Прогноз проницаемости горной породы с помощью символьной регрессии

Бочкарев А. М.
Научный руководитель: д.ф.-м.н. Стрижов В. В.

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

Таганрог,
12 октября 2017 г.

Цель работы

Задача

- Предсказать проницаемость горной породы на основе других измеренных параметров керна.
- Упростить структуру нейронных сетей, используемых для решения задачи регрессии.

Требования к модели

- предсказания должны быть точны — обеспечивать минимально возможное значение заданной функции потерь;
- должно уменьшиться число нейронов сети;
- модель должна быть экспертно интерпретируемой (хотя бы частично).

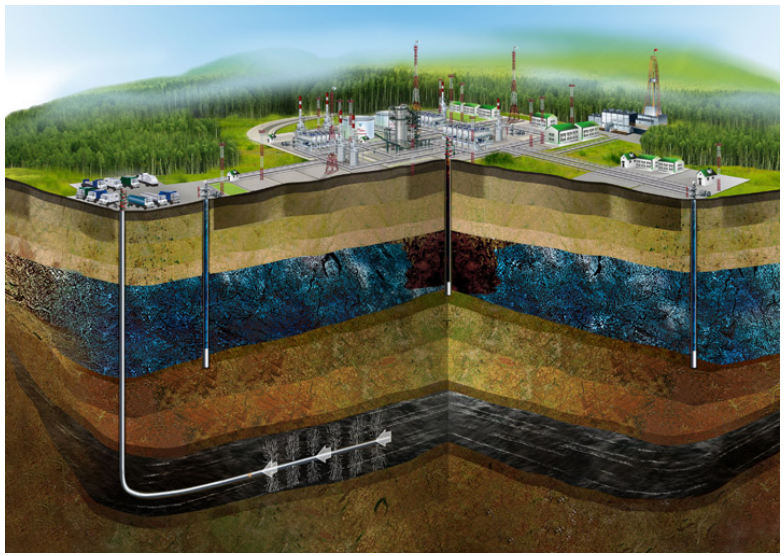
Метод решения

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.



Литература

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. Kulunchakov, A., Strijov V. *Generation of simple structured information retrieval functions by genetic algorithm without stagnation.* — Expert Systems with Applications 85 (2017): 221-230.
4. Р.А. Сологуб *Методы трансформации моделей в задачах нелинейной регрессии* Машинное обучение и анализ данных. 1.14 (2015): 1961-1976.

Постановка задачи

Задана выборка $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \mathbf{x}_i \in \mathbb{R}^n, y_i \in \mathbb{R}$. Необходимо найти функцию $f : \mathbb{R}^n \to \mathbb{R}$ из семейства моделей $\mathcal{F}$, удовлетворяющую минимуму заданной функции потерь Q:

$$f^* = \arg\min_{f \in \mathcal{F}} Q(f, D)$$

$$Q = \sqrt{\sum_{i=1}^{m} (y_i - f(\mathbf{x}_i)^2}$$

Задача символьной регрессии

Ищутся всевозможные допустимые суперпозиции заданных функций над грамматикой G:

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

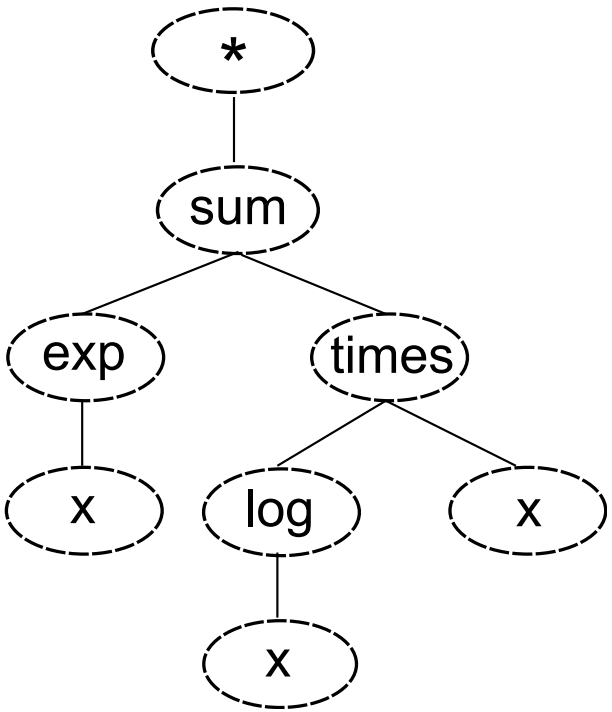
где B – множество бинарных операций $\{+, -, *, /\}$, U – множество унарных операций $\{\ln, x^{\alpha}, \exp\}$, S – множество исходных переменных.

