

Порождение экспертно-интерпретируемых моделей петрофизических измерений в лабораторных исследованиях керна

Бочкарев Артем Максимович

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.

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.

Важность оценивания проницаемости

Проницаемость — это свойство пористой среды пропускать через себя жидкость или газ при перепаде давления.

Классическая модель — зависимость от пористости.



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 также измерения по разным осям.

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*

Пусть дана выборка $D = \{(\mathbf{x}_i, y_i) | i \in \{1, \dots, N\}\}$, где \mathbf{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 Q :

$$\phi^* = \underset{\phi}{\operatorname{arg\,min}} Q(\Phi, D, \mathbf{w}).$$

В качестве функции Q выступает среднеквадратичная ошибка.

Критерии качества модели

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

Будем искать всевозможные суперпозиции над грамматикой G :

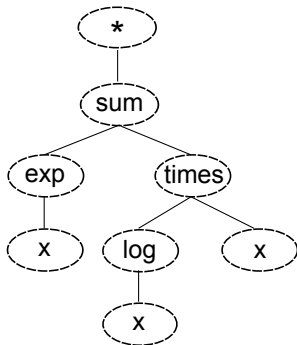
$$B(g, g) | U(g) | S,$$

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

Задача оптимизации

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

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



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

Дерево Γ_f

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

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

Генетический алгоритм

- 1 Инициализация
- 2 Проверка качества решения
- 3 $V_i \leftrightarrow V_j$
- 4 $V_i \rightarrow \hat{V}_i$
- 5 Отбор моделей

Нейронной сети на вход помимо признаков подается топ функций, построенных на предыдущем шаге.

Структура сети

$$\phi(\vec{x}, \vec{w}) = \sigma\left(\sum_{m=1}^M \sigma\left(\sum_{i=1}^{N_1} w_{0i} f_i(\vec{x}) + \sum_{j=1}^{N_2} w_{1j} x_j + w_{00}\right) + w_{01}\right),$$

где M – число нейронов скрытого слоя, N_1 – число построенных в результате символьной регрессии 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 был поставлен на трех выборках:

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

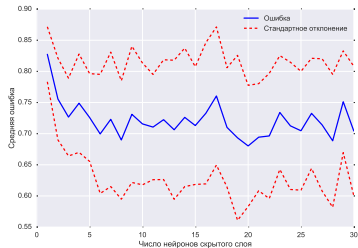
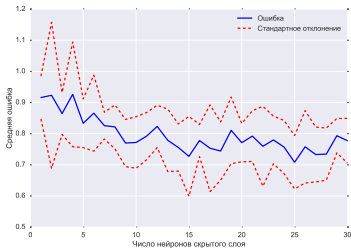
Цели эксперимента

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

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 регрессии.

Номер функции	Функция	R^2
1	$2 - \sqrt{x_3} - \sqrt{x_0} + x_3$	0.7
2	$2x_3 + x_0 + \left(\frac{x_2}{x_3}\right)\sqrt{x_3}$	0.69
3	$2x_3 + x_0 + \sqrt[4]{x_1}$	0.69
4	$2x_3 + x_0 + 1$	0.66
5	$2x_3 + x_0 + e^0$	0.65

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

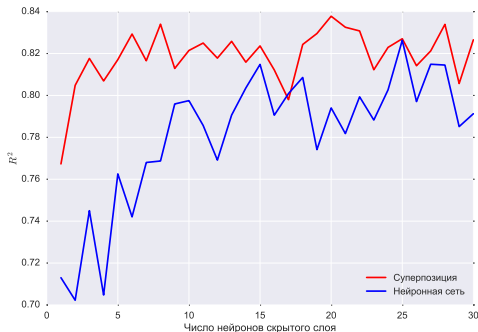


Выводы

Использование суперпозиции обеспечивает меньшую ошибку на кросс-валидации.

Каротажные измерения керна

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



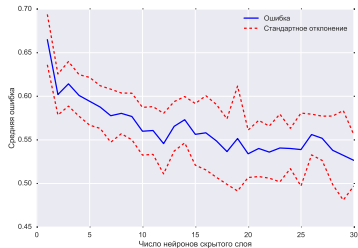
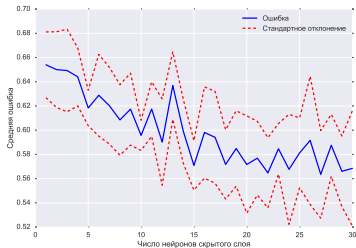
Выводы

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

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 регрессии.

Номер функции	Функция	R^2
1	$\frac{\frac{2x_0+x_2+x_4}{x_2-4}}{\sqrt[4]{x_0}}$	0.58
2	$\frac{x_4 + \frac{x_0+x_2}{\sqrt{x_2}}}{-3}$	0.56
3	$\frac{\sqrt{x_0+x_2+x_4}}{\log x_0 - x_0}$	0.55
4	$\frac{2x_0+x_2+x_4}{-5}$	0.55
5	$\frac{(x_2+5)(x_0+x_4)}{-10}$	0.53

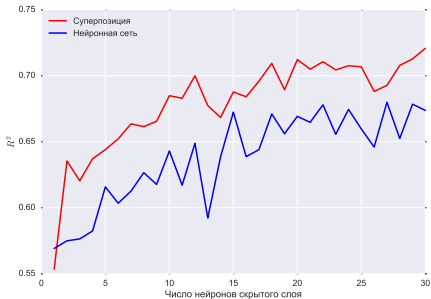
На графиках показана зависимость средней ошибки на кросс-валидации от числа нейронов скрытого слоя.



Выводы

Использование суперпозиции обеспечивает меньшую ошибку на кросс-валидации.

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



Выводы

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

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