

# Прогнозирование оптимальных суперпозиций в задачах регрессии

Иннокентий Андреевич Шibaев

Московский физико-технический институт  
Физтех-школа прикладной математики и информатики  
Факультет управления и прикладной математики  
Кафедра «Интеллектуальные системы»

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

Выпускная квалификационная работа бакалавра

Москва 2018

## Цель

Построение алгоритма восстановления матрицы смежности графа суперпозиции по матрице суперпозиции, предсказанной классификатором.

## Задача

Восстановление суперпозиции, задающей модель по ее описанию, с помощью взвешенного графа с покрашенными вершинами.

## Предложение

Использовать постановку задачи линейного программирования для  $k$ -MST, а затем восстановить отброшенные ограничения.

- Бочкарев А. М., Софронов И. Л., Стрижов В. В. Порождение экспертно-интерпретируемых моделей для прогноза проницаемости горной породы // Системы и средства информатики. 2017.
- Chudak F. A., Roughgarden T., Williamson D. P. Approximate  $k$ -MSTs and  $k$ -Steiner trees via the primal-dual method and Lagrangean relaxation // Mathematical Programming. 2004.
- Hegde C., Indyk P., Schmidt L. A fast, adaptive variant of the Goemans-Williamson scheme for the prize-collecting Steiner tree problem // Workshop of the 11th DIMACS Implementation Challenge. Providence, Rhode Island: Workshop of the 11th DIMACS Implementation Challenge. 2014.

## Дано

Семейство базовых функций  $\mathcal{F} = \{f_1, \dots, f_n\}$ , а также множество  $\mathcal{G} = \{g_1, \dots, g_m\}$ ,  $g_i = g_i(f_1, \dots, f_n)$  их суперпозиций. Дан набор выборок  $\mathcal{A} = \{\mathbf{A}_1, \dots, \mathbf{A}_N\}$ ,  $\mathbf{A}_i = (\mathbf{X}_i, \mathbf{y}_i)$  где  $\mathbf{X}_i$  — признаковое описание  $n_i$  объектов  $i$ -й выборки, а  $\mathbf{y}_i = g_i(\mathbf{X}_i)$ ,  $g_i \in \mathcal{G}$ .

## Задача символьной регрессии

Построить суперпозицию  $g^* = g(f_1, \dots, f_n)$  минимизирующую некоторую функцию потерь

$$g^* = \arg \min_{g \in \mathcal{G}} S(g | \mathbf{w}^*, \mathbf{X}, \mathbf{y}),$$

где  $S$  — заданная функция ошибки,  $\mathbf{w}^*$  — оптимальный набор параметров для модели  $f$  при заданных  $\mathbf{X}, \mathbf{y}$

## Формат описания суперпозиции

Для суперпозиции функций строится граф вычисления, матрица смежности которого и является описанием суперпозиции.

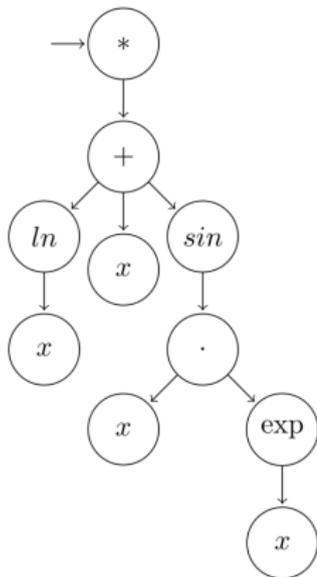
## Задача восстановления матрицы суперпозиции

Дана матрица смежности неориентированного взвешенного графа  $G = (V, E)$  с покрашенными вершинами и выделенной корневой вершиной  $r$ , при этом веса  $w(e_i) = w_i \in [0, 1]$ ,  $e_i \in E$ , а цвета вершин  $c(v_i) = c_i \in \mathbb{N}$ .