Допустимые суперпозиции

1. элементами могут являться только порождающие функции g и свободные переменные;
2. число аргументов элемента суперпозиции равно числу аргументов используемой функции;
3. порядок аргументов соответствует порядку аргументов функции;
4. область определения следующей функции в суперпозиции включает в себя область значений текущей.

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.

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

Генетический алгоритм построения суперпозиций

1: **while** не достигнута требуемая точность **do**
2: Из популяции $\mathcal{M}$ выбирается некоторое подмножество моделей, лучших в смысле функционала Q.
3: Две случайно выбранные модели меняются случайными поддеревьями, условие допустимости должно удовлетворяться.
4: Произвольно выбранное поддерево удаляется и заменяется на новое случайное поддерево при условии фиксированной максимальной сложности.
5: Порожденные модели добавляются в популяцию $\mathcal{M}$.
6: **end while**

Анализ алгоритма

Вводится метрика на парах суперпозиций:

$$\mu(f, f') = \mu(\Gamma_f, \Gamma_{f'})$$

Свойства метрики

1. $\mu(f, f) = 0, \mu(f, f') > 0$, при $f \neq f'$
2. $\mu(f, f') = \mu(f', f)$
3. $\mu(f, f') \leq \mu(f', f'') + \mu(f, f'')$

Окрестность суперпозиции

$$U_r(f) = \{f' \in \mathcal{F} : \mu(f, f') < r\}$$

Анализ алгоритма

Радиус популяции $\mathcal{M}$

$$r(\mathcal{M}) = \arg\min_{r>0}\{\exists f \in \mathcal{F}\ \forall f' \in \mathcal{M} : f' \in U_r(f)\}$$

Эмпирический радиус популяции

$$r_e(\mathcal{M}) = \frac{\sum_{f,f' \in \mathcal{M}} \mu_i(f, f')}{|\mathcal{M}| \sum_{f \in \mathcal{M}} |f_j|}$$

Способы задания метрики

1. Метрика основанная на изоморфизме графов:
   $\mu(T_i, T_j) = |T_i| + |T_j| - 2|T_{ij}|$
2. Расстояние Левенштейна на графах

Анализ алгоритма

Стагнация алгоритма

Эволюционная стагнация - ситуация при котором все суперпозиции попарно близки. В этом случае усложняется генерация принципиально новых суперпозиций и популяция не изменяется от итерации к итерации.

Индикатором стагнации служит радиус популяции $r(\mathcal{M})$.

Решение проблемы стагнации

Если $r_e(\mathcal{M})$ меньше некоторого порога, алгоритм находится в стагнации. Худшие суперпозиции из $\mathcal{M}$ заменяются на случайные.

Комбинация нейронной сети и символьной регресии

Для повышения качества прогноза строится нейросеть, имеющая в качестве нейронов входного слоя суперпозиции признаков, полученные при помощи генетического алгоритма.

$$f(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{m=1}^{M} \sigma\left(\sum_{n=1}^{N} u_n g_n(\mathbf{x}) + u_0\right) + w_0\right),$$

где M – число нейронов скрытого слоя, N – число нейронов входного слоя. В случае обычной нейронной сети $g_n(\mathbf{x})$ – исходные признаки, а в случае суперпозиции – построенные в результате символьной регрессии функции.

Вычислительный эксперимент

Модели

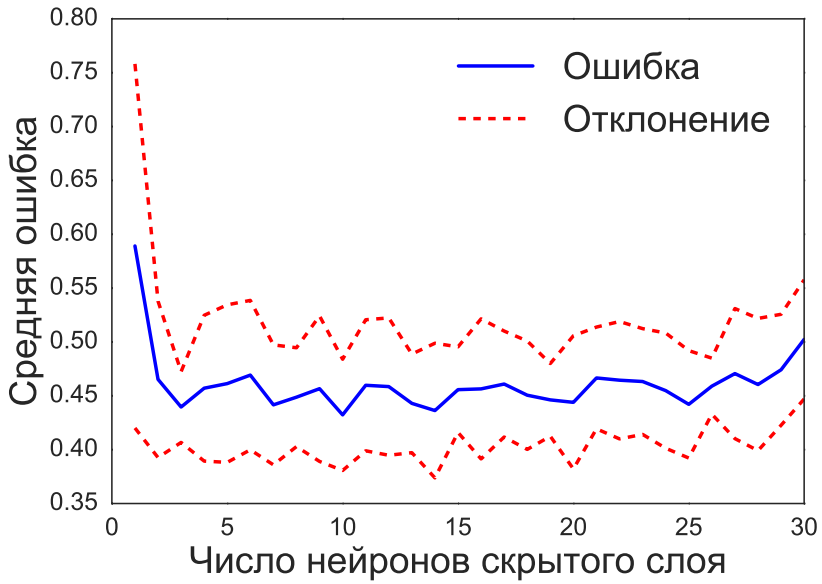
- **Нейронная сеть.** В этом случае значения функций $g_n(\mathbf{x})$ являются исходными признаками. Варьировалось число нейронов скрытого слоя в диапазоне от 1 до 30.
- **Символьная регрессия и нейронная сеть.** Пять раз запускался генетический алгоритм, полученные топ-5 функций использовались в качестве функций $g_n(\mathbf{x})$.
- **Lasso, SVM и градиентный бустинг** сравнивались с нейронной сетью.

