Прогнозирование оптимальных суперпозиций в задачах регрессии

Иннокентий Андреевич Шибаев

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 исследования

Задача

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

Цель

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

Предложение

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

Литература

- 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.

Постановка задачи

Дано

Семейство базовых функций $\mathcal{F} = \{f_1, \ldots, f_n\}$, а также множество $\mathcal{G} = \{g_1, \ldots, g_m\}$, $g_i = g_i(f_1, \ldots, f_n)$ их суперпозиций. Дан набор выборок $\mathcal{A} = \{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$, $\mathbf{A}_i = (\mathbf{X}_i, \mathbf{y}_i)$ 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 $g^* = g(f_1, \ldots, f_n)$ минимизирующую некоторую функцию потерь

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

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

Постановка задачи

Формат описания суперпозиции

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

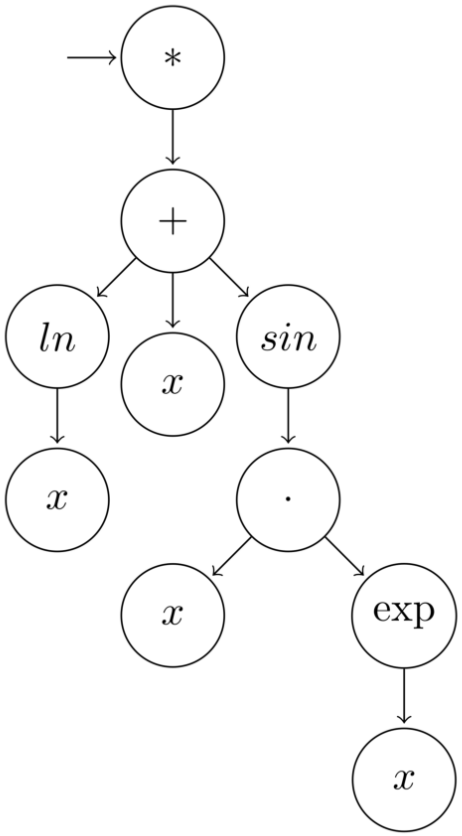
Задача восстановления матрицы суперпозиции

Дана матрица смежности неорентированного 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 накрывающее 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*	$\cdot$	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	$\cdot$	0	0	0	0	0	1	1
1	exp	0	0	0	0	0	0	1

arity	$f(.)$	$*$	$+$	*ln*	*sin*	$\cdot$	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	$\cdot$	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 неорентированный 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 накрывающее в этом графе как минимум k вершин.

Rooted $PCST$ (Prize-Collecting Steiner Tree)

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

Постановка задачи ($k - MST$, $PCST$)

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

$$\underset{\substack{x_e, z_S \\ e \in E, S \subseteq V \setminus \{r\}}}{\text{minimize}} \quad \min \sum_{e \in E} c_e x_e + \lambda \left(\sum_{S \subseteq V \setminus \{r\}} |S| z_S - (n - k) \right)$$

$$\text{s.t.} \quad \sum_{e \in \delta(S)} x_e + \sum_{T : T \supseteq S} z_T \geq 1, \ \forall S \subseteq V \setminus \{r\}, \ v \in S,$$

$$x_e \in [0, 1], \ \forall e \in E$$

$$z_S \in [0, 1], \ S \subseteq V \setminus \{r\}$$

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

Вычислительный эксперимент

Предположения

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

Метрика качества восстановления

В конечном счете нам нужно корректное дерево суперпозиции, так что восстановление мы будем считать лишь полное совпадение с исходной матрицей.

