

РУДНИЧНАЯ АЭРОГАЗОДИНАМИКА И ГОРНАЯ ТЕПЛОФИЗИКА

УДК 622.831.322

DOI:10.7242/echo.2025.2.5

ПРИМЕНЕНИЕ АЛГОРИТМА «ДЕРЕВЬЕВ РЕШЕНИЙ» В ЗАДАЧЕ ЛОКАЛЬНОГО ПРОГНОЗИРОВАНИЯ ГАЗОДИНАМИЧЕСКИХ ЯВЛЕНИЙ

В.О. Лядов

Горный институт УрО РАН, г. Пермь

Аннотация: В статье представлены результаты разработки модели классификации на основе алгоритма дерева решений в приложении к задаче локального прогнозирования газодинамических явлений. Из общей выборки, состоящей из 245 наблюдений с 12 горно-геологическими параметрами, были отобраны 84 наблюдения для обучающей выборки. Остальная часть использовалась в качестве тестовой выборки для полученных моделей. Разработанные модели показали, что максимальная точность модели достигается при максимальной глубине дерева, составляющей более 6 уровней. Из 12 использованных для обучения модели горно-геологических параметров наиболее информативными для дискриминации в рассматриваемом алгоритме оказались 8. Отобранная наиболее эффективная модель решающего дерева показала высокую эффективность на тестовом наборе данных, однако известные недостатки деревьев решений в виде высокой чувствительности к выбросам и высокой нестабильности не позволяют в полной мере произвести анализ взаимосвязей между горно-геологическими параметрами, что ограничивает использование данного типа моделей лишь в качестве дополнительного алгоритма в более комплексном ансамбле алгоритмов для решения задачи классификации.

Ключевые слова: газодинамические явления, локальное прогнозирование, задача классификации, машинное обучение, дерево решений, сильвинит.

Введение

Стабильно нарастающие темпы ведения подземных горных работ на калийных месторождениях вслед за повышением мирового спроса на минеральные удобрения требуют комплексных и эффективных решений вопросов, связанных с добычей калийной руды, и в особенности вопросов безопасности ведения горных работ. Известной проблемой, сопровождающей добычу почти на всех известных калийных месторождениях, являются газодинамические явления (ГДЯ), проявляющиеся в виде внезапных выбросов соляных пород и газа, обрушений пород кровли, сопровождающихся газовыделением, а также явлений комбинированного типа [1-3]. В свою очередь, для определения границ участков, опасных по ГДЯ в пределах конкретных областей шахтных полей, производят локальное прогнозирование газодинамических явлений.

В общем случае локальное прогнозирование ГДЯ сводится к решению задачи классификации горно-геологических параметров с двумя классами. Классическим, имеющим обширный опыт использования в решении подобных задач и хорошо интерпретируемым способом классификации является применение линейного дискриминантного анализа [4]. Однако растущая степень цифровизации горной индустрии и легкодоступность больших вычислительных мощностей указывает на актуальность использования новых, ранее широко не использовавшихся методов машинного обучения для решения задачи классификации в рамках локального прогнозирования газодинамических явлений [5]. В данной статье рассмотрено применение дерева классификации – одного из простейших, но при этом хорошо зарекомендовавшего себя алгоритма машинного обучения для решения задачи классификации.

Алгоритм классификации деревом принятия решений

Дерево классификации или дерево принятия решений – это алгоритм машинного обучения, имеющий в своей базе структуру на основе правил «если-то». Каждый узел в дереве представляет собой вопрос, которому удовлетворяют или не удовлетворяют входные данные. Соответствие или несоответствие условию в узле определяет дальнейший путь по ветвям дерева классификации [6]. Важными достоинствами деревьев решений являются их хорошая интерпретируемость, возможность удобной визуализации и отсутствие необходимости предварительной подготовки данных. В свою очередь, данные модели склонны к переобучению, приводящему к созданию чрезмерно сложных деревьев решений.

Существует множество алгоритмов построения деревьев решений, среди которых наиболее популярными являются ID3, C4.5, CART (Classification and Regression Tree), CHAID (Chi-square Automatic Interaction Detection) [7]. Для изучения эффективности классификации в приложении к задаче локального прогнозирования газодинамических явлений в данной работе используется реализация алгоритма построения деревьев решений CART из библиотеки алгоритмов машинного обучения Scikit-learn 1.6.1. Математический аппарат деревьев решений также сравнительно прост и строится на разделении множества входных данных на подмножества в зависимости от показателей информационных критериев – в случае классификации критерия неопределенности Джини [8].