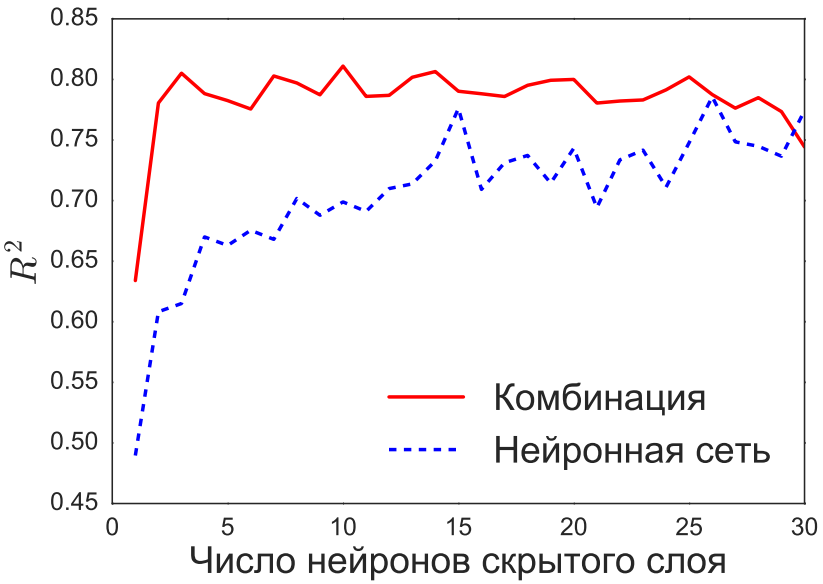
Выборки

- Данные измерения керна
- Airfoil
- Concrete
- Yacht Hydrodynamics

Данные измерения керна

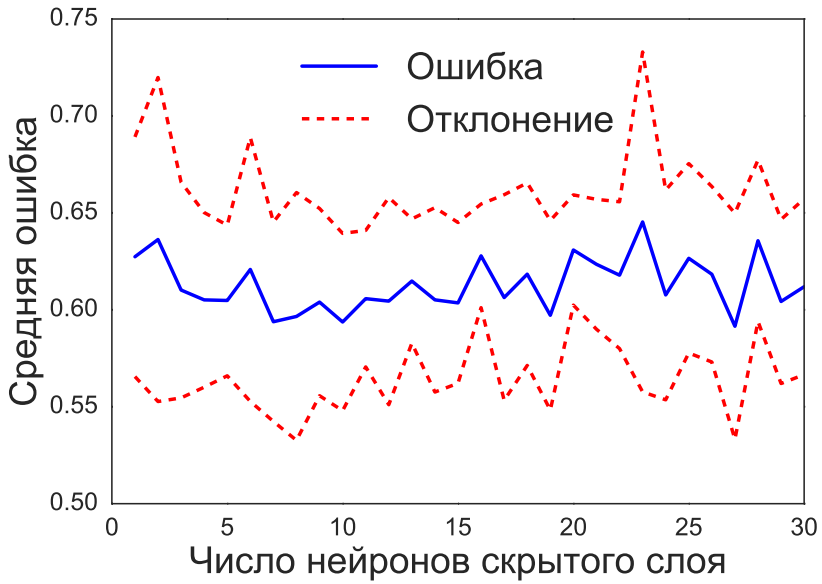
Выборка состоит из 300 образцов керна и 4 признаков: глубины извлечения образца, объемной плотности, минералогической плотности и общей пористости.

