МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, ЧИСЛЕННЫЕ МЕТОДЫ И КОМПЛЕКСЫ ПРОГРАММ. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Игорь Александрович МУРАВЬЕВ[1]
Ирина Гелиевна ЗАХАРОВА[2]

УДК 004.8.032.26

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ТЕРРИГЕННЫХ КОЛЛЕКТОРОВ ПО ХАРАКТЕРУ НАСЫЩЕНИЯ

[1] старший преподаватель кафедры программного обеспечения,
Тюменский государственный университет
to.imuravev@gmail.com

[2] кандидат физико-математических наук,
профессор кафедры программного обеспечения,
Тюменский государственный университет
i.g.zakharova@utmn.ru

Аннотация

Выявление свойств нефтегазовых коллекторов на основе информации, получаемой в результате геофизических исследований скважин, является одним из главных направлений исследований в области геологического и гидродинамического моделирования пласта. Недостаточная эффективность точных математических моделей для анализа данных геофизических исследований скважин, а также большой объем и зашумленность этих данных определяют актуальность использования методов машинного обучения

Цитирование: Муравьев И. А. Исследование возможностей методов машинного обучения для классификации терригенных коллекторов по характеру насыщения / И. А. Муравьев, И. Г. Захарова // Вестник Тюменского государственного университета. Физико-математическое моделирование. Нефть, газ, энергетика. 2019. Том 5. № 1. С. 123-137. DOI: 10.21684/2411-7978-2019-5-1-123-137

для выявления особенностей коллекторов. В статье исследованы возможности классификации терригенных коллекторов с помощью различных методов (метод опорных векторов, дерево решений, градиентный бустинг, случайный лес, многослойная нейронная сеть и др.). Набор данных был сформирован на основе каротажных кривых для 24 скважин одной залежи. Для обучения моделей классификации были использованы предварительно нормированные данные индукционного каротажа, бокового каротажа, нейтрон-нейтронного каротажа по тепловым нейтронам, электрометрии с помощью потенциал-зондов, резистивиметрии, каротажа потенциалов самопроизвольной поляризации, гамма-каротажа и каротажа сопротивления с использованием пяти различных последовательных градиент-зондов. Для оценки точности моделей классификации, построенных различными методами, в каждом случае выполнялась кросс-валидация, оценивалось среднее значение точности и стандартное отклонение. Для метода опорных векторов исследовалось влияние выбора функции ядра (линейная, полиномиальная, сигмоид). В случае нейронной сети варьировались: ее архитектура, включая число скрытых слоев и нейронов, функции активации на различных слоях, а также вероятность дропаута. Качество полученных моделей классификации оценивалось также по значениям элементов матрицы несоответствия. Результаты вычислительных экспериментов показали результативность использования методов машинного обучения, в частности многослойных нейронных сетей, для выявления с высокой точностью (около 90%) коллекторов с нефтью.

Ключевые слова

Терригенный коллектор, геофизические исследования скважин, каротажные кривые, классификация, машинное обучение, нейронная сеть, вычислительный эксперимент.

DOI: 10.21684/2411-7978-2019-5-1-123-137

Введение

Основной целью прикладных геофизических исследований в контексте геологической разведки является поиск залежей полезных ископаемых, в частности углеводородного сырья. Одним из способов изучения строения и состава верхнего слоя земной коры для выявления нефтегазоносных пластов являются геофизические исследования скважин (ГИС). ГИС — это целый комплекс методов для изучения разреза скважины, получивший название «каротаж». Геофизические исследования скважин позволяют получить разностороннюю информацию, отражающую широкий спектр параметров определенной области пласта вблизи ствола скважины, — от температуры, радиоактивности, удельной электропроводности до скорости распространения упругих волн и др. Как следствие, в центре внимания исследователей оказывается задача интерпретации данных ГИС с целью выявления особенностей пласта для его дальнейшего изучения и разработки. При этом основные массивы данных, подлежащих интерпретации, поступают благодаря геофизическим исследованиям скважин, но необходимо принимать во внимание также априорные знания о территории, на которой

производится разведка. Это определяет сложность и нетривиальность построения решения задачи интерпретации данных, поскольку зависимости значений геофизических параметров от показателей зондов носят неочевидный характер [3]. Погрешности измерений, связанные с внешними условиями, естественным шумом, деформациями кабеля и т. п., обуславливают дополнительную неопределенность интерпретации. Наконец, для интерпретации специалисты опираются не только на свои знания о специфике территории и конкретного пласта, на котором проводятся исследования, но и свой обобщенный опыт решения подобных задач.

