Порождение экспертно-интерпретируемых моделей петрофизических измерений в лабораторных исследованиях керна

Бочкарев Артем

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва,
15 июня 2016 г.

Цель работы

Задача

Предсказать проницаемость горной породы на основе других измеренных параметров керна.

Требования к модели

- 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 решения

Проведем символьную регрессию, а затем построим суперпозицию полученных функций и двухслойной нейронной сети.

Образец керна

Важность оценивания проницаемости

Проницаемость — это свойство пористой среды пропускать через себя жидкость или газ при перепаде давления. Классическая модель — зависимость от пористости.

Используемые признаки

Для предсказания проницаемости используются физические свойства горной породы:

- теплопроводность
- температуропроводность
- теплоемкость
- плотность
- минералогическая плотность
- пористость

Для большинства показателей имеются данные для сухой и водонасыщенной породы, а таже измерения по разным осям.

Литература

Исследования петрофизических данных

1. M. N. Mohamad Ibrahim and L. F. Koederitz. *Two-phase relative permeability prediction using a linear regression model.* SPE, 2000
2. A. Al-Anazi and I.D. Gates. *Support-vector regression for permeability prediction in a heterogeneous reservoir: A comparative study.* SPE Reservoir Evaluation, 2010
3. Yulia Maslennikova. *Permeability prediction using hybrid neural network modelling.* SPE Annual Technical Conference and Exhibition, 2013.

Методы порождения моделей

1. Koza John R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* — The MIT Press, 1992.
2. Г. И. Рудой, В. В. Стрижов. *Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных* Информатика и ее применения. — 2013. — Vol. 1
3. Р.А. Сологуб *Методы трансформации моделей в задачах нелинейной регрессии*

Постановка задачи

Пусть дана выборка $D = \{(\mathbf{x}_i, y_i) | i \in \{1, \ldots, N\}\}$, где $\mathbf{x}_i$ – признаковое описание i-го объекта, а y_i – ответ (значение проницаемости) на этом объекте. Необходимо найти такую модель $f \in \mathcal{F}$, которая бы доставляла минимум ошибки Q:

$$f^* = arg \min_f Q(f, D).$$

В качестве функции Q выступает среднеквадратичная ошибка.

Критерии качества модели

- Величина ошибки
- Дисперсия ошибки
- Сложность модели
- Экспертная интерпретируемость

Порождение функций

Будем искать всевозможные суперпозиции над грамматикой G:

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

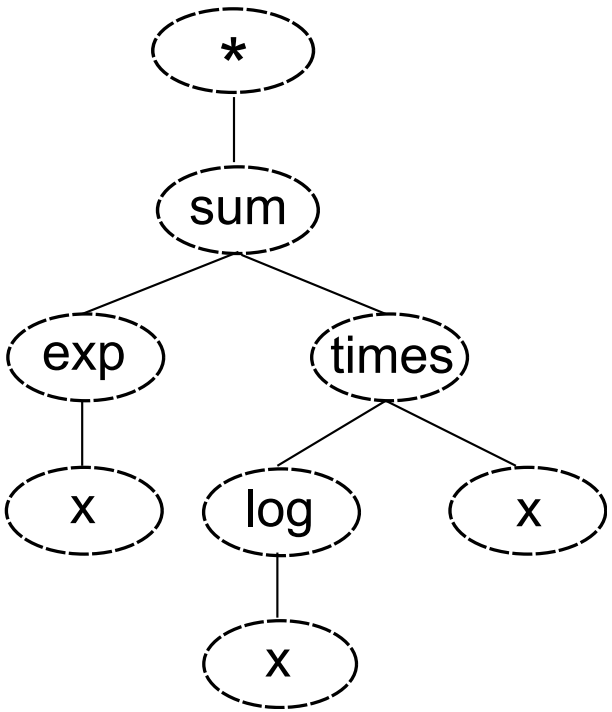
где B – множество бинарных операций $\{+,-,*,/\}$, U – множество унарных операций $\{\ln, x^{\alpha}, \exp\}$, S – множество исходных переменных.

Задача оптимизации

$$f^* = arg \min_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$$

Дерево суперпозиции

Каждой суперпозиции f можно сопоставить дерево суперпозиции Γ_f. Глубиной дерева суперпозиции будем считать длину самого длинного пути от корня до листа дерева.



