Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Бочкарев Артем Максимович

Порождение экспертно-интерпретируемых моделей петрофизических измерений в лабораторных исследованиях керна

010990 — Интеллектуальный анализ данных

Выпускная квалификационная работа бакалавра

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва
2016

Содержание

Аннотация

Работа посвящена предсказанию проницаемости горной породы. Работа посвящена нахождению характеристик горной породы, которые можно предсказать через другие измеренные ранее. Используются данные таких замеров как пористость, плотность и температурные характеристики образца. Наиболее важным для приложений является нахождение проницаемости, поэтому основной упор сделан на предсказание этого параметра. Предлагается способ восстановления проницаемости при помощи порождения суперпозиций элементарных функций от других параметров. Поставлен эксперимент на реальных данных замеров керна. Выполнено сравнение с другими популярными методами восстановления проницаемости. Предложен алгоритм, упрощающий структуру нейронной сети благодаря предварительному использованию символьной регресии.

Ключевые слова: *проницаемость, порождение суперпозиций, анализ керна, символьная регрессия.*

Введение

Актуальность темы Работа разделена на две части. В первой части работы предлагается метод комбинации символьной регресии и нейронной сети для упрощения структуры последней, вычислительный эксперимент поставлен на двух различных выборках. Во второй части рассматривается непосредственно задача предсказания проницаемости горной породы, а также различные подходы, которые применяются для ее решения.

Одной из ключевых характеристик для исследования горной породы является проницаемость (permeability). Прямое измерение многих геофизических параметров производится в специально оборудованных лабораториях и может занимать несколько дней, поэтому особый интерес представляет восстановление важных характеристик по тем, которые можно измерить при помощи каротажа (использование магнитных, либо звуковых датчиков, опускаемых на глубину через скважину), либо получаемым в лабораториях.

Моделью для вычисления проницаемости может служить линейная регрессия [1, 2]. Показана особая важность предобработки данных и отбора признаков для решения задачи [3–5]. Для восстановления проницаемости было предложено использовать байесовский вывод [6], проведено сравнение с Lasso регрессией [7]. Также была продемонстрирована эффективность непараметрической регрессии [8] и алгоритма SVR [9]. В качестве моделей для вычисления проницаемости могут использоваться нейронные сети разной структуры [10–13].

Цель работы Целью работы является предсказание проницаемости горной породы на основе других измеренных параметров керна. Модель должна удовлетворять следующим требованиям:

- предсказания должны быть точны — обеспечивать минимально возможное значение заданной функции потерь;
- прогнозы должны удовлетворять экспертным требованиям — значение функции потерь должно позволять использовать прогноз на практике;
- модель должна быть экспертно интерпретируемой — вид и свойства модели

должны быть понятны эксперту-физику.

Методы исследования Метод символьной регрессии заключается в нахождении одновременно наиболее простой и точной суперпозиции элементарных формул. Суперпозиция порождается при помощи грамматики порождения простых математических функций. Суперпозиция представляется в виде дерева, по которому и восстанавливается искомая формула. Для отбора лучших суперпозиций используется генетический алгоритм. Полученные формулы суперпозиции подаются дополнительно на вход нейронной сети, сравнивается эффективность такой структуры и случая с обычными двумя слоями.

В работе используются три выборки. Первая выборка представляет собой рузультаты замеров керна, проведенных в лабораторных условиях. Всего в данной выборке 90 объектов и 15 признаков (такие как пористость, плотность, температурные показатели и.т.д.) Вторая и третья выборки используются для иллюстрации работы предложенного алгоритма суперпозиции нейронной сети и символьной регресии. Одна выборка является результатами аэродинамических и акустических тестов во время продувки крыла в аэротрубе. Другая выборка также содержит петрофизические данные, однако в отличие от первой, данные являются каротажными замерами.

Основные положения, выносимые на защиту.

1. Построение экспертно-интерпретируемых моделей для лабораторных исследований керна и других физических выборок.

2. Решение прикладной задачи предсказания проницаемости горной породы

3. Решение прикладной задачи предсказания шума при продувке крыла

4. Экспериментальное исследование сложности структуры нейронной сети при использовании результатов символьной регресии

Научная новизна Разработан подход суперпозиции символьной регрессии и нейронной сети. Показана его эффективность на реальных данных, установлено снижение сложности структуры нейронной сети.

Практическая значимость. Предложенные методы позволяют успешно решать задачу предсказания проницаемости горной породы. Разработан програмный модуль, позволяющий проводить анализ любых физических выборок, показана эффективность при предсказании уровня шума при продувке крыла.

1 Суперпозиция символьной регрессии и нейронной сети

Рассматривается работа обычной двухслойной нейронной сети. Предлагается заранее построить символьную регрессию и подобрать аналитические функции, наилучшим образом аппроксимирующие обучающую выборку. Предположение состоит в том, что если использовать полученные функции как дополнительные входы для нейронной сети, можно значительно уменьшить сложность ее структуры, при этом не потеряв в качестве работы алгоритма.