Таким образом, можно заключить, что речь идет об исследовании ключевой задачи data mining, которая состоит в извлечении новых знаний из имеющихся данных (в нашем случае — данных ГИС). Для исследования подобных задач все шире в последнее время применяются методы машинного обучения. Так, в работе [2] Д. О. Гафуров предлагает использовать аппарат искусственных нейронных сетей для выделения литотипов в разрезе скважины для построения геологической модели. Х. Б. Агаев [1] рассматривает метод расчленения геологического разреза по различным кластерам на основе нейронной сети для последующего принятия решения специалистом. Д. Ю. Чудинова, М. Р. Дулкарнаев и др. [5] исследуют применение нейросетевого подхода с целью дифференциации действующего фонда скважин для оценки его структуры и выявления причин низких дебитов скважин. Ф. Тан (F. Tan), Г. Ло (G. Luo) и др. [20] показывают возможности применения методов data mining и подходы к выбору различных пространств признаков для комплексной оценки сложных нефтяных пластов. В своей работе [6] В. Дж. Аль-Мудхафер (W. J. Al-Mudhafar) применяет нейронную сеть для определения литотипов для задачи моделирования проницаемости пористых систем. В статье [4] Н. Б. Паклин и Р. С. Мухамадиев решают задачи выделения пластов-коллекторов и определяют характер насыщения, используя такие методы машинного обучения, как деревья решений и самоорганизующиеся карты Кохонена.

В то же время особенности исходных данных и конечной цели интерпретации, в частности степень детализации при выполнении классификации, не дают универсального ответа в плане выбора конкретного метода. Поэтому целью настоящего исследования выступает сравнительный анализ различных методов машинного обучения для решения задачи обобщенной классификации коллекторов определенного типа, а именно терригенных.

Постановка задачи

На каждом этапе геологоразведочных работ для каждой скважины проводится оперативная интерпретация данных ГИС, целью которой является выделение пластов-коллекторов и оценка их свойств.

Определение типа флюида (нефть, газ, вода или их смесь), который заполняет коллектор, — одна из первостепенных задач, необходимых для получения достоверной модели залежи.

Известно, что существует несколько типов коллекторов: терригенный, карбонатный и др., для которых характерны свои геофизические параметры, и для определения типа флюидонасыщения необходимо использовать целые различные комплексы ГИС. Коллектор терригенного типа — один из наиболее часто встречаемых, поэтому в данной работе ограничимся определением характера флюида, заполняющего терригенный коллектор. Для определения типа флюида в терригенном коллекторе используются методы, в основу которых положена оценка удельного сопротивления породы. В самом простейшем случае можно говорить, что водоносные коллекторы имеют низкое сопротивление, в то время как нефтеносные — высокое. В частности, к методам каротажа для определения характера насыщения коллектора относят GZ (каротаж сопротивления) и BK (боковое каротажное зондирование).

В данной работе для определенности рассматривается однофазовая нефтяная залежь, таким образом, особенность насыщения характеризуется присутствием нефти, воды или их смесью. Тогда постановка исходной задачи будет выражена так: на основе данных комплекса геофизических исследований необходимо определить характер насыщения терригенного коллектора вдоль ствола скважины (вода, нефть или смесь воды и нефти).

Процесс интерпретации данных геофизических исследований скважин на практике реализуется следующим образом: специалист получает данные с зондов <*name* 1>, <*name* 2>, … , <*name N*>, где *N* — число каротажных исследований. Затем проводится анализ показаний в соответствии с рекомендациями и регламентами, после чего делается вывод о составе флюида, насыщающего коллектор. Далее процесс повторяется вдоль всего ствола скважины.

Принимая во внимание все сказанное выше, можно заключить, что речь идет о задаче классификации, которую можно сформулировать следующим образом: задан определенный набор объектов, характеризующихся некоторым вектором признаков, и меток-классов, к которым эти объекты могут относиться; каждому объекту из набора необходимо выставить метку принадлежности к определенному классу.

Для формального описания необходимо представить основные элементы процесса в виде математических объектов. Так, данные ГИС — дискретная функция от одной переменной (глубина ствола скважины):

$$y = f_{name}(d), \tag{1}$$

где *name* — название ГИС, *d* — глубина вдоль ствола.

