САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Математико-Механический факультет
Кафедра информационно-аналитических систем

Применение методов машинного обучения для автоматической разметки результатов геофизического исследования скважин

Курсовая работа студента 546 группы
Чурикова Никиты Сергеевича

Научный руководитель:
Доцент Графеева Н. Г.

Заведующая кафедрой:
Доцент Михайлова Е. Г.

Санкт-Петербург
2017 г.

Содержание

1 Введение

Машинное обучение проникает во многие сферы нашей жизни [11] автоматизируя различные рутинные процессы, вроде поездок на машине и обработки рутинных документов. Поэтому у профессионалов из различных областей естественно возникает желание сократить время работы на не столь увлекательных задачах.

В данном тексте пойдет речь о применении машинного обучения в области геофизики. У специалистов в этой области есть очень трудоемкая задача по выделению на различной глубине в почве пород, основываясь на так называемых методах каротжа. Будет показано, что представляют из себя данные скважин, которые геофизики анализируют, какие наработки, продукты и технологии в данной области уже есть, а также будут приведены наработки и идеи автора по данной задаче.

2 Обзор литературы

Идея применения методов машинного обучения к задаче выделения пород в скважине не новая. Существует достаточно много литературы и статей на эту тему.

Начать разбираться в области применения машинного обучения к классификации литологии стоит с соревнования по данному вопросу [5], которое проводилось сообществом SEG [9]. В этом контесте приводят отличный пример того, как начинать с работать данными по скважинам.

Также благодаря этому конкурсу, существует открытый датасет с разметкой пород. Они также объясняют и показывают, что породы бывают трудноразличимыми и потому ошибка в одну похожую породу допустима.

Также по результатам этого соревнования были написаны интересные статьи, которые кратко описывают научные результаты контеста. Статья [1] подводит итоги и рассказывает о том, как генерировать новые атрибуты. Помимо стандартных подходов, вроде попарных перемножений фичей, они также предлагают считать градиент от атрибутов, воспринимая фичу, как функцию от глубины. Данный подход оказался достаточно удачным и был использован во всех лучших решениях. Помимо этого, работа показывает, что, что лучшим алгоритмом соревнования были деревья основанные на градиентном бустинге [3].

В статье [4] приведена попытка применить популярный алгоритм convolutional neural network (CNN) [6] к данным соревнивания. Но, несмотря на то, что они популярны, и то что атрибуты являются вещественными значениями, на этих данных алгоритм не попал даже в десятку лучших решений. Авторы статьи утверждают, что проблема заключается в недостаточном количестве данных.

Неплохой литературой для начала погружения в геофизику и машинное обучение является книга Мухамедиева Р.И. [8]. В этой работе приведено хороше описание методов каротажа, базовых алгоритмов машинного обучения, а также приводятся рекомендации по подготовке таких специфичных данных. В частности, они не рекомендуют использовать вейвлет преобразования [7], а советуют обратить внимание на следующие этапы предобработки данных:

1. Удаление аномальных значений;
2. Линейная нормировка;
3. Очистка данных по методу «ближайших соседей»;
4. Формирование плавающего окна данных.

3 Постановка задачи

Дана информация об обработке одной или нескольких скважин в некотором месторождении и известно, что на глубине d_i^j встретилась порода y_i^j, где j – номер скважины. Также для каждой скважины j и для каждой глубины i известны значения применявшихся методов исследования скважин x_i^j – **методов каротажа**.

По данным X необходимо сделать прогноз в новой скважине о том, какие породы в ней встретились для каждой глубины i.

Получается, что данную задачу можно интерпретировать, как задачу *классификации*, где X – наблюдаемые значения, а y – целевая переменная.

4 Решение

В данной статье были использованы размеченные данные, полученные во время геофизических работ по исследованию угольных месторождений в республике Коми. Всего в датасете 216 скважин из трех месторождений. Для исследования, были использованы только первое месторождение, в котором 24 скважины.

4.1 Разбиение на обучающую, валидационную и тестовую выборки

Поскольку специфика применения исследования в том, чтобы применить исторические данные к новым, то было принято решение разбить данные из следующим образом:

(train set) В тренировочной выборке, все скважины из первого месторождения исключая одну

(dev set) В валидационной выборке, одна случайная скважина из первого месторождения

4.2 Анализ целевой переменной

На Рис. 1 представлена целевая переменная в первом месторождении. Как видно на изображении, классы несбалансированы, что вносит определенные трудности в работу с ними. Для решения этой проблемы существует различное множество методов, однако мы ограничимся тем, что внутри классификатора будем понижать важность тех классов, которые представлены в большом количестве.

Существует также еще одна потенциальная проблема с размеченными данными. Проблема состоит в том, что их размечает некоторый эксперт. Человек. Т.е. он вносит определенный шум, который определяется его усталостью, неопытностью, ленью - то что называется *Человеческий фактор*. К сожалению, мы пока никак не можем повлиять на это, поскольку в нашем распоряжении имеется лишь эта информация.

