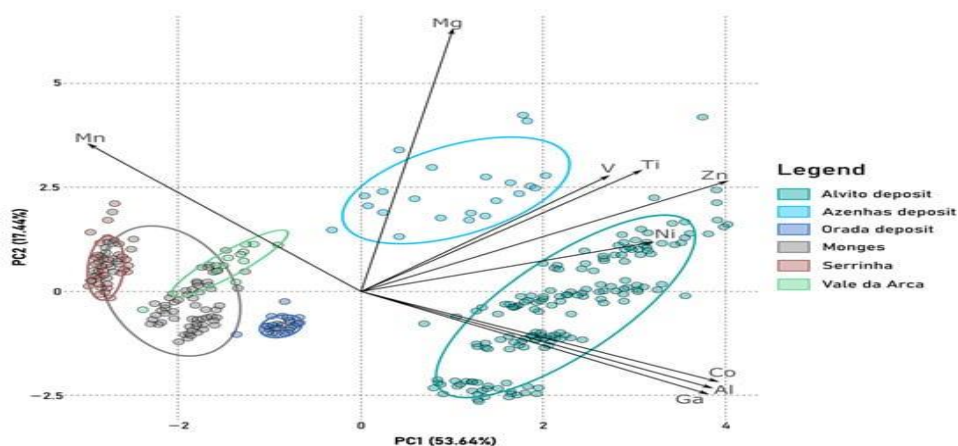


Рис. 5. График PC_1 – PC_2 для набора данных тематического исследования.

Уравнение трех наиболее значимых переменных для PC₁ является:

$$0.41 \times \text{Zn} + 0.40 \times \text{Co} + 0.39 \times \text{Al} \quad (3)$$

и для PC₂:

$$0.65 \times \text{Mg} + 0.37 \times \text{Mn} + 0.30 \times \text{Ti} \quad (4)$$

В данном случае наблюдается явное противопоставление групп Zn, Co и Al группам Mg, Mn и Ti.

6.4. Машинное обучение.

Результаты различных классификаторов лучше понятны при анализе матриц путаницы, которые являются показателем эффективности модели. Критерии для рассмотрения правильной классификации месторождений определяются как мера качества результатов.

Одним из распространенных показателей такой точности является оценка F_1 , которая определяется как:

$$F_1\text{-Оценка} = 2 * (\text{Точность} * \text{повторения}) / (\text{Точность} + \text{повторение})$$

где точность = $TP / (TP + FP)$ и отзыв = $TP / (TP + FN)$, где TP-истинные положительные результаты, FP-ложноположительные результаты, а FN-ложноотрицательные результаты.

Таким образом, точность выше 65% считается правильной классификацией, точность от 55% до 65% считается неубедительной, а точность ниже 55% считается неверной.

Значение важности перестановки используется для определения важности переменной в каждой модели. Этот параметр измеряет изменение производительности модели после перетасовки значений данной переменной. Этот расчет выполняется с использованием функции `varImp()` из пакета 'caret'. Если производительность модели снижается, это указывает на то, что функция важна для составления точных прогнозов. Если производительность остается прежней или улучшается, это говорит о том, что функция менее важна или даже вредна для прогностической способности модели. Положительное значение указывает на то, что производительность модели снизилась после того, как объект был перетасован; следовательно, объект важен для того, чтобы модель делала точные прогнозы. Чем больше положительное значение, тем важнее объект. Отрицательное значение указывает на то, что производительность модели улучшилась или осталась прежней после изменения функции.

Оценка F_1 - это еще один показатель, используемый для оценки производительности модели машинного обучения. Он объединяет показатели точности и отзыва в единый показатель. Точность измеряет долю истинно положительных прогнозов среди всех положительных прогнозов, в то время как отзыв измеряет долю истинно положительных

прогнозов среди всех фактических положительных образцов. Оценка F_1 - это среднее гармоническое значение точности и отзыва. Оценка F_1 варьируется от 0 до 1, причем 1 является наилучшей возможной оценкой. Высокий балл F_1 указывает на то, что модель обладает высокой точностью и высокой отзывчивостью, что означает, что она может правильно идентифицировать положительные образцы и минимизировать ложноположительные результаты и ложноотрицательные результаты. Низкий балл F_1 указывает на то, что модель работает недостаточно хорошо либо из-за низкой точности, низкого отзыва, либо из-за того и другого.