Интерпретацию реализует функция *F* от *N* переменных, возвращающая значения из множества $z \in \{1, 2, 3\}$, где 1 соответствует нефтяному насыщению, 2 — водному, а 3 — смешанному.

$$z = F(f_{name1}(d), f_{name2}(d), \dots, f_{nameN}(d)), \quad z \in \{1, 2, 3\}. \tag{2}$$

Необходимо построить классифицирующую функцию *F*.

Исходные данные

Для построения функции-классификатора было использовано 819 замеров данных ГИС по 24 скважинам одной залежи. Заранее специалистом была проведена разметка данных на основе обычной интерпретации, в частности был выделен терригенный коллектор и определен его характер насыщения. Для формирования пространства признаков были использованы результаты исследований, представленные в таблице 1.

Таблица 1 *Table 1*

Пространство признаков **The feature space of attributes**

Название	Ед. изм.	Среднее значение	Стандартное отклонение	Мин.-макс. значение
Индукционный каротаж (IK)	мСм/м	144,69	23,35	[64,83; 225,38]
Боковой каротаж (BK)	Ом · м	7,33	1,6	[4,61; 20,20]
Нейтрон-нейтронный каротаж по тепловым нейтронам (NNKTb)	у. е.	8,14	11,32	[2,10; 44,84]
Электрометрия с помощью потенциал-зондов (PZ)	Ом · м	10,9	2,61	[7,05; 38,54]
Резистивиметрия (RS)	Ом · м	1,01	0,09	[0,83; 1,33]
Каротаж потенциалов самопроизвольной поляризации (SP)	мВ	107,56	10,24	[83,73; 132,12]
Гамма-каротаж (GK)	у. е.	12,37	20,09	[3,00; 318,00]
Каротаж сопротивления последовательным градиент-зондом № 1 (GZ1)	Ом · м	8,17	1,88	[4,75; 30,71]
Каротаж сопротивления последовательным градиент-зондом № 2 (GZ2)	Ом · м	11,96	6,19	[4,97; 76,15]
Каротаж сопротивления последовательным градиент-зондом № 3 (GZ3)	Ом · м	9,85	4,64	[3,78; 52,96]
Каротаж сопротивления последовательным градиент-зондом № 4 (GZ4)	Ом · м	8,01	2,77	[2,47; 24,70]
Каротаж сопротивления последовательным градиент-зондом № 5 (GZ5)	Ом · м	8,34	3,54	[3,23; 29,90]

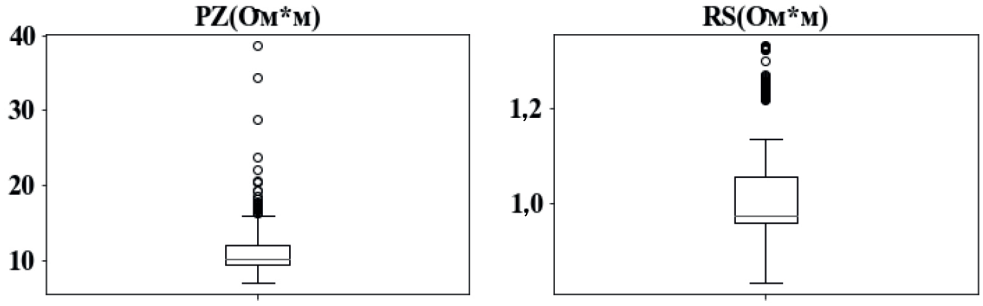

Рис. 1. Характерные графики размаха для обучающей выборки

Fig. 1. Characteristic box plots for a training set

Предварительная обработка данных

Важным аспектом при использовании алгоритмов машинного обучения является подготовка данных. На данном этапе необходимо выявить аномалии в данных, например, пропущенные значения или выбросы, так называемые outliners [14]. Эффективным инструментом обнаружения выбросов значений признаков служат графики размаха (box plot) [21]. На рис. 1 приведены графики, характерные для обучающей выборки. Выбросы на графиках отображаются точками. Для устранения выбросов необходимо преобразовать исходные данные. Одним из способов преобразования данных, позволяющих бороться с выбросами, является квантильное преобразование [10, 17]. Применение данного метода позволяет получить данные, удовлетворяющие заданному распределению.

Еще одним важным этапом является проверка параметров на взаимную корреляцию. Корреляционный анализ показал возможность удаления 4 признаков (GZ1 (Ом · м), IK (мСм/м), BK (Ом · м), NKTb (у. е.)), после чего матрица перестает содержать сильно зависимые признаки. Кроме того, можно скомбинировать зависимые и слабые признаки в сильные путем снижения размерности исходного пространства с помощью метода главных компонент PCA [16]. Благодаря этому число признаков исходной выборки было снижено с 12 до 8.