1.1 Постановка задачи сивольной регресии

Будем искать всевозможные суперпозиции над грамматикой G:

$$B(g,g)|U(g)|S,$$

где B – множество бинарных операций $\{+,-,*,/\}$, U – множество унарных операций $\{\ln, x^\alpha, \exp\}$, S – множество исходных переменных. Допустимой назовем суперпозицию удовлетворяющую следующим требованиям:

- элементами могут являться только порождающие функции g и свободные переменные
- количество аргументов элемента суперпозиции равно арности используемой функции
- порядок аргументов соответствует порядку аргументов функции
- область определения следующей функции в суперпозиции включает в себя область значений текущей

Каждой суперпозиции f можно сопоставить дерево суперпозиции Γ_f: каждой элементарной функции сопоставляется внутренняя вершина, а каждой независимой переменной сопоставляется лист в дереве. Глубиной дерева суперпозиции будем считать длину самого длинного пути от корня до листа дерева.

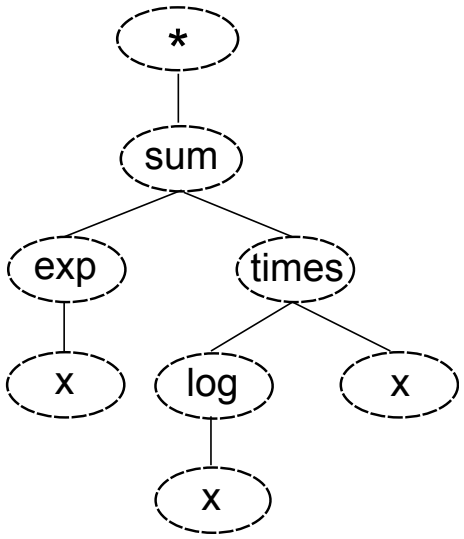

Рис. 1: Пример дерева суперпозиции $f = e^x + x \cdot \ln x$.

Полученную задачу можно определить в качестве задачи оптимизации следующим образом:

$$f^* = arg\min_{f \in F} \sum_{i=1}^{n} (f(x_i) - y_i)^2,$$

где F – пространство всех суперпозиций, n – число объектов в выборке, x_i, y_i – соответственно вектор признакового описания и ответ на i-ом объекте.

1.2 Решение задачи

Заметим, что если хранить дерево суперпозиции в виде матрицы смежности, то поиск графа, оптимизирующего некоторый функционал является NP-сложной задачей. Предлагается приближенное решение, заключающееся в построении искомого дерева в результате поиска генетическим алгоритмом. На этапе инициализации необходимо создать начальную популяцию корректных моделей случайным образом. Опишем действия на каждой итерации алгоритма:

- Выбирается некоторое подмножество моделей, лучших в смысле функционала Q
- Если требуемая точность достигнута, алгоритм прекращает работу
- Происходит скрещивание некоторых из моделей (две модели меняются случайными поддеревьями, корректность должна сохраняться)

- Происходит мутация некоторых моделей (произвольно выбранное поддерево удаляется и заменяется на новое случайное поддерево с фиксированной максимальной сложностью)

- Образуется новое множество моделей, переход к следующей итерации

При использовании генетического алгоритма возможно застревание в локальных минимумах, поэтому во время вычислительного эксперимента будет использоваться мультистарт.

1.3 Общая постановка задачи

Итак, пусть после процедуры генетического алгоритма мы получили функции, наилучшим образом аппроксимирующие выборку. Далее строится нейросеть, имеющая следующую структуру: часть нейронов первого слоя являются стандартными, с нелинейной функцией активации. Нейроны другой части фиксированы и существенно нелинейны, так как являются суперпозициями, полученными при помощи символьной регрессии. Во втором слое находятся обычные нейроны с нейлинейной функцией активации.

Полученная структура может быть описана формулой:

$$\phi(\vec{x}, \vec{w}) = \sigma(\sum_{m=1}^{M} \sigma(\sum_{n=1}^{N_1} \tilde{w_n} f_n(\vec{x}) + \sum_{n=1}^{N_2} w_n x_n + \tilde{w_0}) + w_0),$$

где M – число нейронов скрытого слоя, N_1 – число построенных в результате символьной регресии функций, N_2 – число признаков объектов. Нейронная сеть настраивается методом обратного распространения ошибки на тех же данных, на которых обучалась модель символьной регрессии.

1.4 Вычислительный эксперимент

Вычислительный эксперимент, как и решение задачи, проволился в два этапа. На первом этапе несколько раз запускался генетический алгоритм, в результате чего находились лучшие функции, удовлетворяющие требованию на некоторую максимальную глубину дерева суперпозиции. После этого полученные функции, вместе

со всей обучающей выборкой подавались на вход двухслойной нейронной сети. Алгоритм сравнивался по качеству своей работы с обычной двухслойной нейронной сетью.

1. **Обычная нейронная сеть:** Варьировалось число нейронов скрытого слоя в диапазоне от 1 до 30. Проверка качества проводилась на кросс-валидации по 5 фолдам.