Дерево Γ_f

1. Корень дерева - $*$;
2. $V_i \mapsto g_r$;
3. $\mathrm{val}(V_j) = v(g_{r(i)})$;
4. $\mathrm{dom}(g_{r(i)}) \supset \mathrm{cod}(g_{r(j)})$;
5. аргументы g_r упорядочены;
6. x_i — листья Γ_f.

$$f = e^x + x \cdot (\log x)$$

Решение

Задача оптимизации является NP-сложной, будем решать ее приближенно при помощи генетического поиска.

Генетический алгоритм

1. Выбирается некоторое подмножество лучших моделей
2. Если требуемая точность достигнута, алгоритм прекращает работу
3. Происходит скрещивание некоторых из моделей
4. Происходит мутация некоторых моделей (произвольно выбранное поддерево удаляется и заменяется на новое случайное поддерево)
5. Образуется новое множество моделей, переход к следующей итерации

Суперпозиция нейронной сети и символьной регресии

Нейронной сети на вход помимо признаков подается топ функций, построенных на предыдущем шаге. Нейронная сеть обучается при помощи метода обратного распространения ошибки, дополнительного прореживания не проводилось.

Структура сети

$$\phi(\vec{x}, \vec{w}) = \sigma(\sum_{m=1}^{M} \sigma(\sum_{n=1}^{N_1} \tilde{w}_n f_n(\vec{x}) + \sum_{n=1}^{N_2} w_n x_n + \tilde{w}_0) + w_0),$$

где M – число нейронов скрытого слоя, N_1 – число построенных в результате символьной регресии функций, N_2 – число признаков объектов.

Предположение

Нейронная сеть, имеющая такую структуру требует меньшего числа нейронов скрытого слоя, чем обычная двухслойная нейронная сеть. При этом не происходит потерь в качестве.

Вычислительный эксперимент

Вычислительный эксперимент был поставлен на трех выборках:

1. каротажные измерения
2. выборка airfoil
3. лабораторные исследования керна

Цели эксперимента

- Построить модель, удовлетворяющую требуемому уровню качества
- Убедиться, что добавление символьной регресии позволяет снизить сложность структуры нейронной сети
- Проверить предположение, что проницаемость может быть описана при помощи хорошей экспертно интерпретируемой модели

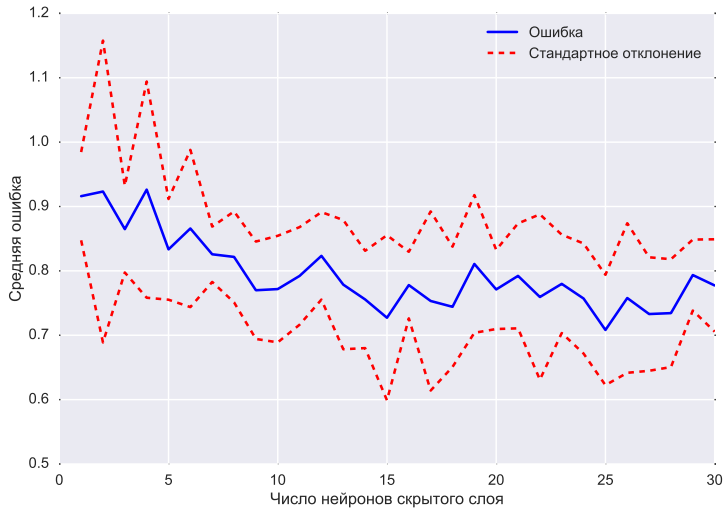
Каротажные измерения керна

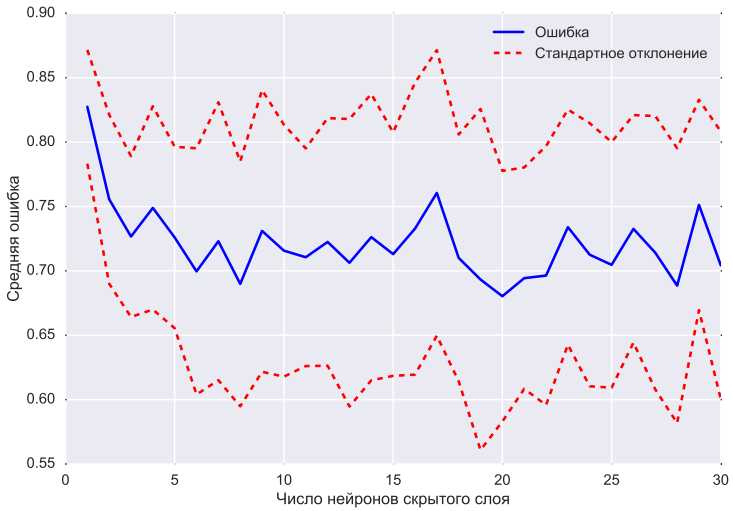
Выборка состояла из 300 объектов и 4 нормализованных признаков. Глубина, на которой проводились измерения, варьируется от 700 метров до 2 километров. Приведен список лучших функций, полученных в результате символьной регресии.