Методы решения задачи классификации

Для изучения применимости аппарата машинного обучения для описанной выше задачи использовались следующие методы:

- метод опорных векторов [19] с различными ядрами: линейным, полиномиальным, сигмоидом;
- метод *k*-ближайших соседей [7];
- дерево решений [8];
- градиентный бустинг на деревьях [13], случайный лес [9];
- нейронные сети [12].

Для каждого алгоритма выделялись основные параметры, влияющие на построение модели. Точность полученной модели оценивалась с помощью процедуры кросс-валидации [15].

Метод опорных векторов является одной из наиболее популярных моделей машинного обучения, суть которого состоит в разделении пространства признаков гиперплоскостями. Для построения разделяющих гиперплоскостей используются различные ядра, которые влияют на точность алгоритма. Еще один из параметров, влияющих на построение гиперплоскостей, — это значение штрафа за неправильно классифицированный объект.

Метод *k*-ближайших соседей является метрическим классификатором, который оценивает похожесть объектов. Зачастую именно данный метод используют для оценки других моделей машинного обучения. Самым главным параметром, влияющим на классификационную способность алгоритма, является число соседей, которые необходимы для принятия решения об отношении объекта к какому-либо классу.

Дерево принятия решений — это метод анализа данных и построения иерархической структуры зависимости выходных параметров от выходного. Самыми важными параметрами являются способ ветвления (т. е. на основании какого закона производить разбиение узла на части) и глубина дерева. Также следует сказать, выбор глубины дерева может защитить от явления переобучения. Рассматривались способы ветвления по критерию энтропии или индексу Джини.

Идея алгоритмов бустинга заключается в комбинировании более слабых моделей в один сильный классификатор. Из этого следует, что одним из параметров является количество простых моделей, участвующих в бустинге. В данной работе рассматривается бустинг с использованием деревьев. Таким образом, еще один параметр — это максимальная глубина дерева. В случае такого бустинга (с использованием деревьев) значение глубины рекомендуется устанавливать небольшим — менее 10. Рассмотрим результаты для случайного леса и градиентного бустинга.

Самым популярным методом машинного обучения на данный момент является нейронная сеть (НС). Существует множество архитектур НС, предназначенных для разных задач. Например, сверточные нейронные сети для анализа изображений, рекуррентные — для анализа текста или временных рядов. Результат классификации во многом зависит от архитектуры нейронной сети. В свою очередь, не существует универсального способа определить «правильную» структуру сети, хотя известны алгоритмы автоматической генерации структуры сети. Для определения архитектуры проводят эксперименты и эмпирически подбирают параметры. К числу таких параметров можно отнести:

— функцию активации,
— число нейронов на каждом слое,
— количество скрытых слоев.

Для повышения точности модели и предотвращения переобучения использовался метод dropout [18].

Результаты вычислительных экспериментов

Наилучший результат показал метод опорных векторов со значением штрафа 3 и 5 для линейного ядра (рис. 2). Средняя точность на валидационной выборке составила 0,88.

Алгоритм *k*-ближайших соседей (применялась метрика Минковского) показал наилучшую (0,99) точность при $k = 3$ (рис. 3).

На рис. 4 изображены зависимости от точности при кросс-валидации от максимальной глубины дерева для каждого критерия ветвления. Таким образом, самая точная модель смогла достичь значения в 0,89.

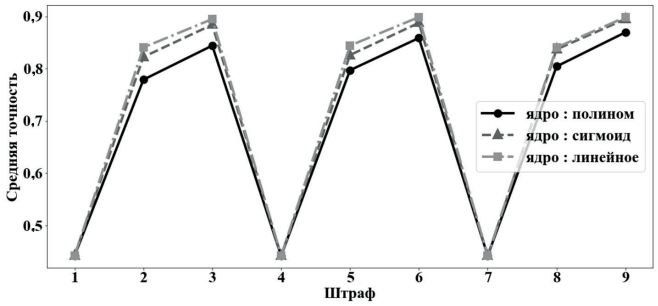

Рис. 2. Зависимость точности метода опорных векторов с различными ядрами от значения штрафа

Fig. 2. The accuracy of the support vectors method with different kernels depending on the value on the penalty value