2. **Символьная регрессия + нейронная сеть:** Прежде чем обучать нейронную сеть, 5 раз запускался генетический алгоритм. Полученные топ-5 функций подавались на вход нейронной сети наряду с остальными признаками.

Вычислительный эксперимент был поставлен на двух выборках. Первой выборкой являлась "Airfoil Self Noise"из репозитоиия UCI [14]. Данные представляют собой результаты аэродинамических и акустических тестов во время продувки крыла в аэротрубе. Данные были использованы в работах [15,16] Всего в выборке 1503 образца и 5 признаков, необходимо предсказать уровень шума в децибелах. При подготовке эксперимента все признаки были приведены к одному масштабу нормализацией, пропусков в данных не было. На рисунках 5 и 6 изображен график среднеквадратичной ошибки в зависимости от числа нейронов скрытого слоя. Красным пунктиром показано стандартное отклонение от среднего. При этом качество на кросс-валидации не уступает качеству, полученному ранее в других работах.

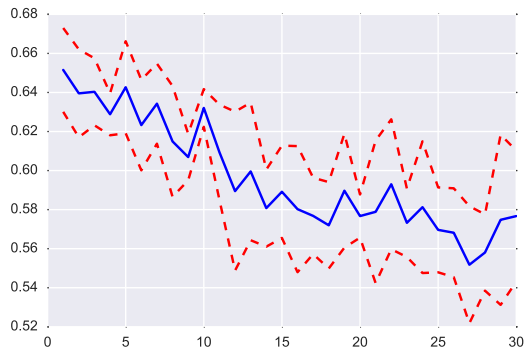

Рис. 2: Двухслойная нейронная сеть

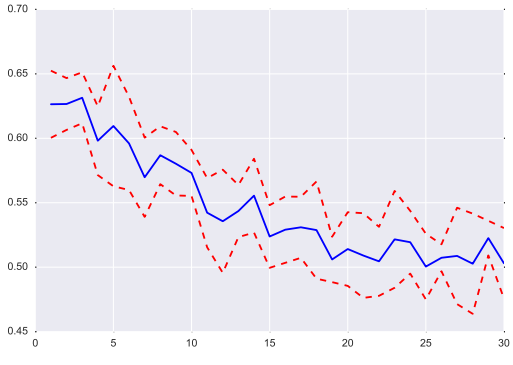

Рис. 3: Комбинация нейронной сети и символьной регресии

На следующем рисунке изображены графики коэффициента детерминации от

числа нейронов для обоих случаев. Видно, что при одинаковом качестве, структура сети существенно упрощается при использовании результатов символьной регресии.

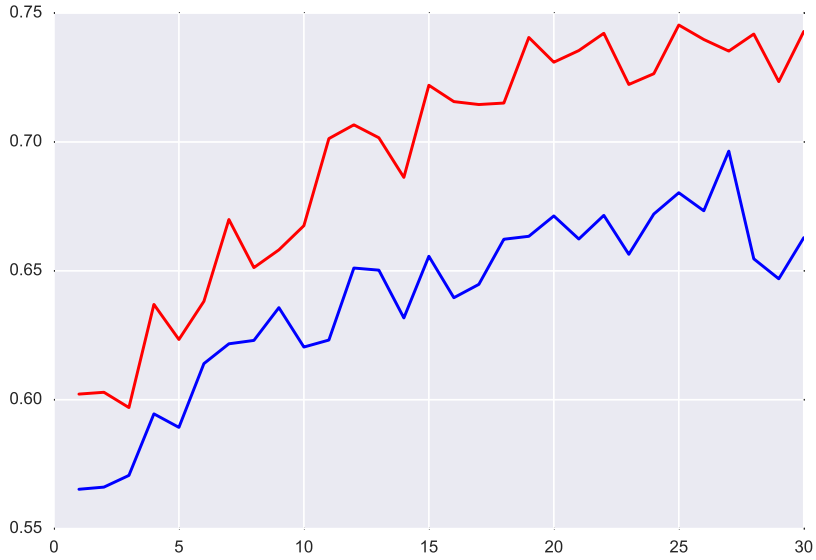

Рис. 4: Коэффициент детерминации

Второй выборкой являлись каротажные данные измерений. Каротаж это измерение параметров горной породы с помощью магнитных и акустических датчиков, опускаемых непосредственно внутрь скважины. Данный вид измерений намного дешевле лабораторных исследований, требующих извлечения образцов на поверхность, но и точность получается меньше. Выборка состояла из 300 объектов и 4 признаков. Глубина, на которой проводились измерения, варьируется от 700 метров до 2 километров. Также как и в предыдущем случае все признаки были приведены к одному масштабу нормализацией. На рисунках 5 и 6 изображен график среднеквадратичной ошибки в зависимости от числа нейронов скрытого слоя. Красным пунктиром показано стандартное отклонение от среднего.