Хотя в будущем, при наличии денег и ресурсов, было бы великолепно попросить n геофизиков разметить эти данные и оценить разброс их ответов.

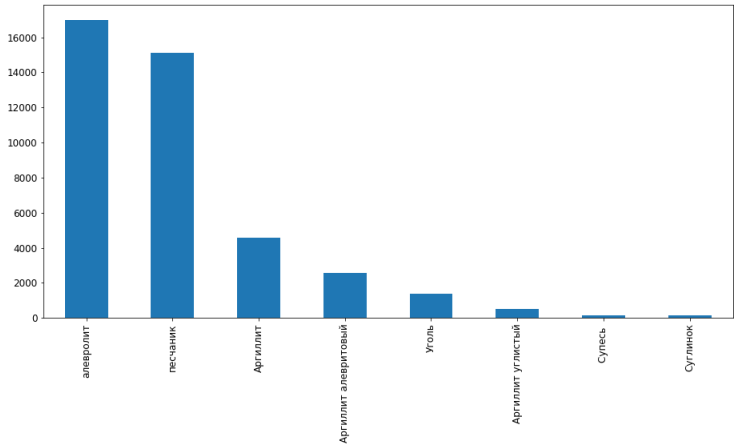

Рис. 1: Породы в первом месторождении

4.3 Анализ наблюдаемых значений

Наблюдаемыми значениями в данной задаче являются значения методов каротажа, которые использовались в скважине. Их существует огромное количество [12], и при различных ситуациях могут использоваться различные их комбинации. В нашем датасете используется от трех до пяти методов. В приложении 1 приведено описание этих методов.

Логично предположить, что чем больше мы проведем геофизических методов исследований, тем более точные данные получатся. Именно поэтому и используют сейчас больше одного метода. Но подобная практика создает сложности для обучения алгоритмов машинного обучения, поскольку текущие алгоритмы не способны работать с пропущенными атрибутами. Эта проблема является отдельной темой исследования. На данном этапе мы ограничимся тем, что будем использовать только те методы, которые есть в обоих месторождениях.

Еще одна проблема с данными, это то что они поступают с датчика, и потому являются шумными данными. В данной работе с этим не было никаких попыток борьбы.

4.4 Классификация

Для классификации было принято решение использовать популярный алгоритм Random Forest [2] и Логистическую регрессию [10]. По результатам классификации получилась следующая матрица неточности, на *dev set*.

По результатам классификации, которые можно увидеть на Рис. 2, 3, получается, что довольно неплохо выделяются уголь, алевролит и песчаник.

С алевролитом и песчаником довольно понятная ситуация. Они представлены в датасете в большом количестве и потому их важность завы-

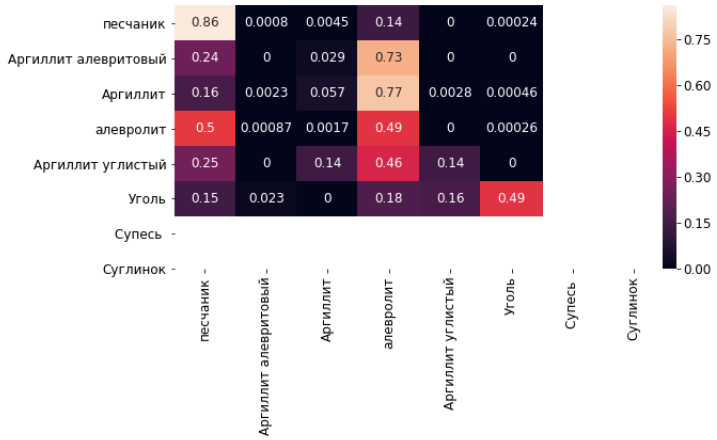

Рис. 2: Random Forest с n_estimators=10, max_depth=8, class_weight='balanced'

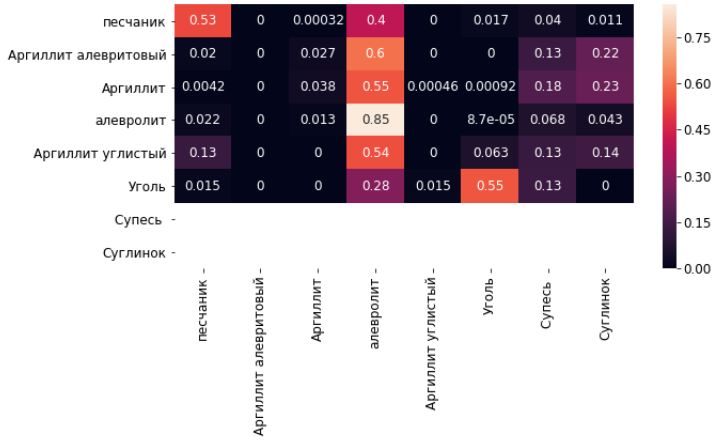

Рис. 3: Logistic Regression с class_weight='balanced'

шается.

С углем можно предположить следующие причины того хорошего предсказания:

- Он является самой интересной породой для исследователей и потому ему уделяется особое внимание;
- Также под него могут подбирать методы каротажа, которые его хорошо выделяют.