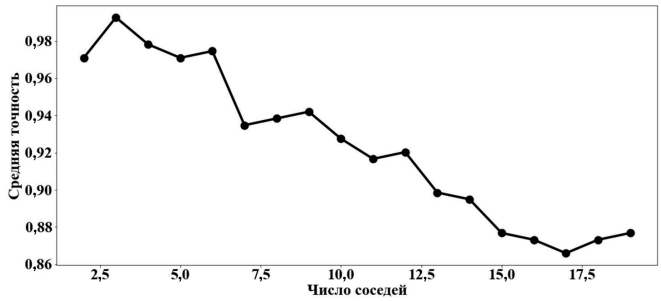

Рис. 3. Зависимость точности метода *k*-ближайших соседей от числа соседних объектов

Fig. 3. The accuracy of the method of *k*-nearest neighbors depending on the number of neighboring objects

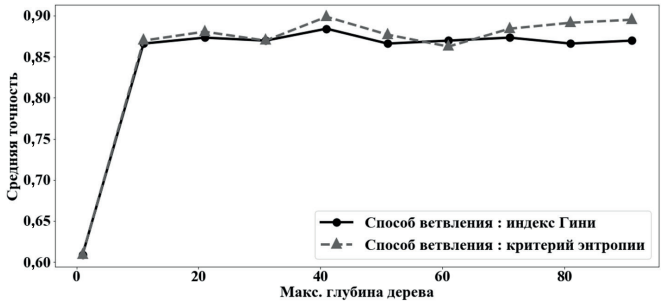

Рис. 4. Зависимость точности дерева решения от его глубины

Fig. 4. The accuracy of the decision tree depending on its depth

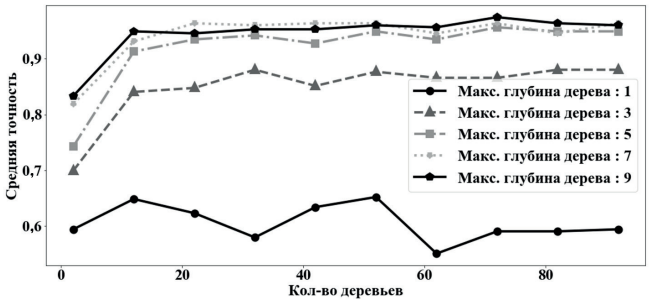

Рис. 5. Зависимость точности алгоритма случайного леса от количества деревьев для различных их глубин

Fig. 5. The accuracy of the random forest algorithm depending on the number of trees for their various depths

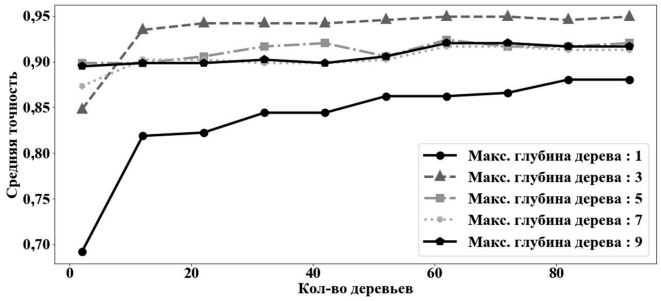

Рис. 6. Зависимость точности метода градиентного бустинга от количества деревьев для различных их глубин

Fig. 6. The accuracy of the gradient boosting method depending on the number of trees for their different depths

Для методов бустинга изучалось влияние количества деревьев на точность модели с заданным параметром глубины. График на рис. 5 отображает зависимость точности метода случайного леса от числа деревьев. Средняя точность 0,95 с наименьшим количеством деревьев 72 достигается при глубине 9. Метод градиентного бустинга показал точность в 0,94 со следующими параметрами: число деревьев — 62, их максимальная глубина — 3 (рис. 6).

Как показал вычислительный эксперимент, для решения задачи классификации объектов с фиксированным набором признаков можно ограничиться тремя скрытыми слоями. После первого слоя был применен метод dropout с порогом 0,25. Функция активации ReLU обеспечивает эффективное обучение сети. Для выходного слоя использовалась функция активации softmax, которая является традиционной для задачи классификации, отличной от бинарной. В результате вычислительных экспериментов была выбрана следующая конфигурация:

— слой из 16 нейронов, функция активации ReLU;
— dropout с вероятностью выброса 0,25;
— слой из 8 нейронов, функция активации ReLU;
— слой из 4 нейронов, функция активации ReLU;
— слой из 3 нейронов, функция активации softmax.

Функция штрафа — категориальная кросс-энтропия, метод оптимизации — Adam.

