УДК 004.852

Восстановление каротажных кривых методами машинного обучения

Зарипова Э. А., Еникеев М. Р.

Уфимский государственный нефтяной технический университет

В докладе рассматриваются алгоритмы машинного обучения для восстановления кривых акустического и плотностного каротажа, на основе другой имеющейся скважинной информации. Выбор именно этих каротажей обосновывается тем, что на их основе можно предсказать литологию.

В статистике хорошо известно интуитивное соображение, согласно которому усреднение результатов наблюдений может дать более устойчивую и надежную оценку, поскольку ослабляется влияние случайных отклонений в отдельном измерении. На аналогичной идее было основано развитие алгоритмов комбинирования моделей, в результате чего построение их ансамблей оказалось одним из самых мощных методов машинного обучения, нередко превосходящим по качеству предсказаний другие методы.

Методом улучшения предсказаний является бустинг (boosting), идея которого заключается в итеративном процессе последовательного построения частных моделей. Каждая новая модель обучается с использованием информации об ошибках, сделанных на предыдущем этапе, а результирующая функция представляет собой линейную комбинацию всего ансамбля моделей с учетом минимизации любой штрафной функции.

В ходе работы сравнивается работа четырех методов обучения, основанных на градиентном бустинге.

Суть метода нейронных сетей заключается в том, что в опорной скважине обучаются нейроны, а именно определяются зависимости между исходными данными и выходной диаграммой. Затем устанавливается возможность использования этих зависимостей для расчета синтетических кривых в ответ на входные данные, схожие, но неидентичные тем, что были использованы при обучении.

Процедура восстановления каротажных кривых заключается:

- Загрузка и подготовка данных,
- Вычисление корреляционной матрицы,
- Нормализация кривых,
- Создание тренировочной и тестовой выборок для валидации получаемого решения,
- Обучение GradientBoostingRegressor(), GridSearchCV(),
- Обучение LGBMRegressor(), GridSearchCV(),
- Обучение XGBoostRegressor(), GridSearchCV(),
- Обучение CatBoostRegressor(), GridSearchCV(),
- Расчет метрик и оценка результата,
- Анализ полученных результатов.

Загрузка и подготовка данных.

На вход подаются данные 286 скважин, чему соответствует 421469 точек (рис. 1). Осуществляется обработка данных:

- Проверка/заполнение пропусков,
- Удаление аномальных данных.

В ходе работы были удалены данные четырех скважин, поскольку они имели разные области определения, что сильно ухудшает результат. Методом подбора пар признаков основой базового решения выбраны каротажи GR, NEU, RT и $RES - DEP$, а целевыми каротажами - DT и DEN - акустическая скорость и плотность, потому что они дали самый лучший результат.

	BS	CALI	DEN	DEPT	DT	GR	NEU	PE	RES_DEP	RES_MED	RES_MIC	RES_SLW	RT	SP	Stratigraphy	Wellname
0	NaN	230.473404	2.618527	2571.60	283.249512	121.194298	0.235311	NaN	4.007868	3.347113	NaN	2.462722	4.136996	40.374001	7.0	0_2
1	NaN	230.293793	2.621109	2571.65	283.019104	123.683601	0.233350	NaN	3.980783	3.329070	NaN	2.467258	4.104412	40.353569	7.0	0_2
2	NaN	230.113403	2.623705	2571.70	282.787506	126.185204	0.231378	NaN	3.953749	3.311036	NaN	2.471825	4.071928	40.333031	7.0	0_2
3	NaN	229.932907	2.626300	2571.75	282.555908	128.686798	0.229407	NaN	3.926900	3.293100	NaN	2.476400	4.039700	40.312500	7.0	0_2
4	NaN	230.266403	2.632016	2571.80	280.377106	126.202202	0.232660	NaN	3.911065	3.286945	NaN	2.491029	4.030546	40.148239	7.0	0_2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
421464	220.699997	220.235992	2.584000	2730.90	245.781006	111.054001	0.271000	NaN	3.244000	3.166000	NaN	3.459000	3.090000	NaN	13.0	2_555
421465	220.699997	221.149002	2.582000	2730.95	246.509995	107.334999	0.267000	NaN	3.231000	3.182000	NaN	3.452000	3.093000	NaN	13.0	2_555
421466	220.699997	221.826004	2.613000	2731.00	247.238007	104.625999	0.269000	NaN	3.218000	3.197000	NaN	3.440000	3.094000	NaN	13.0	2_555
421467	220.699997	222.169006	2.640000	2731.05	247.968994	103.586998	0.283000	NaN	3.204000	3.211000	NaN	3.428000	3.098000	NaN	13.0	2_555
421468	220.699997	222.279007	2.639000	2731.10	248.695999	105.808998	0.297000	NaN	3.188000	3.219000	NaN	3.411000	3.094000	NaN	13.0	2_555

421469 rows × 16 columns

Рис. 1. Входные данные по каротажным кривым для тестового месторождения

Вычисление корреляционной матрицы. Корреляционная матрица - таблица, в которой определяется взаимосвязь признаков датафрейма. Такая матрица используется для того, чтобы определить какие признаки линейно зависят между собой.