5 Выводы

Дипломная работа должна будет содержать следующее:

± Исследование данных угольных скважин

    - Исследована на данный момент только одна скважина

\+ Проверка концепции, что можно прогнозировать породу, используя исторические данные о месторождении.

± Применение различные методы нормализации данных;

    - При обучении Logistic Regression было использовано вычитание среднего и деление на дисперсию

\- Применение других методов классификации

\- Рассмотрена идея того, что методы каротажа могу быть нелинейно зависимы друг от друга. Возможно, эти зависимости можно будет выделить с помощью нелинейных алгоритмов регрессии или нелинейных функций преобразований;

\- Генерация новых признаков посредством их различного комбинирования (перемножение, подсчет градиента, и т.п.)

\- Использована и рассмотрена подробно остальная часть датасета;

6 Заключение

Данная работа является очень актуальной, поскольку данные методы могут помочь геофизикам точнее размечать породы в земле. Были представлены базовые идеи, от которых можно отталкиваться. Несмотря на не очень впечатляющие результаты классификации, алгоритмы машинного обучения способы выделять породы без априорного знания о данных, если есть замеченные данные.

Список литературы

[1] Bestagini, P. A Machine Learning Approach to Facies Classification Using Well Logs / Paolo Bestagini, Vincenzo Lipari, Stefano Tubaro // SEG Technical Program Expandend Abstracts 2017. — 2017. — P. 2137–2142. — http://library.seg.org/doi/abs/10.1190/segam2017-17729805.1.

[2] Breiman, L. Random forests / Leo Breiman // Mach. Learn. — 2001. — Oct. — Vol. 45, no. 1. — P. 5–32. — https://doi.org/10.1023/A:1010933404324.

[3] Chen, T. Xgboost: A scalable tree boosting system / Tianqi Chen, Carlos Guestrin // Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '16. —

New York, NY, USA: ACM, 2016. — P. 785–794. — http://doi.acm.org/10.1145/2939672.2939785.

[4] Facies classification from well logs using an inception convolutional network / Valentin Tschannen, Matthias Delescluse, Mathieu Rodriguez, Janis Keuper. — 2017. — http://arxiv.org/abs/1706.00613.

[5] Geophysics machine learning contest. — 2017. — oct. — https://github.com/seg/2016-ml-contest.

[6] Krizhevsky, A. Imagenet classification with deep convolutional neural networks / Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton // Advances in Neural Information Processing Systems 25 / Ed. by F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger. — Curran Associates, Inc., 2012. — P. 1097–1105. — http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[7] Mallat, S. A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way / Stphane Mallat. — 3rd edition. — Academic Press, 2008.

[8] Muchamediev, R. I. Machine Learning methods applied to geophysics research / Ravil Ilgizovich Muchamediev. — Riga, 2016. — Vol. 200 p.

[9] Society of explorational geophysicists. — 2017. — oct. — http://seg.org.

[10] Sparse multinomial logistic regression: Fast algorithms and generalization bounds / B. Krishnapuram, L. Carin, M. A. T. Figueiredo, A. Hartemink // IEEE Trans. on Pattern Analysis and Machine Intelligence. — 2005. — June. — Vol. 27, no. 6. — P. 957–968.

[11] The top 10 ai and machine learning use cases everyone should know about. — 2016. — sept. — https://goo.gl/bBbZHh.

[12] Well logging methods. — 2017. — oct. — https://goo.gl/hbi7qa.

7 Приложение 1

- **КС** — кажущееся сопротивление с нефокусированными зондами. Самый распространённый метод данной группы, являющийся скважинным аналогом метода электрического профилирования в электроразведке

- **резистивиметрия (REZ)** С помощью этого метода измеряют удельное электрическое сопротивление жидкости, заполняющей в данный момент скважину. Жидкость может быть представлена как

буровым раствором (его сопротивление заранее известно), так и пластовыми флюидами (нефть, пресная или минерализованная вода), а также их смесью

- **БК** — боковой каротаж. Отличие от классического КС заключается в фокусировке тока зондом

- **ГК** — гамма-каротаж. Очень простой и распространённый метод, измеряющий только естественное гамма-излучение от пород, окружающих скважину. Существует его чуть более усложнённый вариант — спектрометрический гамма-каротаж (СГК или ГК-С), который позволяет различить попавшие в детектор геофизического зонда гамма-кванты по их энергии. По этому параметру можно точнее судить о характере слагающих толщу пород.

  - Основная расчетная величина – мощность экспозиционной дозы в микрорентгенах в час (МЭД, мкР/ч). Измеряемая величина определяется концентрацией, составом и пространственным распределением ЕРЭ, плотностью ρ и эффективным атомным номером Zэфф пород.
  - Входит в число обязательных методов

- **ГГК** — гамма-гамма каротаж. Геофизический зонд облучает породу гамма-излучением, в результате которого порода становится радиоактивной и в ответ тоже излучает гамма-кванты. Именно эти кванты и регистрируются зондом.