Построить максимальное остовное дерево покрывающее в этом графе как минимум  $k$  вершин так, чтобы порядок вершины  $v_i$  после накрытия был равен либо 0, либо  $c_i + 1$  (для корневой вершины порядок равен 1).

# Постановка задачи (формат суперпозиции)

$$f(x) = \ln(x) + x + \sin(x \cdot e^x)$$



arity	f(.)	*	+	ln	sin	·	exp	x
1	*	0	1	0	0	0	0	0
3	+	0	0	1	1	0	0	1
1	ln	0	0	0	0	0	0	1
1	sin	0	0	0	0	1	0	0
2	·	0	0	0	0	0	1	1
1	exp	0	0	0	0	0	0	1

arity	f(.)	*	+	ln	sin	·	exp	x
1	*	0.2	0.7	0.5	0.4	0.5	0.3	0.2
3	+	0.3	0.2	1.	0.8	0.6	0.3	0.7
1	ln	0.3	0.2	0.1	0.	0.1	0.5	0.5
1	sin	0.1	0.4	0.4	0.5	0.9	0.2	0.5
2	·	0.3	0.	0.3	0.5	0.	0.8	0.6
1	exp	0.2	0.3	0.4	0.1	0.5	0.4	0.4

# Постановка задачи ( $k - MST, PCST$ )

## Rooted $k - MST$ ( $k$ -Minimum spanning tree)

Дан неориентированный взвешенный граф  $G = (V, E)$  с выделенной корневой вершиной  $r$ , при этом веса  $w(e_i) = w_i > 0, e_i \in E$ .

Построить минимальное остовное дерево покрывающее в этом графе как минимум  $k$  вершин.

## Rooted $PCST$ (Prize-Collecting Steiner Tree)

Дан неориентированный взвешенный (и вершины и ребра имеют веса) граф  $G = (V, E)$  с выделенной корневой вершиной  $r$ , при этом веса  $w(e_i) = w_i > 0, e_i \in E$  (стоимости) и  $c(v_i) = c_i > 0, v_i \in V$  (призы).

Построить дерево  $T$  максимизирующее функционал прибыли:

$$P = \sum_{v_i \in T} c(v_i) - \sum_{e_i \in E(T)} w(e_i).$$

## Задача ЛП для $PCST$ ( $k - MST$ ) с релаксированными условиями целочисленности

$$\begin{aligned} & \underset{x_e, z_S}{\text{minimize}} && \min \sum_{e \in E} c_e x_e + \lambda \left( \sum_{S \subseteq V \setminus \{r\}} |S| z_S - (n - k) \right) \\ & \text{s.t.} && \sum_{e \in \delta(S)} x_e + \sum_{T: T \supseteq S} z_T \geq 1, \quad \forall S \subseteq V \setminus \{r\}, v \in S, \\ & && x_e \in [0, 1], \quad \forall e \in E \\ & && z_S \in [0, 1], \quad S \subseteq V \setminus \{r\} \end{aligned}$$

В нерелаксированной постановке  $x_e \in \{0, 1\}$ , и  $x_e = 1$  означает что соответствующее ребро взято в дерево  $F$ . Аналогично,  $z_S \in \{0, 1\}$ ,  $z_S = 1$  для множества  $S = V \setminus F$

## Базовые

- *DFS*
- *BFS*
- Алг. Прима

## Основанные на *k-MST*

- *k - MST* через *PCST*
- *k - MST + DFS*
- *k - MST + BFS*
- *k - MST + алг. Прима*

Для каждого из них исследовались неориентированный и ориентированный варианты.

## Предположения

- Наборы арностей функций имеют биномиальное распределение (много функций малой арности).
- Одна переменная (обобщение на несколько переменных — двумя способами).

## Качество восстановления

Необходимо восстановить корректное дерево суперпозиции, так что восстановлением считаем лишь полное совпадение с исходной матрицей.

$$\text{Acc}(R, N, \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} [R(N(M)) = M],$$

где  $N$  — функция зашумления, а  $R$  — алгоритм восстановления.