Как значение важности перестановки, так и оценка F_1 применимы только к справочному набору данных и рассматриваются для оценки производительности моделей.

6.4.1. Моделирование эталонных наборов данных.

Как указано выше, базовый набор данных был разделен на набор данных Train и тестовый набор данных. Результаты классификации тестового набора данных после обучения модели с помощью набора данных Train дают следующие результаты:

- Случайный лес (RF)

Модель RF была запущена с использованием 500 деревьев, что является хорошим компромиссом между временем вычислений и точностью. Существует 100%-ный прогноз положительных значений для типов месторождений IOCG и скарна с хорошей точностью для месторождений порфира, IOA и BIF. Идентификация месторождений с помощью виртуальной машины имеет наихудшее значение точности.

Чтобы проиллюстрировать одно из деревьев принятия решений, созданных для модели RF, на рисунке 6 показан пример одного из деревьев; в данном случае того, у которого меньше ветвей. Обратите внимание, что в этом дереве используются не все химические элементы, в котором только значения Ti, Ni, Mg, Ga, V и Mn соответствуют узлам принятия решений, используемым для классификации. В данном случае наиболее релевантными элементами являются Ti и Ni.

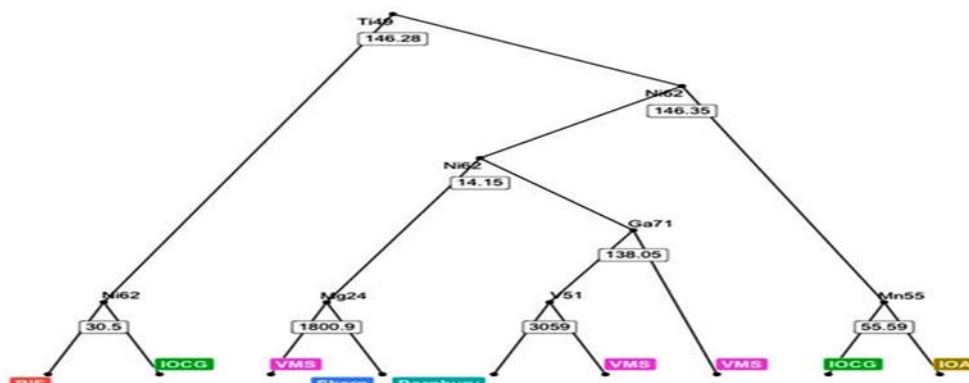


Рис. 6. Пример случайного дерева решений, созданного на основе справочного набора данных.

Что касается переменной важности, измеренной методом перестановок, Mg имеет более высокое значение (0,25), за которым следуют Zn (0,04) и Ni (0,02); все остальные переменные имеют меньшую важность ($<0,01$) в этой модели.

Общая точность с учетом взвешенного показателя F_1 для этой модели составляет 92%.

- Наивный Байес

Результаты с использованием модели NB отображаются в матрице путаницы. С помощью этой модели классификатора месторождения IOA, IOCG и скарны имеют идеальный истинно положительный прогноз. BIF и порфир имеют хорошую точность, в то время как типы месторождений VMS имеют более низкие значения.

Переменная важность, рассчитанная с использованием метода перестановок, указывает на более высокую важность, связанную с Ni (0,04), за которой следуют Ti (0,02) и V (0,02). Все остальные переменные имеют ограниченное значение, учитывая прогнозирование этой модели.

Оценка F_1 с общей взвешенной точностью для модели NB составляет 89%.