Полученные модели с оптимальными параметрами были протестированы на данных, не попавших в обучающую выборку. Итоговые результаты приведены в таблице 2.

Таким образом, можно заключить, что наилучший результат из числа рассмотренных методов показали модели с применением алгоритма *k*-ближайших соседей и нейронной сети. В случае классификации, отличной от бинарной, необходимо следить не только за точностью модели в целом, но и для каждого класса в отдельности. Визуально оценить распределение значений между классами можно с помощью матрицы несоответствий [11] (таблицы 3 и 4).

Таблица 2 *Table 2*

Результаты **Results**

Название	**Точность при кросс-валидации**	**Стандартное отклонение**	**Точность на тестовой выборке**
Метод опорных векторов	0,87	0,10	0,62
Метод *k*-ближайших соседей	0,99	0,10	0,76
Дерево решений	0,87	0,15	0,67
Случайный лес	0,89	0,10	0,60
Градиентный бустинг	0,90	0,04	0,61
Нейронная сеть	0,99	0,10	0,76

Таблица 3 *Table 3*

Матрица несоответствий для нейронной сети **Confusion matrix for the neural network**

	Нефть	Вода	Нефть + вода
Нефть	0,89	0,02	0,09
Вода	0,04	0,62	0,34
Нефть + вода	0,28	0,05	0,67

Таблица 4 *Table 4*

Матрица несоответствий для метода *k*-ближайших соседей **Confusion matrix for the method of *k*-nearest neighbors**

	Нефть	Вода	Нефть + вода
Нефть	0,79	0,04	0,17
Вода	0,05	0,82	0,13
Нефть + вода	0,20	0,08	0,72

Исходя из полученных результатов, можно сделать вывод о том, что классификаторы достаточно уверенно определяют нефтяные коллекторы. Однако классификаторы «путают» чистый флюид и смесь нефти с водой. Физически это объяснимо, поскольку смесь нефти и воды может быть в различных объемах.

Заключение

Проведенный вычислительный эксперимент показал возможность применения методов машинного обучения в задаче интерпретации данных ГИС. Полученные результаты свидетельствуют о том, что классификаторы с высокой долей вероятности определяют нефтяные коллекторы, но слабо дифференцируют чистый флюид и смесь. Поэтому данные методы можно использовать для выделения коллекторов только с присутствием нефти (т. е. сведение задачи к бинарной классификации) для поддержки принятия решения в качестве экспресс-интерпретации. Еще несколькими вариантами устранения указанного недостатка являются: а) увеличение числа классов, что позволяет разделять смеси с учетом пропорций флюидов; б) использование дополнительных методов обработки выхода классификатора, например, оценивания характера насыщения комплексно для всего коллектора или использование скользящего окна для устранения ошибочных значений.

СПИСОК ЛИТЕРАТУРЫ

1. Агаев Х. Б. Применение кластерного анализа для расчленения геологического разреза по данным каротажа скважины / Х. Б. Агаев // Каротажник. 2013. № 5 (227). С. 3-11.
2. Гафуров Д. О. Геологическая интерпретация с применением обучаемых нейронных сетей в «НейроИнформГео» данных ГИС Талаканского нефтегазоконденсатного месторождения / Д. О. Гафуров // Известия Томского политехнического университета. 2006. Том 309. № 3. С. 32-37.
3. Косков В. Н. Геофизические исследования скважин и интерпретация данных ГИС: учеб. пособие / В. Н. Косков, Б. В. Косков. Пермь: Изд-во Пермского государственного технического университета, 2007. 317 с.
4. Паклин Н. Б. Использование обучающихся алгоритмов для интерпретации данных ГИС / Н. Б. Паклин, Р. С. Мухамадиев // Бурение и нефть. 2005. № 5. С. 38-40.
5. Чудинова Д. Ю. Дифференциация скважин в зонах с остаточными запасами нефти с использованием нейросетевого моделирования / Д. Ю. Чудинова, М. Р. Дулкарнаев, Ю. А. Котенев, Ш. Х. Султанов // Экспозиция Нефть Газ. 2017. № 4 (57). С. 46-50.
6. Al-Mudhafar W. J. Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms / W. J. Al-Mudhafar // Journal of Petroleum Exploration and Production Technology. 2017. Vol. 7. No 4. Pp. 1023-1033. DOI: 10.1007/s13202-017-0360-0
7. Altman N. S. An introduction to kernel and nearest-neighbor nonparametric regression / N. S. Altman // The American Statistician. 1992. Vol. 46. No 3. Pp. 175-185. DOI: 10.1080/00031305.1992.10475879