На рисунке 2 представлена корреляционная матрица параметров каротажных кривых для тестового месторождения.

	BS	CALI	DEN	DEPT	DT	GR	NEU	PE	RES_DEP	RES_MED	RES_MIC	RES_SLW	RT	SP	Stratigraphy	GR/NEU
BS	1	0.026	0.0081	0.0095	-0.032	0.0021	-0.026	-0.0021	0.0056	-0.015	0.0098	-0.019	0.0074	-0.11	-0.014	0.021
CALI	0.026	1	0.1	0.024	0.012	0.13	0.022	0.21	-0.01	-0.0038	-0.029	-0.0011	-0.015	0.04	0.053	0.012
DEN	0.0081	0.1	1	-0.00058	-0.16	0.52	-0.046	0.86	0.03	0.058	0.087	0.067	0.0083	0.054	0.3	0.091
DEPT	0.0095	0.024	-0.00058	1	-0.077	0.018	0.019	-0.0079	0.011	0.0059	-0.0019	0.004	0.018	-0.045	0.099	0.00024
DT	-0.032	0.012	-0.16	-0.077	1	0.4	0.66	-0.044	-0.3	-0.28	-0.31	-0.27	-0.3	0.049	-0.026	-0.076
GR	0.0021	0.13	0.52	0.018	0.4	1	0.53	0.65	-0.19	-0.16	-0.18	-0.15	-0.21	0.067	0.14	0.039
NEU	-0.026	0.022	-0.046	0.019	0.66	0.53	1	0.078	-0.34	-0.32	-0.36	-0.31	-0.33	0.034	0.027	-0.11
PE	-0.0021	0.21	0.86	-0.0079	-0.044	0.65	0.078	1	0.056	0.082	0.11	0.086	0.034	0.055	0.29	0.45
RES_DEP	0.0056	-0.01	0.03	0.011	-0.3	-0.19	-0.34	0.056	1	0.66	0.48	0.5	0.86	-0.002	0.01	0.081
RES_MED	-0.015	-0.0038	0.058	0.0059	-0.28	-0.16	-0.32	0.082	0.66	1	0.44	0.79	0.59	-0.0018	0.0036	0.081
RES_MIC	0.0098	-0.029	0.087	-0.0019	-0.31	-0.18	-0.36	0.11	0.48	0.44	1	0.46	0.48	-0.0074	-0.018	0.48
RES_SLW	-0.019	-0.0011	0.067	0.004	-0.27	-0.15	-0.31	0.086	0.5	0.79	0.46	1	0.48	0.0062	-0.00037	0.095
RT	0.0074	-0.015	0.0083	0.018	-0.3	-0.21	-0.33	0.034	0.86	0.59	0.48	0.48	1	-0.0024	0.013	0.069
SP	-0.11	0.04	0.054	-0.045	0.049	0.067	0.034	0.055	-0.002	-0.0018	-0.0074	0.0062	-0.0024	1	0.058	0.038
Stratigraphy	-0.014	0.053	0.3	0.099	-0.026	0.14	0.027	0.29	0.01	0.0036	-0.018	-0.00037	0.013	0.058	1	0.013
GR/NEU	0.021	0.012	0.091	0.00024	-0.076	0.039	-0.11	0.45	0.081	0.081	0.48	0.095	0.069	0.038	0.013	1

Рис. 2. Корреляционная матрица между каротажными кривыми. Красный цвет - корреляция между параметрами отсутствует, синий цвет - наблюдается линейная зависимость между параметрами

Нормализация кривых.

Нормализация каротажных кривых на основании распределений по опорной скважине позволяет привести кривые ГИС к одному масштабу измерения. Преобразование одной

кривой в единицы измерения другой производится с помощью регрессионного уравнения. Уравнение регрессии связи показаний рассматриваемых методов строится по опорным пластам. Эти пласты связаны общим свойством, влияние второго свойства отсутствует. Нормализация производится по опорным пластам, и в этом случае будет отмечаться расхождение кривых в интервалах разреза, обладающих вторым свойством.

Процесс нормализации осуществляется в два этапа: находится уравнение нормализации, затем проводится сама нормализация. В большинстве случаев уравнение нормализации между двумя кривыми представлено в виде прямой $y = a + bx$, где y – базовая кривая, масштаб которой не меняется; x – трансформируемая кривая; и b – коэффициенты, означающие сдвиг нулевых линий между кривыми и изменение масштаба нормализуемой кривой.