- Машина опорных векторов (SVM)

Результаты классификации модели SVM представлены матрицей путаницы. Этот классификатор содержит идеальную классификацию для месторождений IOA, IOCG и скарна; BIF и порфир имеют правильную классификацию, а месторождения VMS имеют неправильные результаты классификации.

Важность перестановки различных переменных указывает на то, что Ga (0,15), Al (0,12), Co (0,11) и V (0,10) являются наиболее значимыми переменными. Тем не менее, все остальные, за исключением Ti, имеют значения выше 0,04.

Взвешенная точность с использованием значения F_1 -score составляет 83%.

- Мультиномиальная логистическая регрессия (MLR)

Результаты использования модели MLR, выражены в виде матрицы путаницы. Этот классификатор дает отличные результаты классификации для IOA и IOCG, представляя хорошие результаты для всех других типов месторождений.

Значения важности перестановки указывают на то, что Co (0,33), Mg (0,32), Ga (0,29) и Ni (0,28) имеют более высокие значения и, следовательно, играют наиболее важную роль в определении модели. Тем не менее, другие переменные имеют значимые значения, где Ti имеет наименьшее значение (0,07).

Общая взвешенная точность, измеренная с использованием оценки F_1 для этой модели, составляет 88%.

- К-Ближайший сосед (KNN)

Результаты с использованием модели классификатора KNN представлены в виде матрицы путаницы. В данном случае месторождения ИОА имеют идеальную классификацию, тогда как BIF и skarn имеют точность выше 80%. IOCG, porphyry и VMS имеют правильную классификацию, но имеют более низкие значения.

Что касается важности перестановки различных переменных, результаты показывают, что Mg (0,34), Al (0,31) и Ti (0,28) имеют более высокие значения, тогда как Ni, Co, Zn и Ga имеют значение ниже 0,02.

Оценка F_1 для этой модели составляет 77%.

- Искусственная нейронная сеть (ANN)

Результаты с использованием подхода ANN model представлены в виде матрицы путаницы. Это наихудшая модель, где ни один из типов месторождений не получает совершенной классификации. Тем не менее, IOA, skarn, BIF и IOCG имеют точность, равную или превышающую 80%, типы месторождений порфира имеют правильную классификацию, но с низким значением точности, а классификация VMS неадекватна.

Важность перестановок с использованием этой модели указывает на то, что Mg (0,33), Mn (0,31), V (0,30), Al (0,26) и Ti (0,26) имеют более высокую важность. Другие химические элементы имеют очень низкое значение в классификации моделей.

Оценка F_1 при использовании модели ANN является самой низкой и составляет 76%.

Наиболее эффективными моделями в порядке убывания оценки F_1 являются случайный лес (93%), наивный байесовский алгоритм (89%), мультиномиальная логистическая регрессия (88%), машина опорных векторов (83%), k-ближайший сосед (77%) и искусственные нейронные сети (76%).

6.4.2. Моделирование набора данных тематического исследования.

Полный справочный набор данных, включающий 271 наблюдение, был использован для обучения моделей машинного обучения (ML), которые будут применяться в наборе данных тематического исследования. Затем эти модели были использованы для прогнозирования классификации месторождений Ossa-Morena. Перекрестные таблицы между месторождением OMZ и предлагаемой классификацией с использованием модели ML оказались лучшим методом оценки полученных результатов.

Показатель достоверности (Assert.) каждой модели используется для количественной оценки процента значений, которые относятся к одному и тому же типу модели. Более высокие значения указывают на то, что большинство наблюдений относятся к одному и тому же типу месторождения. Обратите внимание, что это не означает точной классификации

месторождения и должно интерпретироваться как показатель однородности классификации. Эти значения представлены в последней колонке следующих таблиц. В этих таблицах выделенные значения относятся к наиболее рекомендуемой классификации с использованием модели. В некоторых случаях нет четкого решения, указывающего на две возможные классификации; например, месторождение Монжес в равной степени относится к типам BIF и VMS с 40 наблюдениями, предложенными для каждого типа месторождения.

- Модель случайной классификации лесов (RF).