8. Breiman L. Classification and Regression Trees / L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen. New York: Routledge, 2017. 368 p. DOI: 10.1201/9781315139470
9. Breiman L. Random forests / L. Breiman // Machine Learning. 2001. Vol. 45. No 1. Pp. 5-32. DOI: 10.1023/A:1010933404324
10. Compare the effect of different scalers on data with outliers // Scikit-learn: Machine Learning in Python. URL: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html (дата обращения: 27.12.2018).
11. Fawcett T. An introduction to ROC analysis / T. Fawcett // Pattern Recognition Letters. 2006. Vol. 27. No 8. Pp. 861-874. DOI: 10.1016/j.patrec.2005.10.010
12. Hagan M. T. Neural Network Design / M. T. Hagan, H. B. Demuth, M. H. Beale. Boston: PWS Publishing Company, 1996.
13. Hastie T. Boosting and additive trees / T. Hastie, R. Tibshirani, J. Friedman // The Elements of Statistical Learning. 2nd edition. New York: Springer, 2009. Ch. 10. Pp. 337-384. DOI: 10.1007/978-0-387-84858-7_10
14. Iglewicz B. How to Detect and Handle Outliers / B. Iglewicz, D. C. Hoaglin. American Society for Quality Control (ASQC), Statistics Division. 1993. 87 p.
15. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection / R. Kohavi // Proceedings of the 14th International Joint Conference on Artificial Intelligence. 1995. Vol. 2. Pp. 1137-1145.
16. Pearson K. On lines and planes of closest fit to systems of points in space / K. Pearson // The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901. Ser. 6. Vol. 2. Pp. 559-572. DOI: 10.1080/14786440109462720
17. Quantile transformer // Scikit-learn: Machine Learning in Python. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html (дата обращения: 27.12.2018).
18. Srivastava N. Dropout: a simple way to prevent neural networks from overfitting / N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov // Journal of Machine Learning Research. 2014. Vol. 15. Pp. 1929-1958.
19. Suykens J. A. K. Least squares support vector machine classifiers / J. A. K. Suykens, J. Vandewalle // Neural Processing Letters. 1999. Vol. 9. No 3. Pp. 293-300. DOI: 10.1023/A:1018628609742
20. Tan F. Evaluation of complex petroleum reservoirs based on data mining methods / F. Tan, G. Luo, D. Wang, Y. Chen // Computational Geosciences. 2017. Vol. 21. No 1. Pp. 151-165. DOI: 10.1007/s10596-016-9601-4
21. Tukey J. W. Exploratory Data Analysis / J. W. Tukey. Pearson, 1977. 712 p.

Igor A. MURAVEV[1]
Irina G. ZAHAROVA[2]

UDC 004.8.032.26

STUDYING THE CAPABILITIES OF MACHINE LEARNING METHODS FOR THE CLASSIFICATION OF THE CHARACTER OF SATURATION OF TERRIGENOUS RESERVOIRS

[1] Senior Lecturer, Department of Software,
University of Tyumen
to.imuravev@gmail.com

[2] Cand. Sci. (Phys.-Math.), Professor, Department of Software,
University of Tyumen
i.g.zakharova@utmn.ru

Abstract

Identifying the properties of oil and gas reservoirs based on information obtained from well logging is one of the main areas of research in the field of geological and hydrodynamic modeling of the reservoir. The insufficient effectiveness of accurate mathematical models for analyzing well survey data, as well as the large volume and noise of these data, determines the relevance of using machine learning methods to identify reservoir features.

This article investigates the possibility of classification of terrigenous collectors using various methods, including support vector machine, decision tree, gradient boost, random forest, and multilayered neural network. The data set was formed on the basis of well logging curves for 24 wells of one reservoir. For training classification models, pre-normalized data from inductive logging, lateral log, neutron-neutron logging on thermal neutrons, borehole electrical measurements, resistivity logging, spontaneous potential logging, gamma logging, and resistance logging were used with five different gradient sondes. To assess the accuracy of classification models constructed using various methods, in each case, cross-validation was performed, the average value of accuracy and standard deviation were estimated. For the support vector method, the influence of the choice of core function (linear, polynomial, and

Citation: Muravev I. A., Zaharova I. G. 2019. "Studying the capabilities of machine learning methods for the classification of the character of saturation of terrigenous reservoirs". Tyumen State University Herald. Physical and Mathematical Modeling. Oil, Gas, Energy, vol. 5, no 1, pp. 123-137.
DOI: 10.21684/2411-7978-2019-5-1-123-137

sigmoid) was investigated. In the case of a neural network, its architecture varied, including the number of hidden layers and neurons, activation functions on different layers, and the probability of a dropout. The quality of the obtained classification models was also evaluated by the values of the elements of the confusion matrix.