Пусть данные в узле дерева t представлены значениями Q_t с n_t элементов. Для каждого потенциального разделения в узле $\theta = (j, t_m)$ состоит из признака j и предела признака t_m , чьи показатели и разделяют данные на левое Q_t^l и правое Q_t^r подмножества:

$$Q_t^l(\theta) = \{(x, y) | x_j \leq t_m\} \quad (1)$$

$$Q_t^r(\theta) = Q_t / Q_t^l(\theta) \quad (2)$$

Качество потенциального разделения в рассматриваемом узле t рассчитывается с помощью показателя функции неопределенности Джини:

$$Gini(Q_t) = 1 - \sum_{i=1}^n p_i^2, \quad (3)$$

где Q – результирующее множество, n – число классов в нем, p_i – вероятность i -го класса, выраженная как относительная частота примеров соответствующего класса. Значения показателя неоднородности Джини меняются от 0 до 1. Если показатель равен нулю, значит все образцы результирующего множества относятся к одному классу, в свою очередь, если показатель равен единице, то классы представлены в равных пропорциях и равновероятны. Оптимальным считается то разделение, для которого значение показателя неоднородности Джини минимизировано. После разделения исходного множества все операции повторяются в новообразованных подмножествах, пока не будет достигнута максимальная глубина дерева или не кончатся элементы.

Эффективность классификационной модели

Общая выборка данных, используемых для исследования эффективности классификации деревьями решений, включала в себя 245 наблюдений по всей площади шахтного поля калийного рудника с 12 следующими горно-геологическими параметрами: средняя газоносность силвинитовых пород по свободным газам, среднее

содержание метана в свободных газах, среднее содержание водорода в свободных газах, среднее суммарное содержание тяжелых углеводородов (C_2-C_5) в свободных газах, средняя газоносность сильвинитовых пород по связанным газам, среднее содержание метана в связанных газах, среднее содержание водорода в связанных газах, среднее суммарное содержание тяжелых углеводородов (C_2-C_5) в свободных газах, среднее содержание хлорида калия (KCl) в сильвинитовых породах, среднее содержание хлорида магния ($MgCl_2$) в сильвинитовых породах, среднее суммарное содержание сульфата кальция ($CaSO_4$) и нерастворимого остатка, средняя мощность сильвинитового слоя промышленного пласта. Для обучения модели из общей выборки были отобраны 84 наблюдения, а для тестовой выборки использовались оставшийся объем в 161 наблюдение, что является близким к оптимальному соотношению данных (30% для обучения к 70% для тестирования) при использовании одной исходной выборки.

Важнейшим параметром при обучении классификационной модели дерева решений является максимальная глубина дерева, поскольку качественная модель должна иметь наибольшую эффективность и при этом не быть чрезмерно сложной. Максимальной глубиной дерева называется количество уровней (итераций алгоритма) в получаемой классификационной модели. В качестве метрик эффективности созданных моделей классификации были использованы точность (англ. Precision или Positive Predictive Value) и полнота (англ. Recall). Точностью называют долю истинно положительных классификаций модели к общему числу положительных классификаций модели. Полнота же определяется как число истинно положительных классификаций модели по отношению к общему числу положительных наблюдений в исходной выборке, данный показатель можно называть способностью классификационной модели обнаруживать определенный класс. Показатели точности и полноты для моделей деревьев решений с различной максимальной глубиной представлены в таблице 1.

Таблица 1

Сводка показателей точности и полноты обученных моделей

Максимальная глубина дерева, уровней	2	3	4	5	6	7	8
Точность модели, д. ед.	0,228	0,66	0,33	0,36	0,44	0,39	0,36
Полнота модели, д. ед.	0,76	0,47	0,88	0,94	1	1	1

В зависимости от поставленной задачи необходимо отдавать приоритет точности или полноте модели. В данном случае, в задаче локального прогнозирования газодинамических явлений, точность является приоритетным показателем в силу необходимости минимизации ложных положительных прогнозов. Как видно из таблицы, максимальная точность может быть достигнута при значении максимальной глубины в 7 уровней. Дальнейшее увеличение максимальной глубины дерева решений нецелесообразно из-за чрезмерного усложнения и дальнейшего падения показателя точности классификационной модели. Матрица ошибок полученной модели представлена в таблице 2. Визуализация полученного дерева решений представлена на рисунке 1.