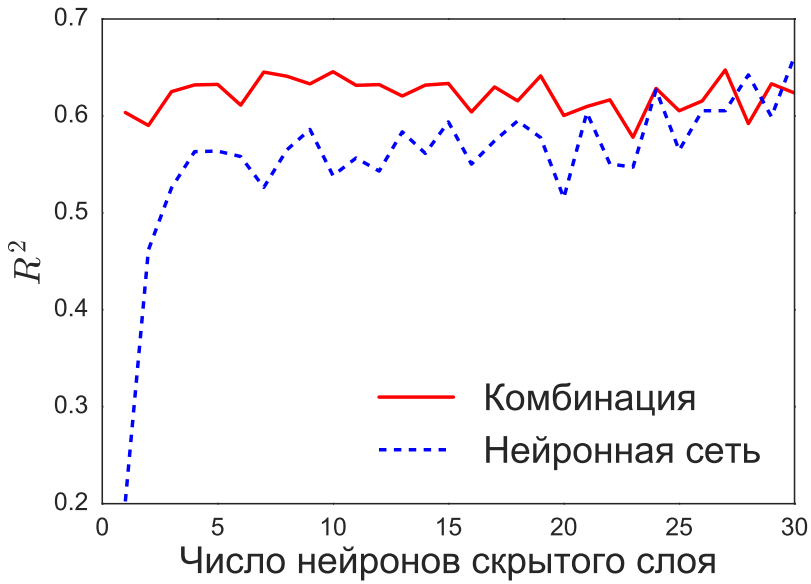




Airfoil

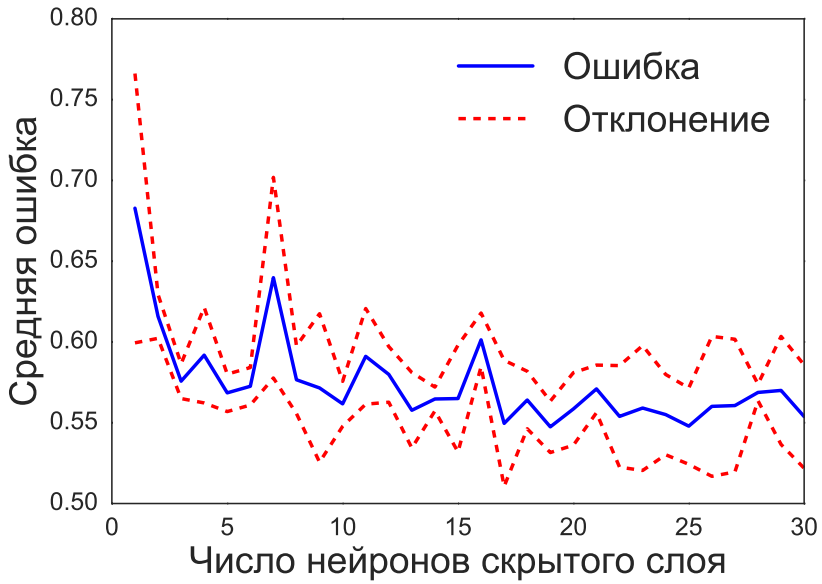
Всего в выборке 1503 образца и 5 признаков.
Требуется предсказать уровень шума в децибелах.

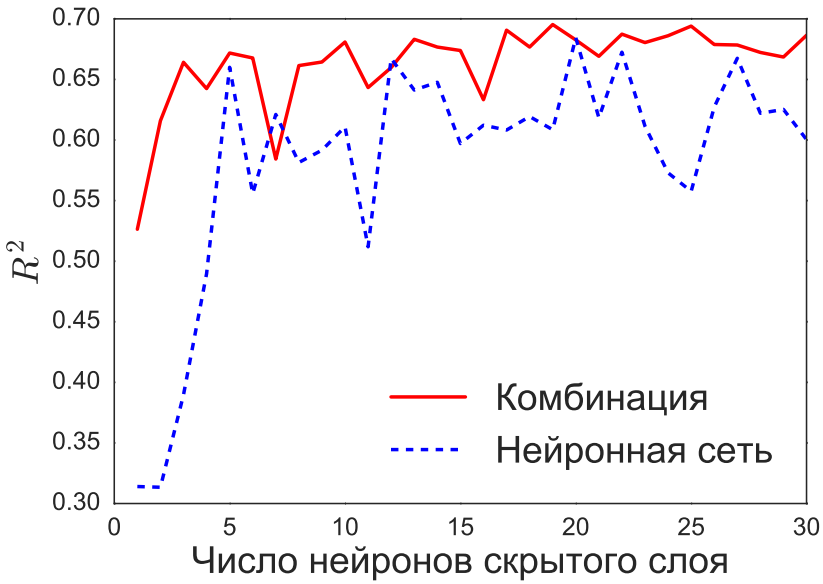




Concrete

Всего в выборке 1030 объектов и 8 признаков.
Требуется предсказать максимально возможную нагрузку на образец.