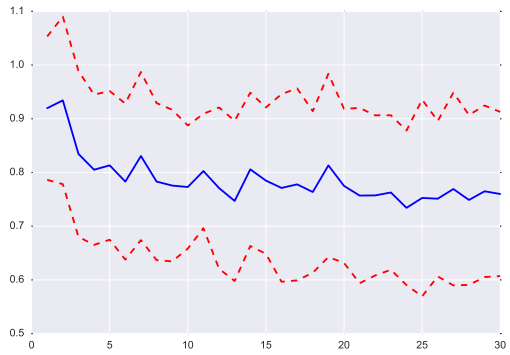

Рис. 5: Двухслойная нейронная сеть

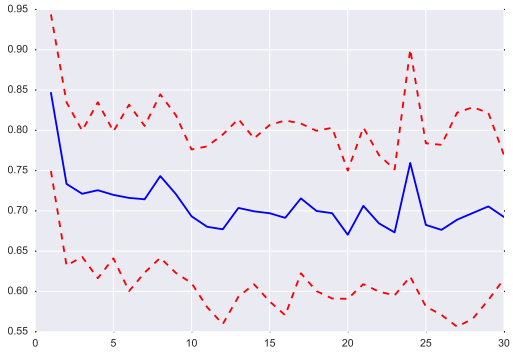

Рис. 6: Комбинация нейронной сети и символьной регресии

На следующем рисунке изображены графики коэффициента детерминации от числа нейронов для обоих случаев. Видно, что качество суперпозиции нейронной сети и генетического алгоритма во всех случаях выше, чем просто при использовании сети..

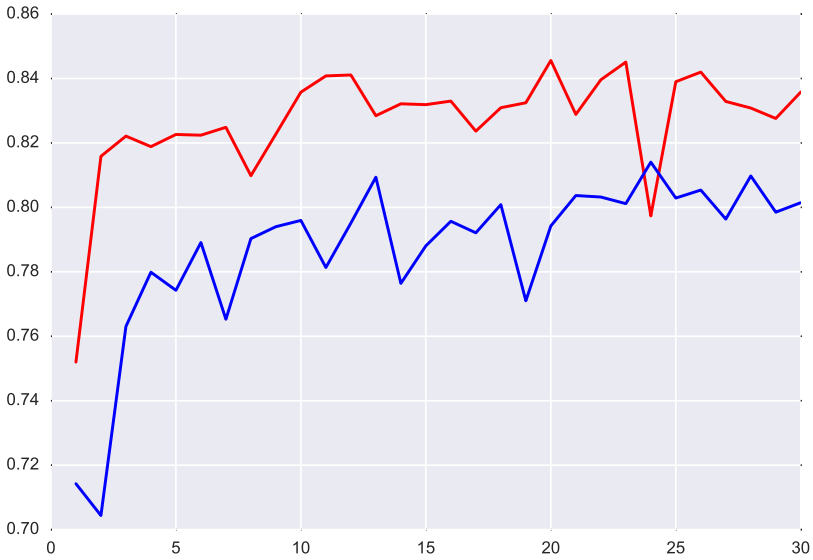

Рис. 7: Коэффициент детерминации

2 Анализ данных исследования керна

В этой части мы рассмотрим выборку, состоящую из образцов керна. Керн – часть горной породы, извлеченная из скважины. В современных лабораториях над образцами керна можно произвести намного более точные измерения, нежели при каротаже. Недостатком является то, что зачастую такие исследования требуют дорогостоящих приборов и много времени. В данной части работы мы попытаемся проанализировать выборку на наличие закономерностей. Особый интерес с точки зрения прикладных исследований представляет собой вычисление проницаемости образца, что и будет нашей основной задачей. В выборке сожержится 235 образцов и 18 признаков, для многих из которых, в том числе и для проницаемости, значения не проставлены.

2.1 Нахождение общих закономерностей

Попытаемся установить, есть ли вообще в нашей признаки, которые хорошо восстанавливаются по остальным. Под словом "хорошо"мы здесь будем подразумевать результат, который устраивал бы по точности физиков, работающих с этими образцами. Расмотрим температурные показатели породы, к ним относятся теплопроводность и температуропроводность по разным осям. Попробуем произвести регрессию каждого такого признака по всем остальным, при этом в качестве дополнительных признаков возьмем элементарные функции от исходных признаков, такие как $\sqrt{x}, \log x, x^2$ и.т.д. Проведем Ridge регрессию, при которой будем варьировать коэффициент регуляризации. Лучшие результаты представлены в табл. 1. Все эксперименты проводились с кросс-валидацией по 5 фолдам.

Таблица 1: Результат регрессии для каждого признака

Признак	α	ϵ
TC horizontal dry	$7.04 \cdot 10^{-4}$	0.053
TD horizontal dry	$7.06 \cdot 10^{-5}$	1.32
TC horizontal water	$5.17 \cdot 10^{-5}$	0.688
TD horizontal water	$3.34 \cdot 10^{-3}$	0.416
Density dry	$7.13 \cdot 10^{-4}$	0.410
Density water	$1.08 \cdot 10^{-4}$	2.072

Ошибка ϵ в данном случае является является относительной, таким образом по крайней мере один признак (TC horizontal dry) можно предсказать по всем остальным с точностью, сравнимой с погрешностью измерения. Ниже представлен путь изменения весов признаков при изменении коэффициента регуляризации для данного случая. Жирной линией выделена получающаяся при этом относительная ошибка.