Таблица 2

Матрица ошибок наиболее эффективной классификационной модели

Группы наблюдений		Фактические данные	
		Опасно (Danger)	Неопасно (Safe)
Модель	Опасно (Danger)	17	21
	Неопасно (Safe)	0	207

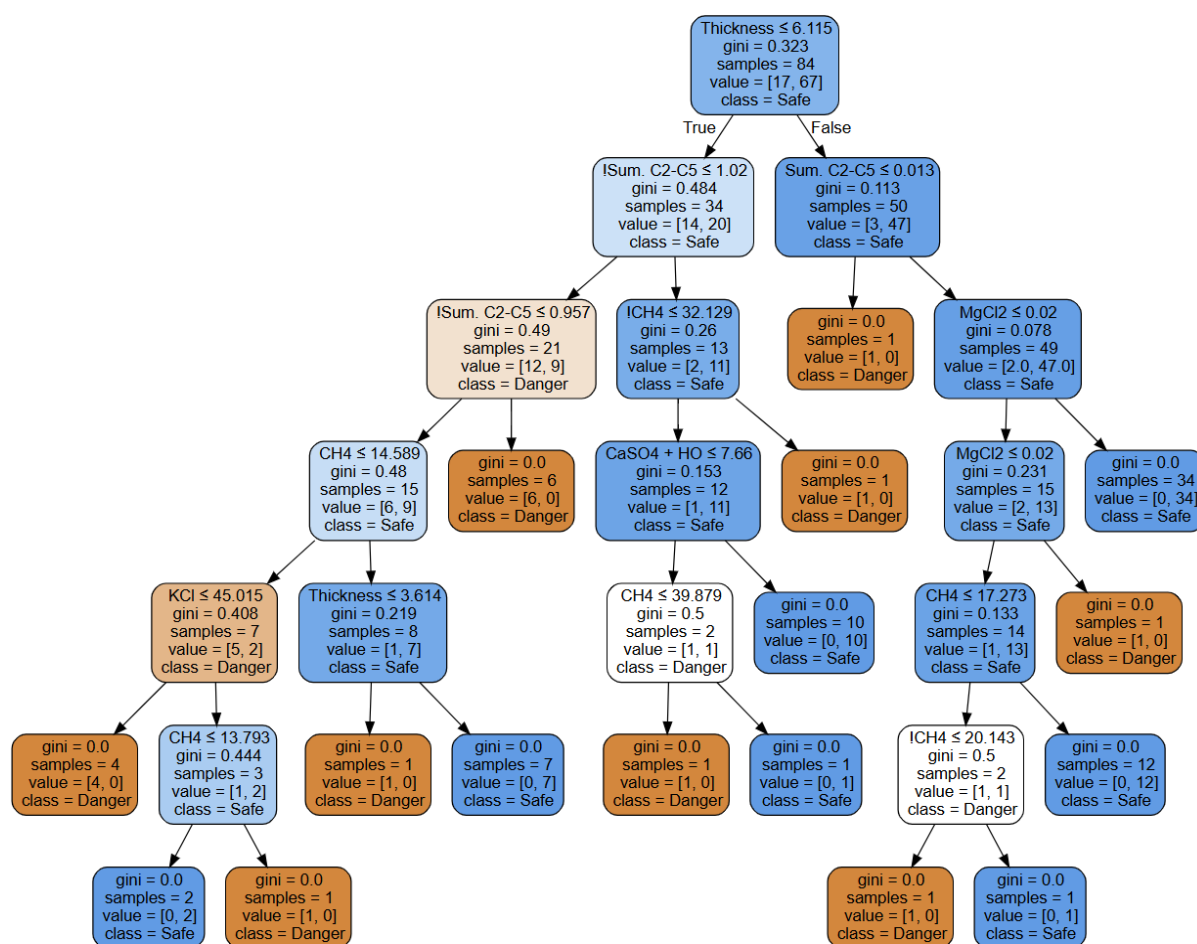


Рис. 1. Визуализация наиболее эффективной классификационной модели

Из всех входных параметров в результате обучения наиболее информативными для дискриминации по классам оказались: мощность сильвинитового пласта (Thickness), среднее суммарное содержание тяжелых углеводородов в свободных газах сильвинитовых пород (Sum. C₂-C₅), среднее суммарное содержание тяжелых углеводородов в связанных газах сильвинитовых пород (!Sum. C₂-C₅), среднее содержание метана в связанных газах сильвинитовых пород (!CH₄), среднее суммарное со-