Модель RF для месторождений Алвито и Азенхас указывает на тип месторождения скарнов, хотя в Алвито большое количество наблюдений приходится на тип месторождения порфира. Орада классифицируется как месторождение типа VMS. Монжес, Серринья и Вале-да-Арка считаются месторождениями типа BIF; однако наблюдения Монжеса в равной степени классифицируются как месторождение типа VMS.

Повышенная надежность достигается для месторождения Азенхас в виде скарна, а также Серриньи и Вале-да-Арка в виде BIF.

- Наивная байесовская классификационная модель (NB).

Классификация модели NB для тематического исследования несколько похожа на режим RF. Однако Алвито и Азенхас почти поровну относятся к порфировым и скарновым месторождениям. Месторождения Орада и Монжес в основном классифицируются как месторождения VMS. Серринья и Вале-да-Арка классифицируются как месторождения BIF.

Что касается самоутверждения, то Орада, Серринья и Вале-да-Арка работают лучше.

- Модель машинной классификации с опорными векторами.

Алвито и Азенхас классифицируются как скарновые месторождения, хотя во многих наблюдениях Азенхас считается порфировым. Орада классифицируется как месторождение BIF, тогда как Монжес, Серринья и Вале-да-Арка классифицируются как месторождения порфирового типа.

Что касается ассертивности, то месторождения Алвито, Серринья и Вале-да-Арка показывают лучшие результаты.

- Модель классификации с мультиномиальной логистической регрессией (MLR).

Эта модель указывает, что Алвито и Азенхас относятся к типу скарновых месторождений. Месторождение Орада считается месторождением BIF. Монжес, Серринья и Вале-да-Арка классифицируются как месторождения порфирового типа.

Что касается самоутверждения, то месторождения Алвито, Азенхас, Орада и Вале-да-Арка показывают лучшие результаты. Эта модель также в целом демонстрирует лучшее самоутверждение.

- Модель классификации К-ближайших соседей (KNN).

Алвито и Азенхас классифицируются как скарновые месторождения. Орада относится как к типу BIF, так и к типу порфировых месторождений. Монжес и Вале-да-Арка в основном классифицируются как порфировые, тогда как Серринья определяется как месторождение скарнов.

Что касается самоутверждения, то месторождения Алвито и Вале-да-Арка показывают лучшие результаты.

- Модель классификации искусственной нейронной сети (ANN).

Алвито классифицируется как месторождение скарнов, но также попадает в классификацию BIF совместно с Azenhas. Орада классифицируется как месторождение скарнов, тогда как Монжес классифицируется как порфировый, а Серринья и Вале-да-Арка классифицируются как месторождения VMS.

Что касается асертивности, то Орада, Серринья и Вале-да-Арка - это месторождения, где эта модель работает лучше.

7. Обсуждение.

7.1. Уменьшение размеров с помощью PCA.

Продemonстрировано, что подход с уменьшением размерности, а именно PCA, является хорошим инструментом для понимания поведения различных химических элементов в системе, предоставляя представление о том, как они коррелируют, и о том, как учитывать индивидуальное или групповое поведение для создания дискриминантных диаграмм. Наши результаты хорошо согласуются со справочными химическими элементами, которые лучше различают магнетит из разных металлогенических сред.

Преобладание Ni, Co и V как наиболее значимых элементов в справочном наборе данных соответствует результатам, в котором предлагаемая диаграмма включает Co и V в качестве важных элементов для выделения рудных месторождений. Ni объясняется в случае месторождения Алвито как элемент, который, возможно, отражает более сильное влияние магматических флюидов из близкого габбро-диоритового комплекса.

График PC₁–PC₂ (рис. 4) эффективно различает большинство типов месторождений, за исключением месторождений порфира и VMS, которые перекрывают несколько других типов месторождений. Результаты не включают ни одной пробы с месторождений VMS и содержат ограниченное количество проб магнетита порфирового происхождения. Магнетит из порфировых и скарновых месторождений, и различение этих двух типов месторождений с использованием состава магнетита возможно, но с большим разбросом данных в порфировых

месторождениях, тогда как в скарновых месторождениях данные сгруппированы. Диаграмма PC₁–PC₂ показывает некоторую степень перекрытия, аналогичную тому, что было обнаружено в тематических исследованиях.