Номер функции	Функция	Ошибка
1	$2 - \sqrt{x_3} - \sqrt{x_0} + x_3$	1
2	$2x_3 + x_0 + \left(\frac{x_2}{x_3}\right)^{\sqrt{x_3}}$	1.07
3	$2x_3 + x_0 + \sqrt[4]{x_1}$	1.08
4	$2x_3 + x_0 + 1$	1.09
5	$2x_3 + x_0 + e^0$	1.09

Каротажные измерения керна

На графиках показана зависимость средней ошибки на кросс-валидации от числа нейронов скрытого слоя.





Выводы

Использование суперпозиции обеспечивает меньшую ошибку на кросс-валидации.

Каротажные измерения керна

На изображена зависимость коэффициента детерминации от числа нейронов в скрытом слое.

Выводы

Использование результатов символьной регресии позволяет упростить структуру нейросети без потерь в качестве.

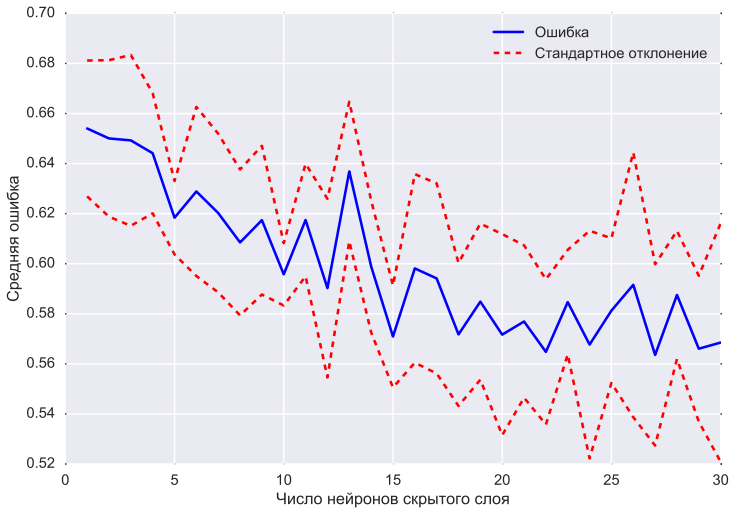
Airfoil(UCI)

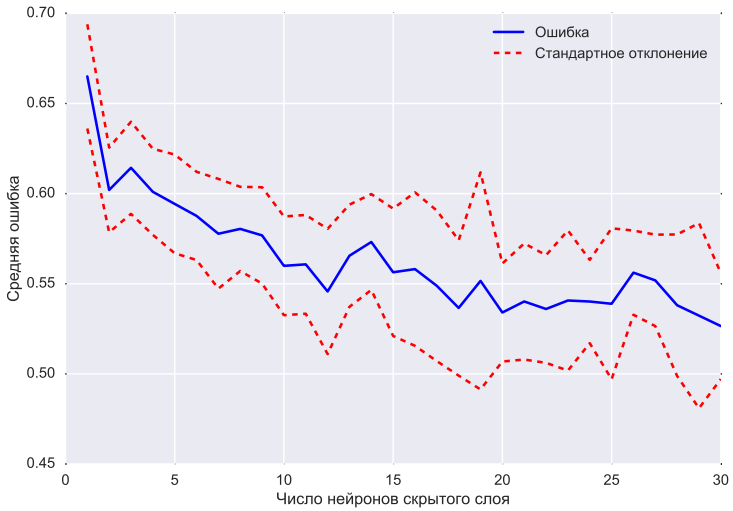
Данные представляют собой результаты аэродинамических и акустических тестов во время продувки крыла в аэротрубе. Всего в выборке 1503 образцов крыла и 5 нормализованных признаков, необходимо предсказать уровень шума в децибелах. Приведен список лучших функций, полученных в результате символьной регресии.