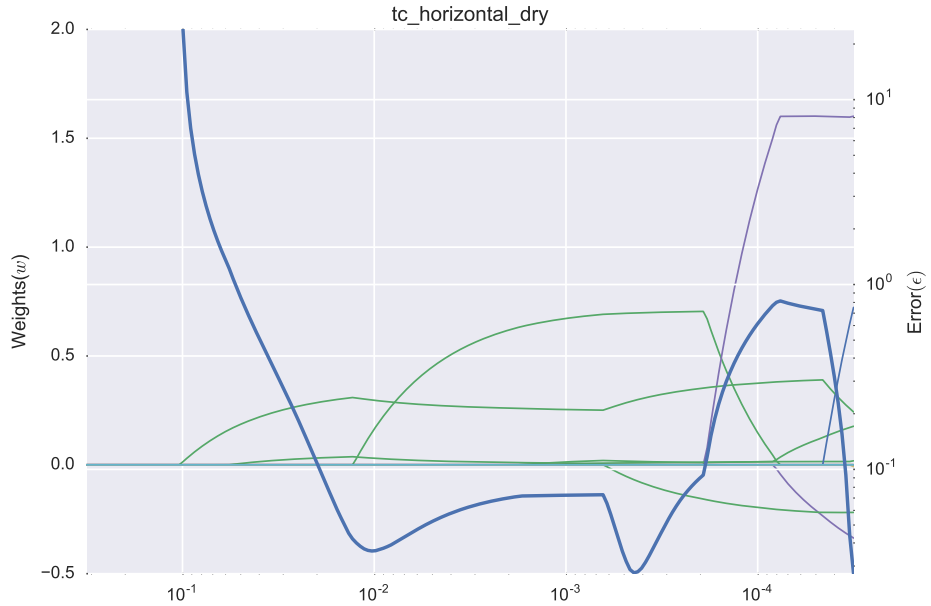

Рис. 8: Путь весов признаков в зависимости от α

2.2 Предсказание проницаемости

Перейдем теперь к наиболее важной с практической точки зрения задачи – предсказанию проницаемости образца. Именно этот параметр помогает определить насколько может быть перспективно добыча ископаемых в данном месте, а также влияет на определение типа попроды. Сформулировать задачу можно следующим образом. Пусть дана выборка $\mathbb{D} = \{(\vec{x}_i, y_i)|i \in \{1, \ldots, N\}\}$, где $\vec{x}_i$ – признаковое описание i-го объекта, а y_i – ответ (значение проницаемости) на этом объекте. Необходимо найти такую модель $\phi \in \Phi$, которая бы доставляла минимум ошибки некоторого функционала качества Q:

$$\phi^* = arg\min_{\phi} Q(\phi, \mathbb{D}).$$

В качестве функционала может выступать среднеквадратичная ошибка, либо коэффицент корреляции.

В нашей выборке значение проницаемости проставлено только на 90 образцах, поэтому будем работать с ними. Попробуем сначала применить несколько основных стандартных методов машинного обучения к решению этой задачи. Прежде всего производится предобработка данных, все признаки приводятся к одному масштабу нормализацией, также предсказывается логарифм проницаемости. После этих операций среднее значение целевой переменной в выборке равно -0.28, а стандартное отклонение равно 0.78.

Двухслойная нейронная сеть. Данный подход широко используется в задачах предсказания проницаемости. Мы будем варьировать число нейронов скрытого слоя в диапазоне от 1 до 20, при этом обучение будет проходить в течение 50 эпох. Здесь и далее все эксперименты поставлены используя кросс-валидацию по 10 фолдам, при этом синей кривой на графиках изображена средняя ошибка, красным пунктиром – стандартное отклонение ошибки.