Наиболее релевантными элементами для набора данных тематического исследования являются Zn, Co и Al для PC₁ и Mg, Mn и Ti для PC₂. Эти элементы также хорошо согласуются с выводами, которые включают изучение скарновых и порфировых месторождений меди. Sn является исключением и рассматривается как относящееся к региональному геологическому контексту. В тематическом исследовании геологический контекст благоприятствует обогащению Zn, что может быть связано с относительным обогащением Pb-Zn гидротермальных систем в исследуемом районе и относительным истощением Sn.

Mg и Mn определили, как элементы, которые вносят наибольший вклад в формирование скарновых месторождений, а Si, Ti и Al - как элементы, на которые в большей степени влияют магматические флюиды. Zn и Co, как халькофильные элементы, зависят от распределения сульфидов и магнетита. Эти элементы хорошо коррелируют с элементами, которые определяют PC₁ и PC₂ для тематического исследования, четко отделяя скарновые месторождения (Алвито и Азенхас) от связанных с «выдыханием» осадочных отложений, измененных гидротермальными флюидами (Монжес, Серринья и Вале-да-Арка). Кластер месторождений Орада находится ближе к этим поздним месторождениям, что определяет необходимость дальнейшего изучения этих магнетитов и их типов месторождений.

7.2. Модели ML в тестовом наборе данных.

Протестированные методы ML, то есть случайный лес, наивный байесовский алгоритм, метод опорных векторов, мультиномиальная линейная регрессия, k-ближайший сосед и искусственные нейронные сети, имеют общую точность выше 70% при использовании оценки F₁, что указывает на хороший прогноз.

Модель случайного леса показала лучшие результаты с точностью 92% при использовании показателя F₁-score. Тестовый набор данных из более чем 17 тыс. наблюдений позволил получить точность 81%. Эта более низкая точность по сравнению с полученными результатами не является показателем лучшей производительности модели, но может быть вызвана значительно большим набором данных, то есть количеством анализов магнетита, использованных в качестве обучающего набора данных для модели. Это неизбежно снижает точность, поскольку при классификации типов месторождений подразумевается больше неограниченных наблюдений. Результаты показывают, что общая точность, полученная с использованием методов ML, позволяет использовать набор данных Train с меньшим количеством наблюдений, более сбалансированный с точки зрения типов месторождений, с

исключительными результатами по сравнению с классическим диаграммным подходом. V, Ni и Mg являются наиболее важными элементами в модели, тогда как в RF-модели преобладают Mg, Zn и Ni.

Модели SVM, используя 876 наблюдений в магматических и гидротермальных системах, получили оценки F_1 , которые находятся в пределах этого диапазона. Средний балл F_1 составляет 85%, тогда как тематический - 83%, следовательно, они хорошо. Mg, Mn, Al, Ti, V и Co определили в качестве ключевых факторов для выделения месторождений. Наиболее важными переменными, использующими SVM-модель, являются Ga, Al, Co и V, совпадающие по большинству элементов.

Для отличия IOA от типов месторождений IOCG использовали RF-модель с набором последовательностей из 913 наблюдений LA-ICP-MS в сочетании с 877 наблюдениями EPMA. Модель получила точность 91%, что хорошо согласуется с тематическим показателем F_1 в 92% при использовании той же модели машинного обучения. Si, Mg и V определили как наиболее важные элементы в классификации, а тематическая RF-модель идентифицирует Mg, Zn и Ni как наиболее важные. Это может быть связано с большим количеством типов месторождений, рассматриваемых в этой модели.

Учитывая выбранные химические элементы, эталонный набор данных демонстрирует сложность идентификации типов месторождений VMS. Именно модель ML показала лучшие результаты при классификации месторождений VMS, показав модель MLR с 86% истинно положительных значений.