## Тестовое множество

- 50 наборов арностей (длиной 5 —20)
- 20 функций для каждого набора
- 5 зашумлений каждой функции
- Шум — равномерно распределенный, в отрезке  $[0, 1]$  с шагом 0.02
- Калибровка — линейная в отрезок  $[0, 1]$

## Цель эксперимента

Сравнить  $k - MST$  подходы с остальными

# Результаты экспериментов, неориентированный случай

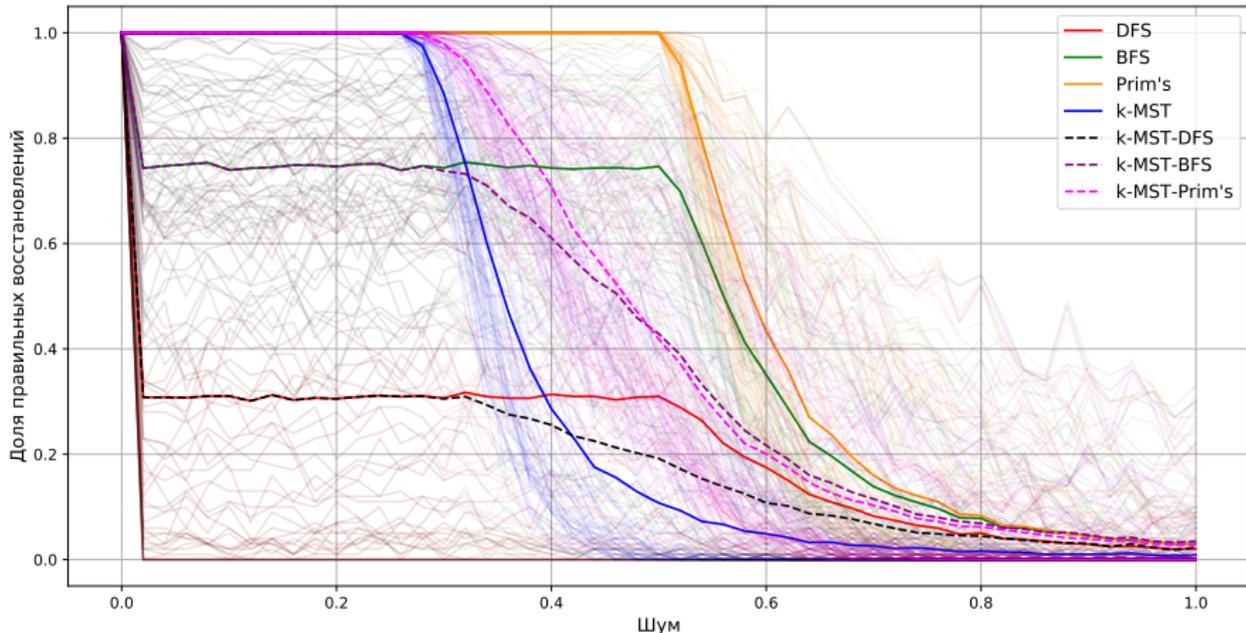


Рис. 1: Зависимость доли правильных восстановлений от силы шума, функции арности 1 – 2. *k-MST* алгоритмы используют симметризованную матрицу смежности