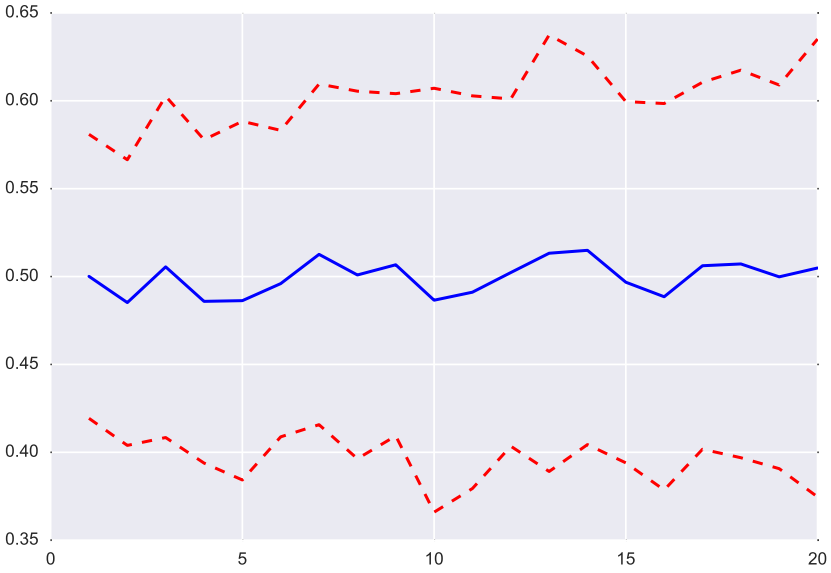

Рис. 9: Двухслойная нейронная сеть

Линейная регрессия. Линейная регрессия также часто используется при предсказании проницаемости. Мы будем использовать Ridge-регрессию, варьируя коэффициент проницаемости от 0 до 10.

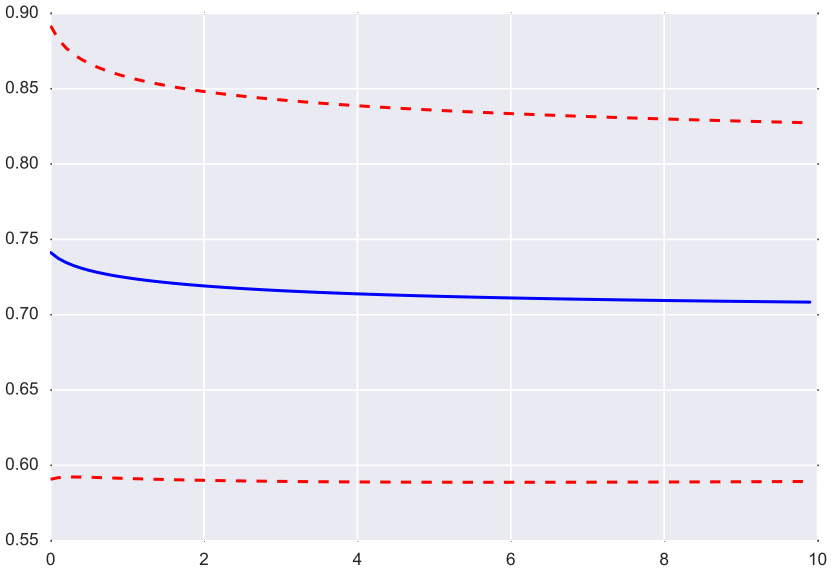

Рис. 10: Линейная регрессия

Случайный лес. В данном случае мы будем варьировать число деревьев в лесу в диапазоне от 10 до 300.

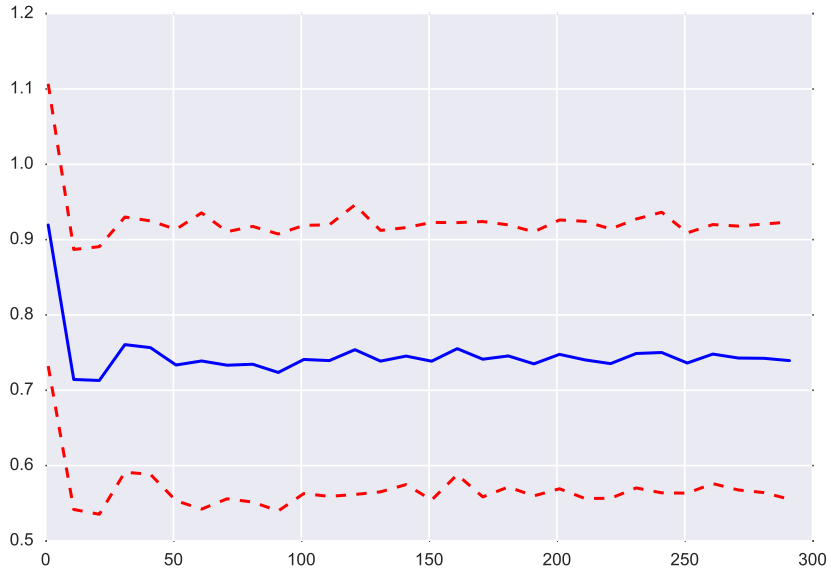

Рис. 11: Случайный лес

Градиентный бустинг Будем варьировать число деревьев от 50 до 1000.