держание сульфата кальция и нерастворимого остатка в сильвинитовых породах ($\text{CaSO}_4 + \text{HO}$), среднее содержание хлорида магния в сильвинитовых породах (MgCl_2), среднее содержание хлорида калия в сильвинитовых породах (KCl), среднее содержание метана в свободных газах сильвинитовых пород (CH_4). Применение подобной модели сводится к подаче данных на верхний узел (лист) и движение по дереву решений до тех пор, пока мы не окажемся в каком-либо из крайних узлов, предсказанный моделью класс подаваемых данных будет соответствовать классу конечного узла. Полученная модель показала высокую эффективность на тестовом наборе данных, однако известная высокая чувствительность к входным данным и ограниченная выразительность не позволяет в полной мере произвести анализ взаимосвязей параметров между собой. В свою очередь, использование деревьев решений в ансамбле с другими алгоритмами распознавания образов позволит повысить эффективность работы комплексных решений задачи классификации при локальном прогнозировании газодинамических явлений.

Выводы

Созданная на основе дерева решений модель классификации показала высокую эффективность при максимальной точности и сравнительно высоком значении полноты при максимальной глубине в 6 уровней. Однако известные недостатки деревьев решений в виде высокой чувствительности к выбросам и высокой нестабильности не позволяет в полной мере произвести анализ взаимосвязей между горно-геологическими параметрами, что ограничивает использование данного типа моделей лишь в качестве дополнительного алгоритма в более комплексном ансамбле алгоритмов для решения задачи классификации.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Барях А.А., Андрейко С.С., Федосеев А.К. О механизме локализации очагов газодинамических явлений в почве сильвинитовых пластов // Вестн. ПНИПУ: Геология. Нефтегазовое и горное дело. – 2017. – Т. 16, № 3. – С. 247-254. – DOI: 10.15593/2224-9923/2017.3.5.
2. Андрейко С.С., Лядов В.О., Папулов А.С., Абашин В.И. Анализ геологических условий проявления газодинамических явлений Гремячинского ГОКа // Изв. Тульского гос. ун-та. Науки о Земле. – 2024. – № 1. – С. 465-474.
3. Иванов О.В., Нестерова С.Ю., Лядов В.О., Лукьянец Е.В. Региональное прогнозирование опасных по газодинамическим явлениям зон в условиях шахтного поля рудника Талицкого ГОКа // Горный журнал. – 2023. – № 11. – С. 84-88. – DOI: 10.17580/gzh.2023.11.14.
4. Лядов В.О. Актуализация регионального прогноза зон, опасных по газодинамическим явлениям, в условиях Гремячинского месторождения калийных солей // Горное эхо. – 2024. – № 2 (95). – С. 66-71. – DOI: 10.7242/echo.2024.2.12.
5. Лядов В.О. Оценка эффективности применения нейронных сетей при региональном прогнозировании зон газодинамической опасности в калийных рудниках // Актуальные проблемы недропользования: тез. докл. XIX Междунар. форум-конкурса студентов и молодых ученых / Санкт-Петерб. Горный ун-т. – СПб., 2023. – Т. 1. – С. 195-197.
6. Brodley C.E., Utgoff P.E. Multivariate decision trees // Machine learning. – 1995. – V. 19. – P. 45-77.
7. Jijo B.T., Abdulazeez A.M. Classification based on decision tree algorithm for machine learning // Journal of applied science and technology trends. – 2021. – V. 2. – № 1. – P. 20-28. – DOI: 10.38094/jastt20165.
8. Кисляков А.Н. Алгоритм бинарной классификации на основе графов принятия решений в задачах кредитного скоринга // Модели, системы, сети в экономике, технике, природе и обществе. – 2021. – № 1. – С. 29-41.