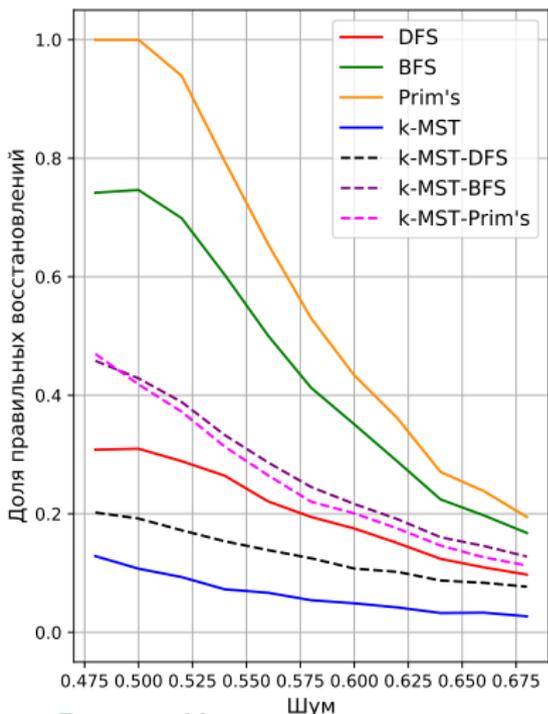


Рис. 2: Неориентированный вариант, качество при шуме  $\sim 0.5$

## Качество при шуме $\sim 0.5$

Шум	.50	.52	.54	.56	.58
<i>DFS</i>	.31	.29	.26	.22	.19
<i>BFS</i>	.75	.70	.60	.50	.41
Прим	<b>1.0</b>	<b>.94</b>	<b>.79</b>	<b>.66</b>	<b>.53</b>

и для неориентированных вариантов алгоритмов на основе *k-MST*:

Шум	.50	.52	.54	.56	.58
<i>k-MST</i>	.11	.09	.07	.07	.05
+ <i>DFS</i>	.19	.17	.15	.14	.12
+ <i>BFS</i>	.43	.39	.33	.29	.25
+ <i>Prim's</i>	.42	.37	.31	.26	.22

# Результаты экспериментов, ориентированный случай

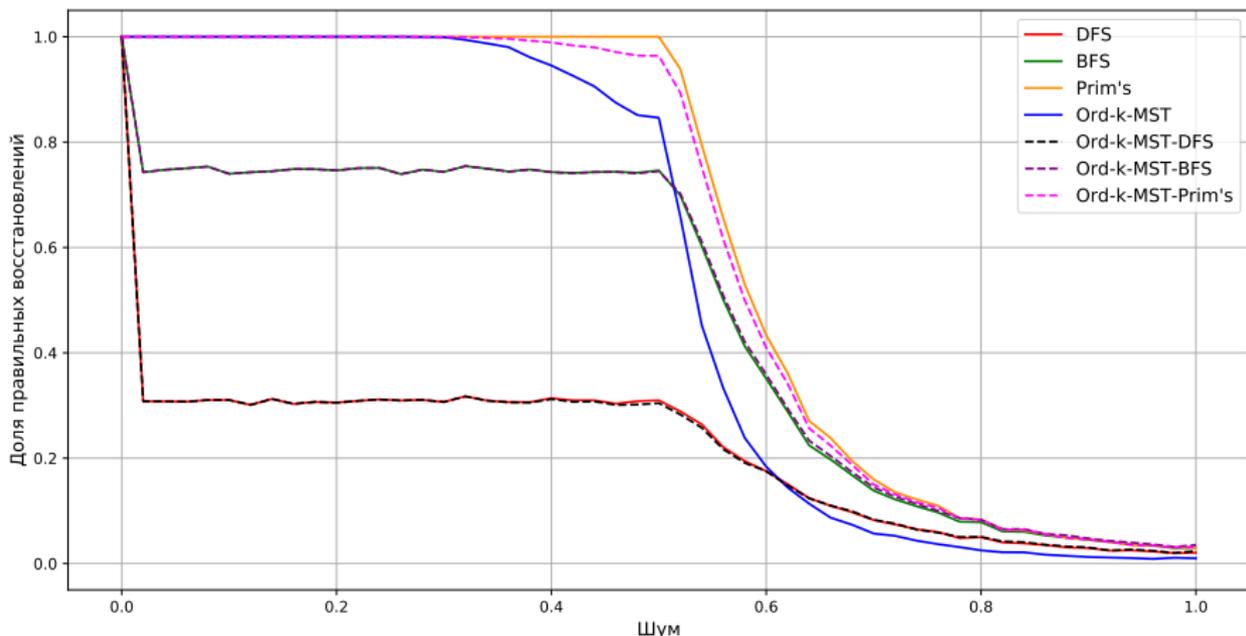


Рис. 3: Зависимость доли правильных восстановлений от силы шума, функции арности 1 – 2.  $k$ -MST алгоритмы используют исходную матрицу смежности