Используется два вида трансформации – логарифмический $\lg(x)$ и гиперболический $\frac{1}{\sqrt{x}}$, позволяющие привести уравнение нормализации к приближенному уравнению прямой линии. Коэффициенты и b определяются по опорным пластам.

В ходе работы применяется стандартная нормировка StandardScaler() с логарифмической трансформацией и линейное нормирование.

Создание тренировочной и тестовой выборок для валидации получаемого решения.

Тренировочный набор - поднабор для того, чтобы тренировать модель.

Тестовый набор - поднабор для тестирования модели.

Тестовый набор должен быть достаточно большим, чтобы результат был статистически достоверным.

Разбиваем данные скважин на тренировочную и тестовую части в пропорции 0,8 и 0,2 соответственно.

Обучение.

GradientBoostingRegressor - ГБ строит аддитивную модель в прямом поэтапном способе, что позволяет оптимизировать произвольные дифференцируемые функции потерь. На каждом этапе дерево регрессии помещается на отрицательный градиент учитывая функцию потери.

LGBMRegressor – это платформа для повышения градиента, использующая алгоритмы обучения на основе дерева. Он предназначен для распространения и эффективности со следующими преимуществами: более быстрая скорость обучения и высокая эффективность, более низкое использование памяти, лучшая точность, поддержка параллельного и GPU обучения, способен обрабатывать крупномасштабные данные

XGBoost - это оптимизированная распределенная библиотека повышения градиента, разработанная для обеспечения высокой эффективности , гибкости и переносимости . Он реализует алгоритмы машинного обучения в рамках GradientBoosting . XGBoost обеспечивает параллельное усиление дерева (также известное как GBDT, GBM), которое быстро и точно решает многие проблемы с данными.

CatBoost - создает решающую модель прогнозирования в виде ансамбля слабых моделей прогнозирования, обычно деревьев решений. Он строит модель поэтапно, позволяя оптимизировать произвольную дифференцируемую функцию потерь.

Обьект *GridSearchCV()* занимается подбором оптимальных параметров для модели из заданного списка.

Расчет метрик и оценка результата.

Алгоритм расчета финальной метрики: для каждой скважины рассчитывается результирующий коэффициент детерминации (для каждого каротажа в отдельности - акустического и плотностного), далее в результате осреднения получаются средние значения R^2 по всем скважинам для каждого каротажа (2 числа), которые в дальнейшем также усредняются.

Анализ полученных результатов. Самую высокую точность предсказания дал ме-

	GradientBoosting		*LGBM*		*CatBoost*		*XGBoost*	
	training	testing	training	testing	training	testing	training	testing
$\mathbf{R^2}$	0.43	0.65	0.57	0.76	0.39	0.74	0.47	0.42
Best_params	learning_rate: 0.1, n_estimators: 60		learning_rate: 0.1, n_estimators: 60, num_leave: 50		depth: 4, eval_metric: RMSE, leaf_estimation_iterations: 4, learning_rate: 0.05		colsample_bytree: 0.8, learning_rate: 0.31, max_depth: 1, min_child_weight: 1, n_estimator: 60, subsample: 1	

Рис. 3. Результаты целевой метрики

тод LGBMRegressor() - 0.76 с параметрами скорость обучения - 0.1, количество деревьев - 60, максимальное количество листьев дерева - 50. Таким образом, в дальнейшем для улучшения точности прогнозирования параметры методов будут модифицироваться.

Литература

1. Еникеев М. Р., Фазлытдинов М. Ф., Еникеева Л. В., Губайдуллин И. М. Прогноз обводненности на проектируемых к бурению скважинах методами машинного обучения. Сборник трудов ИТНТ-2019: V междунар. конф. и молодеж. шк. «Информ. технологии и нанотехнологии». Самара: Новая техника. 2019. Т. 4: Науки о данных. 2019. С. 434-444.

2. Косков В. Н., Косков В. В. Геофизические исследования скважин и интерпретация данных ГИС. Учеб. пособие. Пермь: Перм. гос. техн. ун-т. 2007. 122 с.

3. Еникеев М. Р., Губайдуллин И. М., Малеева М. А. Информационно-вычислительная аналитическая система для оценки и прогнозирования коррозионных процессов на поверхности стали и алюминия // Системы и средства информатики. 2017. Т. 27. № 3. С. 155-170

4. Заворотько Ю. М. Геофизические методы исследования скважин. Москва: «Недра»; 1983. 208 с.

MSC2020 68T05

Recovery of logging curves using machine learning methods

E. A. Zaripova, M. R. Enikeev

Ufa State Petroleum Technological University [1]