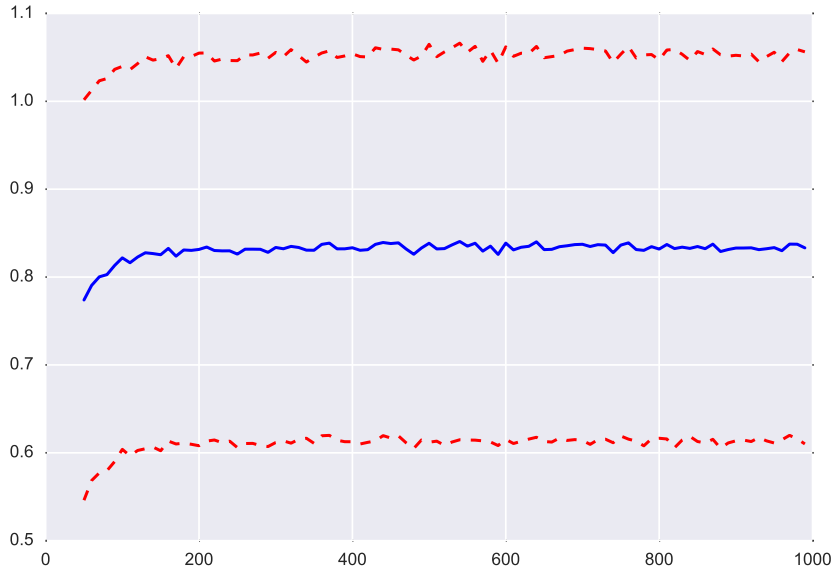

Рис. 12: Градиентный бустинг

K ближайщих соседей. Число ближайших соседей берется в диапазоне от 1 до 20.

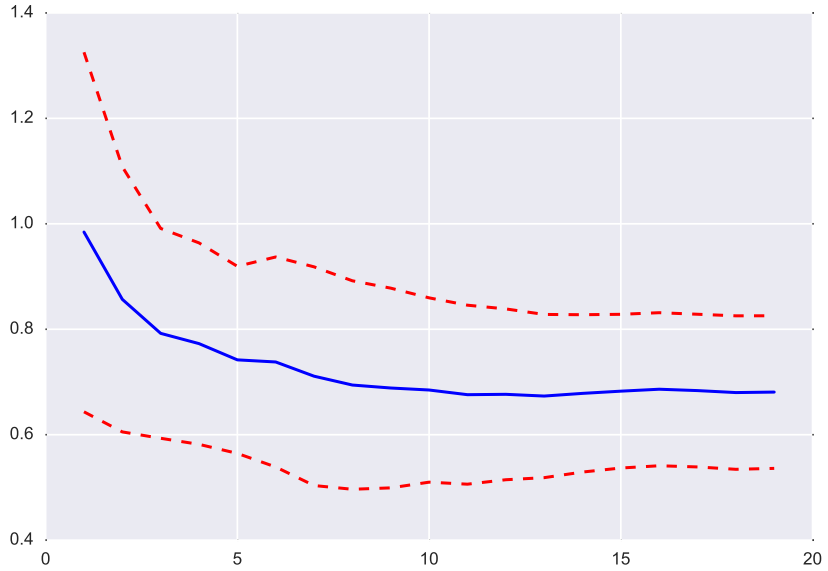

Рис. 13: K ближайщих соседей

По этим графикам видно, что средняя величина ошибки большая, дисперсия ошибки также является большой. Все графики очень быстро выходят на почти полное насыщение, поэтому возникает предположение, что проницаемость в принципе плохо восстанавливается по имеющимся в данной выборке признакам. Общие результаты представлены в таблице.

Таблица 2: My caption

Алгоритм	Средняя ошибка на кросс-валидации	Стандартное отклонение
Двухслойная нейронная сеть	0.47	0.08
Линейная регрессия	0.71	0.11
Случайный лес	0.73	0.18
Градиентный бустинг	0.77	0.22
KNN	0.67	0.15

Возможно, значительная погрешность вносится из-за того, что образцы в нашей

выборки берутся с разной глубины, то есть могут принадлежать разным попродам, а значит не могут описываться одной моделью достаточно хорошо. Попробуем прежде чем предсказывать величину проницаемости провести кластеризацию имеющейся выборки. Если какой-то из кластеров будут хорошо выделяться при варьировании числа кластеров, то можно предположить, что объекты в нем принадлежат одной породе и построенная модель будет лучше. Кластеризацию проведем алгоритмом K-means, варьируя число кластеров от 2 до 5. Визуализируем результат следующим образом: по горизонтали отложим номер образца, образцы одного кластера будем обозначать одним цветом, а по вертикали отложим случаи различного числа кластеров.