В таблице 3 обобщены полученные результаты и проведено их сравнение с тестовыми данными.

Табл. 3.

Сравнение результатов тематического и тестового исследований.
Значения точности измеряются в баллах F_1 .

ML Model	[13]	[14]	[15]	This Study
RF	91% Si, Mg, V		81% V, Ni, Mg	92% Mg, Zn, Ni
NB			58% Not defined	89% Ni, Ti, V
SVM		85% Mg, Mn, Al, Ti, V, Co		83% Ga, Al, Co, V
MLR				88% Co, Mg, Ga, Ni
KNN			69% Not defined	77% Mg, Al, Ti
ANN				76% Mg, Mn, V

Этот анализ позволяет рассматривать RF как лучшую модель для определения типа месторождения, за которой следуют модели NB и MLR.

7.3. ML применительно к набору данных тематического исследования.

Для набора данных тематического исследования обстановка несколько иная. Модель RF имеет более низкую среднюю достоверность (69%), а MLR - самую высокую (89%), но следует помнить, что это не показатель точности модели, а показатель того, как модель сохраняет свою классификацию в той же группе.

Согласно модели RF, Азенхас, Серринья и Вале-да-Арка являются месторождениями с более высокой степенью достоверности (>50%). Азенхас и Алвито классифицируются как скарны, в то время как Серринья и Вале-да-Арка классифицируются как BIF. Классификация Монжа находится в тупике между BIF и VMS. Принимая во внимание тестовую классификацию, классификация месторождений Алвито и Азенхас, использующих состав магнетита, как месторождений скарнов, верна. Что касается Монжа, Серриньи и Вале-да-Арка, то их отнесение к BIF, а также к VMS также можно считать правильным, поскольку их эггалационно-осадочное происхождение в сочетании с поздними гидротермальными изменениями приводит к переменному микроэлементному составу в этих магнетитах. Более того, нерешительность в отношении месторождения Алвито между скарнами и порфиром (т. е. с поступлением магматического флюида) также указывает на вероятную смешанную природу этого месторождения с сильным влиянием соседней габро-диоритовой свиты. Классификация месторождения Орада как VMS очень неоднозначна, что также свидетельствует о несколько неоднозначном характере этого объекта.

Модель NB дает результат несколько схожий с RF. Однако для месторождения Азенхас явно относится к месторождению порфиров. Классификация месторождения Орада в этой модели явно относится к типу VMS.

Модели SVM и MLR дают схожие результаты идентификации, явно рассматривая Алвито и Азенхас как скарны, тогда как Монж, Серринья и Вале-да-Арка относятся к типу месторождений порфира. Обе модели классифицируют Ораду как месторождение BIF. Таким образом, для Алвито и Азенхаса существует четкое согласие относительно классификации скарнов, а также альтернативная классификация Монжеса, Серриньи и Вале-да-Арка как типов месторождений порфира. Это может быть связано с тем, что эти модели придают большее значение таким элементам, как Co и Ga, которые обычно содержатся в более высокой концентрации в магнетите из порфировых месторождений по сравнению с магнетитом из месторождений BIF (табл. 1). Модель KNN похожа на модели SVM и MLR; однако она классифицирует Серринью как скарновое месторождение, возможно, из-за более высокого

значения Mg, которое является наиболее важным элементом в этой модели и характерно для скарновых месторождений в результате метасоматических изменений.

Модель ANN предоставляет существенно отличающиеся классификации, которые трудно сопоставить с какой-либо интерпретацией моделей. Алвито и Орада классифицируются как месторождения скарнов, а Монжес - как месторождения порфиров. Азенхас классифицируется как месторождение BIF, и, наконец, Серринья и Вале-да-Арка классифицируются как «виртуальные машины». Это показатель того, что результаты этой модели должны быть отброшены или протестированы с более точно настроенными параметрами.

Следует отметить, что на диаграмме (рис. 3) все месторождения, рассмотренные в тематическом исследовании, относятся к типу скарновых месторождений, и большинство используемых моделей способны различать и лучше идентифицировать этот тип месторождений.