# Результаты экспериментов, ориентированный случай

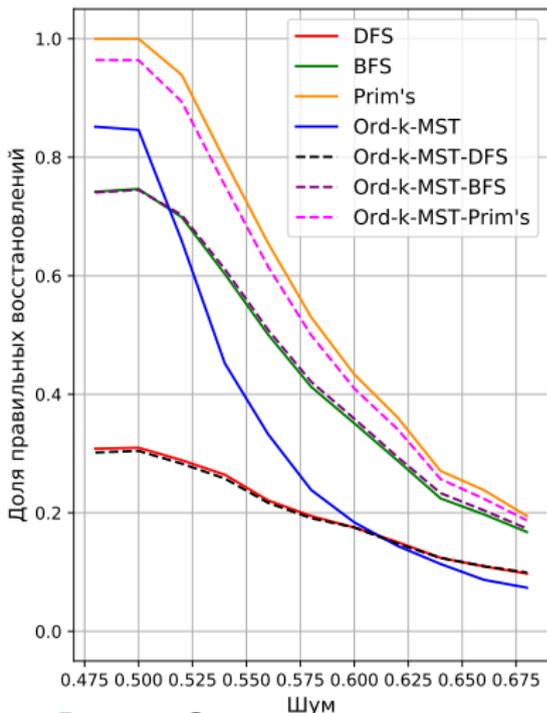


Рис. 4: Ориентированный вариант, качество при шуме  $\sim 0.5$

## Качество при шуме $\sim 0.5$

Шум	.50	.52	.54	.56	.58
<i>DFS</i>	.31	.29	.26	.22	.19
<i>BFS</i>	.75	.70	.60	.50	.41
Прим	<b>1.0</b>	<b>.94</b>	<b>.79</b>	<b>.66</b>	<b>.53</b>

и для ориентированных вариантов алгоритмов на основе *k-MST*:

Шум	.50	.52	.54	.56	.58
<i>k-MST</i>	.85	.66	.45	.33	.24
+ <i>DFS</i>	.30	.28	.26	.22	.19
+ <i>BFS</i>	.74	.70	.61	.51	.42
+ <i>Prim's</i>	.96	.89	.75	.62	.50

Алгоритм	Малые арности время, с.	Большие арности время, с.
<i>DFS</i>	51	48
<i>BFS</i>	53	51
<i>Prim's</i>	85	67
<i>k-MST</i>	182	151
<i>k-MST-DFS</i>	196	162
<i>k-MST-BFS</i>	204	166
<i>k-MST-Prim's</i>	241	187
<i>Ord-k-MST</i>	168	127
<i>Ord-k-MST-DFS</i>	175	131
<i>Ord-k-MST-BFS</i>	171	133
<i>Ord-k-MST-Prim's</i>	195	149

Таблица 1: Полное время работы на датасете (250000 запусков)

- Алгоритм Прима работает лучше всех остальных, и полностью восстанавливает при шуме до 0.5
- Ориентированный вариант алгоритма на основе  $k$ -MST работает существенно лучше неориентированного
- $k$  – MST препроцессинг существенно повышает качество работы алгоритма зависящего от порядка обхода ( $BFS$ )
- Для  $k$  – MST параметр  $\lambda$  в  $PCST$  надо положить равным 0.5
- При наличии даже малого кол-ва базовых функций больших арностей качество восстановления суперпозиций для любого из алгоритмов снижается.