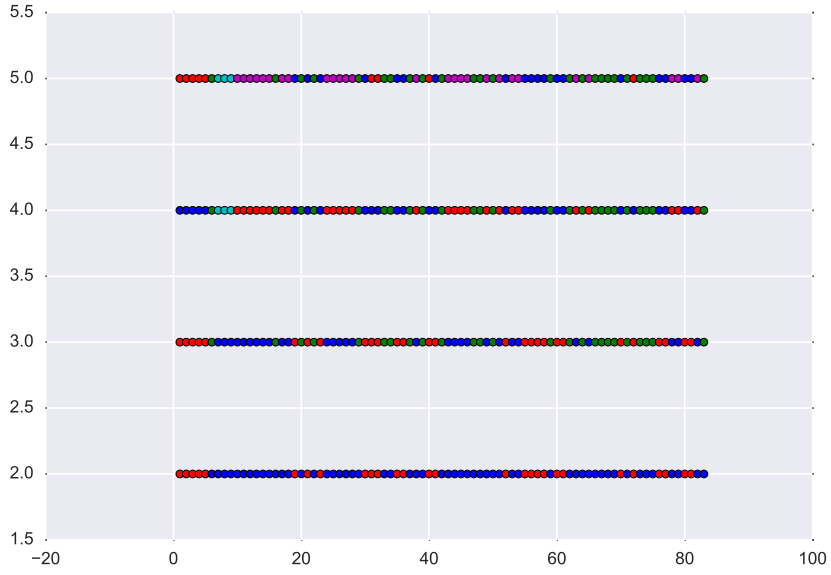

Рис. 14: Кластеризация

Выберем вручную наиболее устойчивый кластер, в нем содержится 30 объектов выборки. Будем искать модель для этого кластера с помощью символьной регресии, чтобы в случае хорошего качества она могла бы быть интерпретируема физиками. В таблице ниже показаны топ-5 моделей, найденных после многократных повторных стартов эволюции, с максимальной глубиной дерева равной 4.

Номер функции	Функция	R^2
1	$x_0 \cdot 3^{x_2}$	0.43
2	$x_0 \cdot e^{x_2}$	0.42
3	$\frac{-x_0}{5^{x_3}}$	0.42
4	$x_0 \cdot 3^{x_2}$	0.41
5	$\frac{\frac{x_2}{2}}{-1-x_3}$	0.40

Исходя из качества и после обсуждения с экспертами неустойчивого вида полученных функций, мы приходим к выводу, что имеющихся данных не хватает для того, чтобы определять проницаемость с желаемой точностью.

3 Заключение

Предложен алгоритм построения суперпозиции двухслойной нейронной сети и символьной регрессии. При помощи символьной регресии построены экспертно-интерпретируемые модели, удовлетворяющие требуемой точности. Проведен эксперимент на трех выборках. На одной из двух петрофизических выборок и выборке Airfoil предложенный алгоритм дал достаточную хорошую точность. Третья выборка плохо описывается любыми моделями, вероятно дело в неполноте данных. Показано существенное упрощение структуры двухслойной нейронной сети при добавлении функций, полученных в результате сивольной регрессии, в качестве нейронов входного слоя.

Список литературы

[1] M. N. Mohamad Ibrahim and L. F. Koederitz. Two-phase relative permeability prediction using a linear regression model. *SPE*, 2000.

[2] A.T. Watson, P.C. Richmond, and T.M. Tao P.D. Kerig. A regression-based method for estimating relative permeabilities from displacement experiments. *SPE Reservoir Engineering*, 1988.

[3] Mauro Cozzi, Livio Ruvo, Paolo Scaglioni, and Anna Maria Lyne. Core-data preprocessing to improve permeability-log estimation.

[4] M. Cozzi, L. Ruvo, P. Scaglioni, and A.M. Lyne. Core data preprocessing to improve permeability log estimation. *Society of Petroleum Engineers*, 2006.

[5] Kabiru O. Akande. Comparative analysis of feature selection-based machine learning techniques in reservoir characterization. *Society of Petroleum Engineers*, 2015.

[6] Watheq Al-Mudhafar. Integrating bayesian model averaging for uncertainty reduction in permeability modeling. *Offshore Technology Conference*, 2015.

[7] Watheq J. Al-Mudhafar. Comparison of permeability estimation models through bayesian model averaging and lasso regression. *Society of Petroleum Engineers*, 2015.

[8] Guoping Xue, Akhil Datta Gupta, Peter Valko, and Tom Blasingame. Optimal transformations for multiple regression: Application to permeability estimation from well logs. *SPE Formation Evaluation*, 1997.

[9] A. Al-Anazi and I.D. Gates. Support-vector regression for permeability prediction in a heterogeneous reservoir: A comparative study. *SPE Reservoir Evaluation*, 2010.

[10] E.M.El-M. Shokir, A.A. Alsughayer, and A. Al-Ateeq. Permeability estimation from well log responses. *Journal of Canadian Petroleum Technology.*

[11] B. Guler, T. Ertekin, and A.S. Grader. An artificial neural network based relative permeability predictor. *Journal of Canadian Petroleum Technology*, 42(4), April 2003.

[12] K. Aminian, S. Ameri, A. Oyerokun, and B. Thomas. Prediction of flow units and permeability using artificial neural networks. *SPE 83586*, 2003.

[13] Yulia Maslennikova. Permeability prediction using hybrid neural network modelling. *SPE Annual Technical Conference and Exhibition*, 2013.

[14] Airfoil self noise. https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise.

[15] K. Lau, R. López, and E. Oñate. A neural networks approach to aerofoil noise prediction. *Publication CIMNE No-335*, 2009.

[16] Roberto Lopez Gonzalez. *Neural Networks for Variational Problems in Engineering*. PhD thesis, Department of Computer Languages and Systems Technical University of Catalonia Department of Computer Languages and Systems Technical University of Catalonia, 2008.