8. Выводы.

Из полученных результатов можно сделать методологические и металлогенические выводы. Они сводятся к следующему:

1. Использование ограниченного набора данных о геохимии микроэлементов в магнетитах в сочетании с большинством моделей машинного обучения продемонстрировало, что является надежным инструментом для обеспечения точной классификации типов рудных месторождений.

2. Различные протестированные методы машинного обучения обеспечивают точную классификацию типов месторождений; однако случайный лес, наивный Байес и мультиномиальная логистическая регрессия определены как наиболее точные. Искусственные нейронные сети обладают наихудшей производительностью.

3. Объединяя все предложенные классификации, Алвито и Азенхас уверенно классифицируются как скарны. Монж, Серринья и Вале-да-Арка классифицируются либо как порфир, либо как BIF. Orada классифицируется либо как BIF, либо как VMS.

4. Применение моделей машинного обучения к тематическим исследованиям дает значимые результаты, которые могут быть использованы не только для интерпретации типов рудных месторождений, но и для выявления смешанного или более сложного происхождения.

5. Применение этих методов при ГРП является мощным инструментом, который проливает новый свет на интерпретацию магнетита, что может дать подсказки для целей поисков и разведки ПИ.

ИСТОЧНИКИ:

1. Shi Li, Jianping Chen and Chang Liu. «OVERVIEW ON THE DEVELOPMENT OF INTELLIGENT METHODS FOR MINERAL RESOURCE PREDICTION UNDER THE BACKGROUND OF GEOLOGICAL BIG DATA». *Minerals* 2022, 12.
2. Nan Lin, Yongliang Chen, Haiqi Liu and Hanlin Liu. «A COMPARATIVE STUDY OF MACHINE LEARNING MODELS WITH HYPERPARAMETER OPTIMIZATION ALGORITHM FOR MAPPING MINERAL PROSPECTIVITY». *Minerals* 2021, 11.
3. Yongzai Xi, Yongbo Li, Junjie Liu, Shan Wu, Ning Lu 1,2,3, Guixiang Liao and Qiule Wang, «APPLICATION OF ANALYTIC HIERARCHY PROCESS IN MINERAL PROSPECTING PREDICTION BASED ON AN INTEGRATED GEOLOGY-AEROGEOPHYSICS-GEOCHEMISTRY MODEL». *Minerals* 2023,13.
4. Ying Wang, Nolwenn Coint, Eduardo Teixeira Mansur, Pedro Acosta-Gongora, Ana Carolina Rodrigues Miranda, Aziz Nasuti and Vikas Chand Baranwal. «LEVERAGING DOMAIN EXPERTISE IN MACHINE LEARNING FOR CRITICAL METAL PROSPECTING IN THE OSLO RIFT: A CASE STUDY FOR FE-TI-P-RARE EARTH ELEMENT MINERALIZATION». *Minerals* 2024, 14.
5. Shuai Zhang, Emmanuel John M. Carranza, Changliang Fu, Wenzhi Zhang and Xiang Qin. «INTERPRETABLE MACHINE LEARNING FOR GEOCHEMICAL ANOMALY DELINEATION IN THE YUANBO NANG DISTRICT, GANSU PROVINCE, CHINA». *Minerals* 2024, 14.
6. Jin Guo and Wen-Yan He. «APPLICATION OF MACHINE LEARNING TO CHARACTERIZE METALLOGENIC POTENTIAL BASED ON TRACE ELEMENTS OF ZIRCON: A CASE STUDY OF THE TETHYAN DOMAIN». *Minerals* 2024, 14.
7. Pedro Nogueira and Miguel Maia. «MAGNETITE TALKS: TESTING MACHINE LEARNING MODELS TO UNTANGLE ORE DEPOSIT CLASSIFICATION—A CASE STUDY IN THE OSSA-MORENA ZONE (PORTUGAL, SW IBERIA)». *Minerals* 2023, 13.