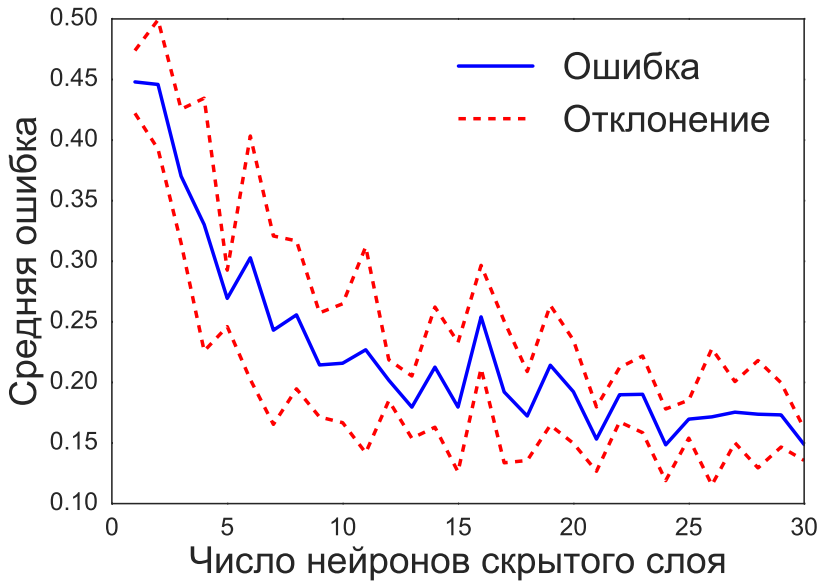


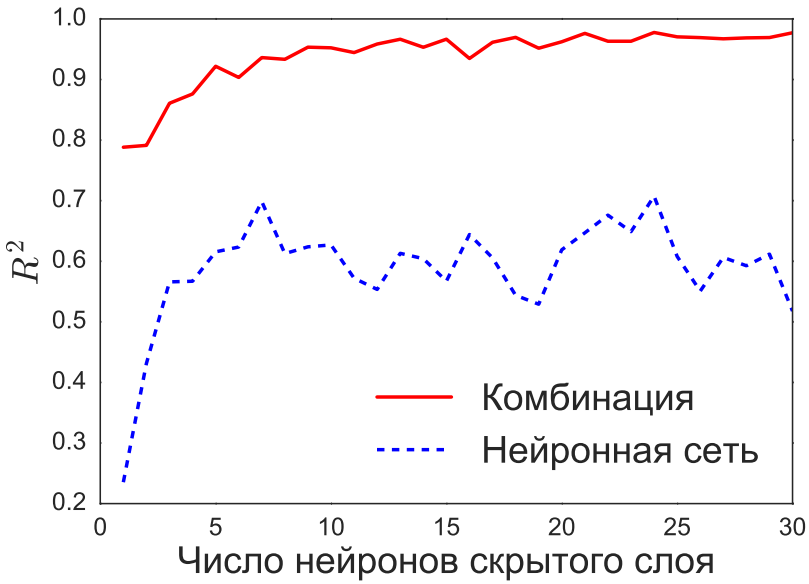


Yacht Hydrodynamics

В выборке представлено 308 объектов и 7 признаков.

Найденная суперпозиция: $g(\mathbf{x}) = \frac{e^{x_5} - \sqrt{2}}{\sqrt{2}}$





Выборка	Lasso	SVM	Бустинг	NN	NN + symb
Керн	$0,64 \pm 0,03$	$0,42 \pm 0,08$	$0,46 \pm 0,03$	$0,47 \pm 0,02$	$0,44 \pm 0,05$
Airfoil	$0,7 \pm 0,01$	$0,48 \pm 0,01$	$0,39 \pm 0,02$	$0,6 \pm 0,02$	$0,61 \pm 0,01$
Concrete	$0,63 \pm 0,02$	$0,51 \pm 0,02$	$0,50 \pm 0,01$	$0,54 \pm 0,03$	$0,55 \pm 0,01$
Yacht	$0,43 \pm 0,01$	$0,31 \pm 0,03$	$0,24 \pm 0,02$	$0,51 \pm 0,02$	$0,15 \pm 0,02$

Результаты

- Предложен алгоритм построения комбинации нейронной сети и символьной регрессии
- Проведен вычислительный эксперимент на четырех выборках
- Показано существенное уменьшение числа нейронов скрытого слоя нейронной сети при использовании символьной регрессии