Номер функции	Функция	Ошибка
1	$\frac{\frac{2x_0+x_2+x_4}{x_2-4}}{\sqrt[4]{x_0}}$	0.74
2	$\frac{x_4+\frac{x_0+x_2}{\sqrt{x_2}}}{-3}$	0.75
3	$\frac{\sqrt{x_0}+x_2+x_4}{\log x_0-x_0}$	0.77
4	$\frac{2x_0+x_2+x_4}{-5}$	0.81
5	$\frac{(x_2+5)(x_0+x_4)}{-10}$	0.82

Airfoil

На графиках показана зависимость средней ошибки на кросс-валидации от числа нейронов скрытого слоя.



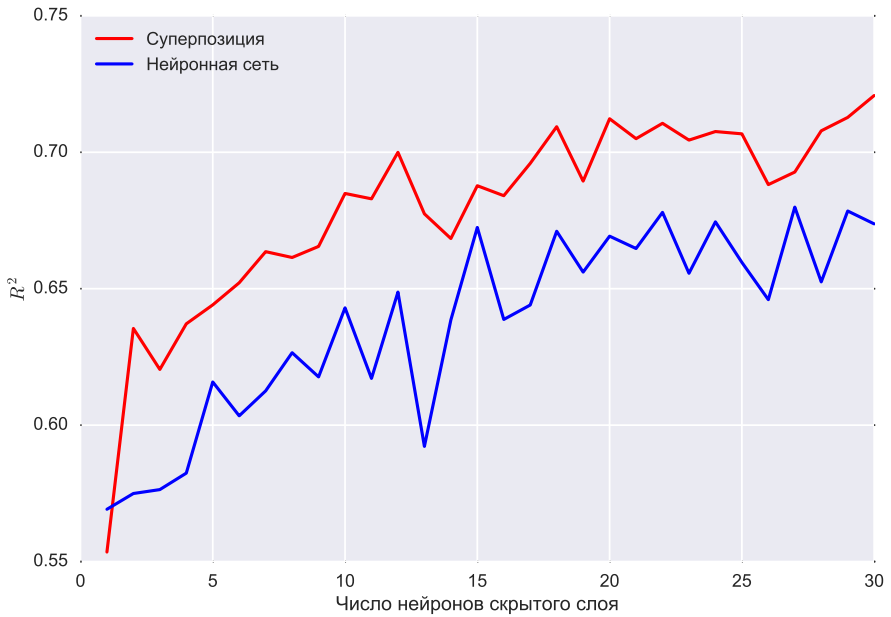


Выводы

Использование суперпозиции обеспечивает меньшую ошибку на кросс-валидации.

Airfoil

На изображена зависимость коэффициента детерминации от числа нейронов в скрытом слое.



Выводы

Использование результатов символьной регресии позволяет упростить структуру нейросети без потерь в качестве. Качество работы нашего алгоритма не уступает качеству работы алгоритмов других авторов, использовавших эту выборку.

Лабораторные исследования керна

В выборке содержится 235 образцов и 18 признаков. В ходе исследований была построена не только модель, описанная ранее, но были опробованы такие подходы как обычная нейронная сеть, случайный лес, линейная регрессия, градиентный бустинг. Качество ни одной из моделей не удовлетворяло экспертным требованиям, было решено сделать предварительную кластеризацию выборки, чтобы построить модель проницаемости хотя бы для одного типа породы. В таблице показаны лучшие функции, отобранные при помощи символьной регресии и их качество.

Номер функции	Функция	Ошибка
1	$\sin({x_3}^4) - \sin(x_0 - 1)$	0.36
2	$sin(4 \cdot {x_2}^9)$	0.54
3	$\sin(\sqrt{7}^{(x_0+x_2)})$	0.54
4	$\cos(\log(x_1) \cdot (x_0 - 8))$	0.56
5	$\sin(\exp x_0 \cdot {x_2}^4)$	0.59

Заключение

- Построена экспертно-интерпретируемая модель проницаемости горной породы
- Установлено, что структура нейронной сети упрощается при предварительном проведении символьной регресии
- Построенная модель удовлетворяет требуемым критериям качества на двух выборках из трех
- Одна выборка плохо описывается любыми моделями, скорее всего данных чтобы определить проницаемость недостаточно