The results of computational experiments have shown the effectiveness of the use of machine learning methods and, in particular, multilayer neural networks to identify with high accuracy (about 90%) of reservoirs with oil.

Keywords

Terrigenous reservoir, well logging, classification, machine learning, neural net, computing experiment.

DOI: 10.21684/2411-7978-2019-5-1-123-137

REFERENCES

1. Agaev Kh. B. 2013. “The use of cluster analysis for the dissection of a geological section according to well logging data”. Karotazhnik, no 5 (227), pp. 3-11. [In Russian]
2. Gafurov D. O. 2006. “Geological interpretation with the use of trained neural networks in the NeuroInformGeo data of GIS data from the Talakan oil and gas condensate field” Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering, vol. 309, no 3, pp. 32-37. [In Russian]
3. Koskov V. N., Koskov B. V. 2007. Geophysical Studies of Wells and Interpretation of GIS Data. Perm: Perm State Technical University. [In Russian]
4. Paklin N. B., Muhamadiev R. S. 2005. “Using learning algorithms for interpreting GIS data”. Burenie and Neft, no 5, pp. 38-40. [In Russian]
5. Chudinova D. Yu., Dulkarnaev M. R., Kotenev Yu. A., Sultanov Sh. Kh. 2017. “Differentiation of wells in areas with residual oil reserves using neural network modeling”. Exposition Oil and Gas, no 4 (57), pp. 46-50. [In Russian]
6. Al-Mudhafar W. J. 2017. “Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms”. Journal of Petroleum Exploration and Production Technology, vol. 7, no 4, pp. 1023-1033. DOI: 10.1007/s13202-017-0360-0
7. Altman N. S. 1992. “An introduction to kernel and nearest-neighbor nonparametric regression”. The American Statistician, vol. 46, no 3, pp. 175-185. DOI: 10.1080/00031305.1992.10475879
8. Breiman L., Friedman J., Stone C. J., Olshen R. A. 2017. Classification and Regression Trees. New York: Routledge. DOI: 10.1201/9781315139470
9. Breiman L. 2001. “Random forests”. Machine Learning, vol. 45, no 1, pp. 5-32. DOI: 10.1023/A:1010933404324
10. Scikit-learn: Machine Learning in Python. “Compare the effect of different scalers on data with outliers”. Accessed 27 December 2018. https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

11. Fawcett T. 2006. “An introduction to ROC analysis”. Pattern Recognition Letters, vol. 27, no 8, pp. 861-874. DOI: 10.1016/j.patrec.2005.10.010
12. Hagan M. T., Demuth H. B., Beale M. H. 1996. Neural Network Design. Boston: PWS Publishing Company,
13. Hastie T., Tibshirani R., Friedman J. “2009. “Boosting and additive trees”. In: The Elements of Statistical Learning, ch. 10, pp. 337-384. 2nd edition. New York: Springer. DOI: 10.1007/978-0-387-84858-7_10
14. Iglewicz B., Hoaglin D. C. 1993. How to Detect and Handle Outliers. American Society for Quality Control (ASQC), Statistics Division.
15. Kohavi R. 1995. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137-1145.
16. Pearson K. 1901. “On lines and planes of closest fit to systems of points in space”. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, ser. 6, vol. 2, pp. 559-572. DOI: 10.1080/14786440109462720
17. Scikit-learn: Machine Learning in Python. “Quantile transformer”. Accessed 27 December 2018. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html
18. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. 2014. “Dropout: a simple way to prevent neural networks from overfitting”. Journal of Machine Learning Research, vol. 15, pp. 1929-1958.
19. Suykens J. A. K., Vandewalle J. 1999. “Least squares support vector machine classifiers”. Neural Processing Letters, vol. 9, no 3, pp. 293-300. DOI: 10.1023/A:1018628609742
20. Tan F., Luo G., Wang D., Chen Y. 2017. “Evaluation of complex petroleum reservoirs based on data mining methods”. Computational Geosciences, vol. 21, no 1, pp. 151-165. DOI: 10.1007/s10596-016-9601-4
21. Tukey J. W. 1977. Exploratory Data Analysis. Pearson.