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

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

Вычислительный эксперимент

Варианты алгоритма

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

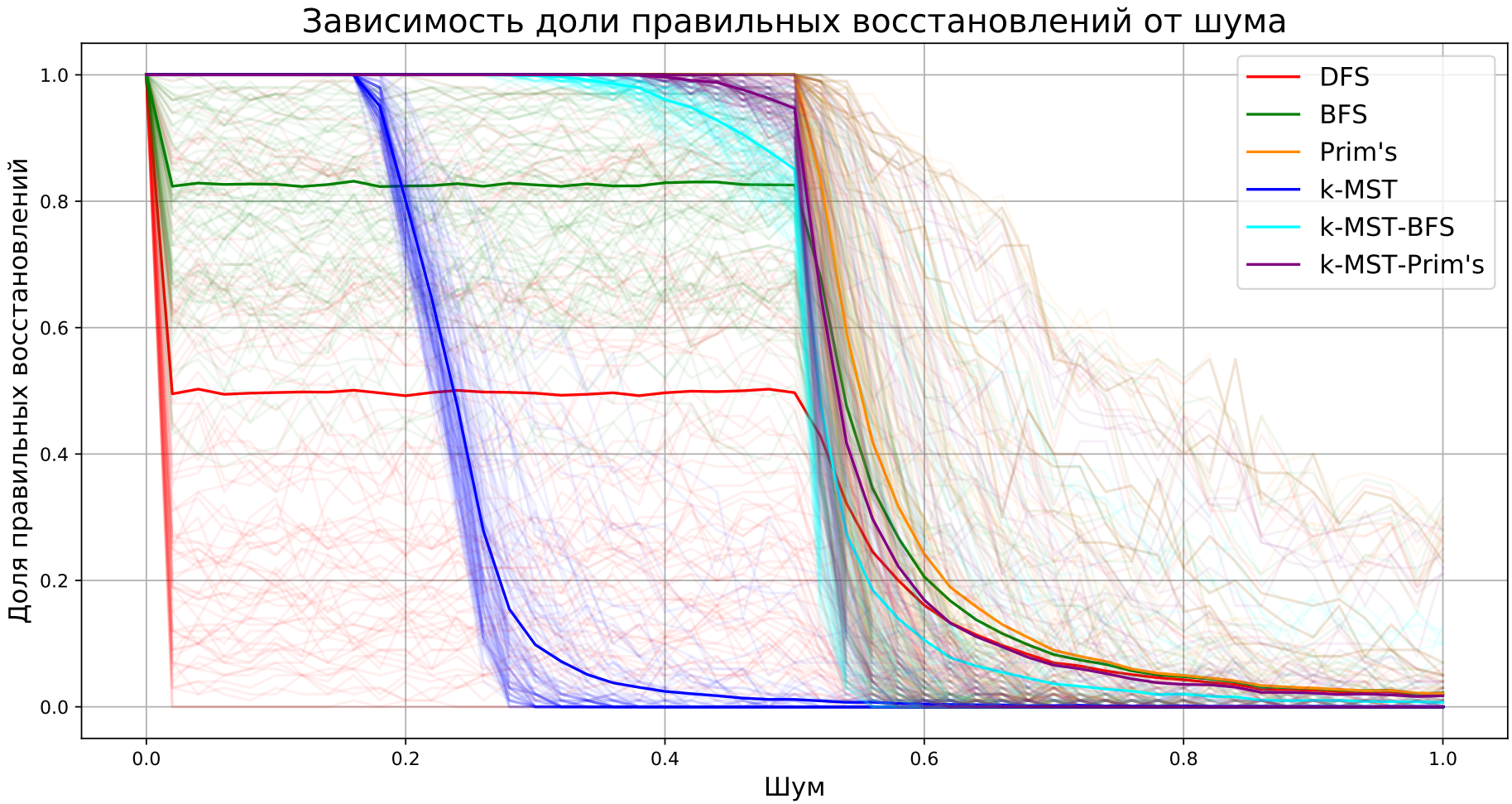
Тестовое множество

- 100 наборов арностей (длиной 5 —20)
- 20 функций для каждого набора
- 5 зашумлений каждой функции
- Шум — равномерно распределенный

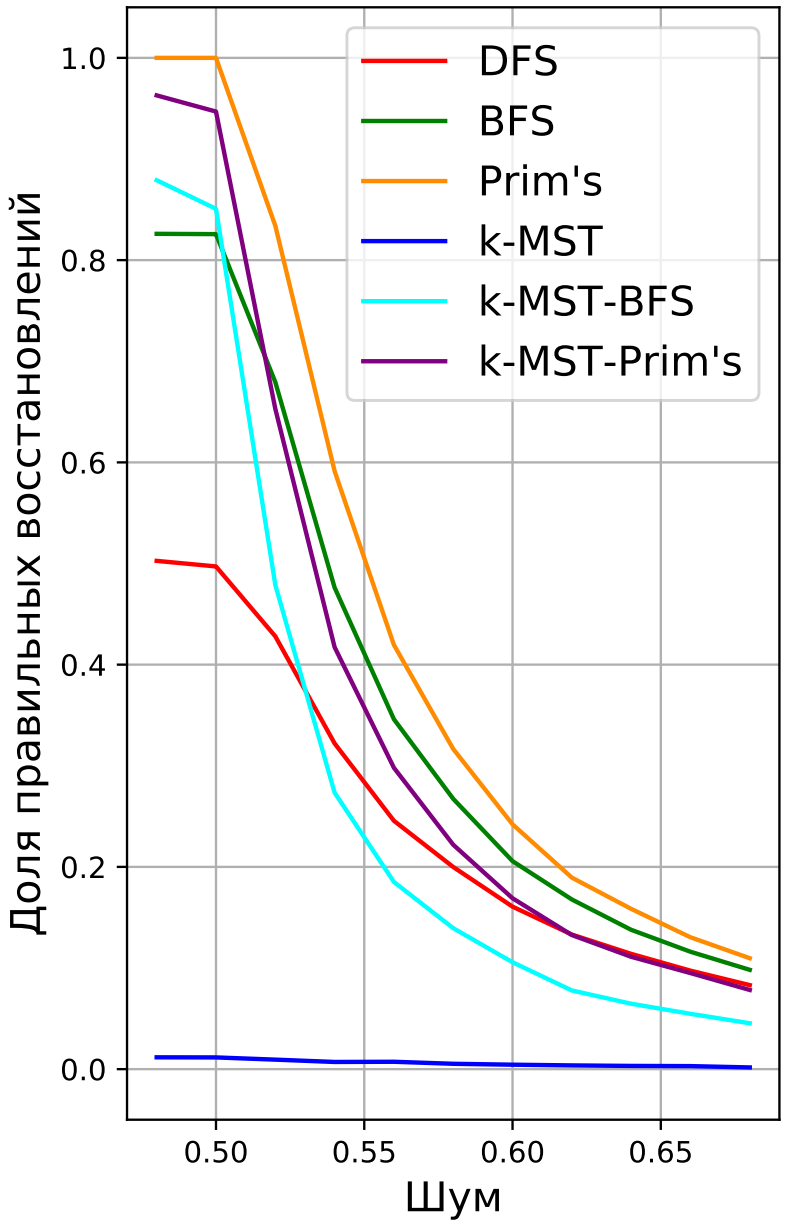
Цель эксперимента

Сравнить $k - MST$ подходы с остальными

Результаты экспериментов, малые арности



Результаты экспериментов, малые арности



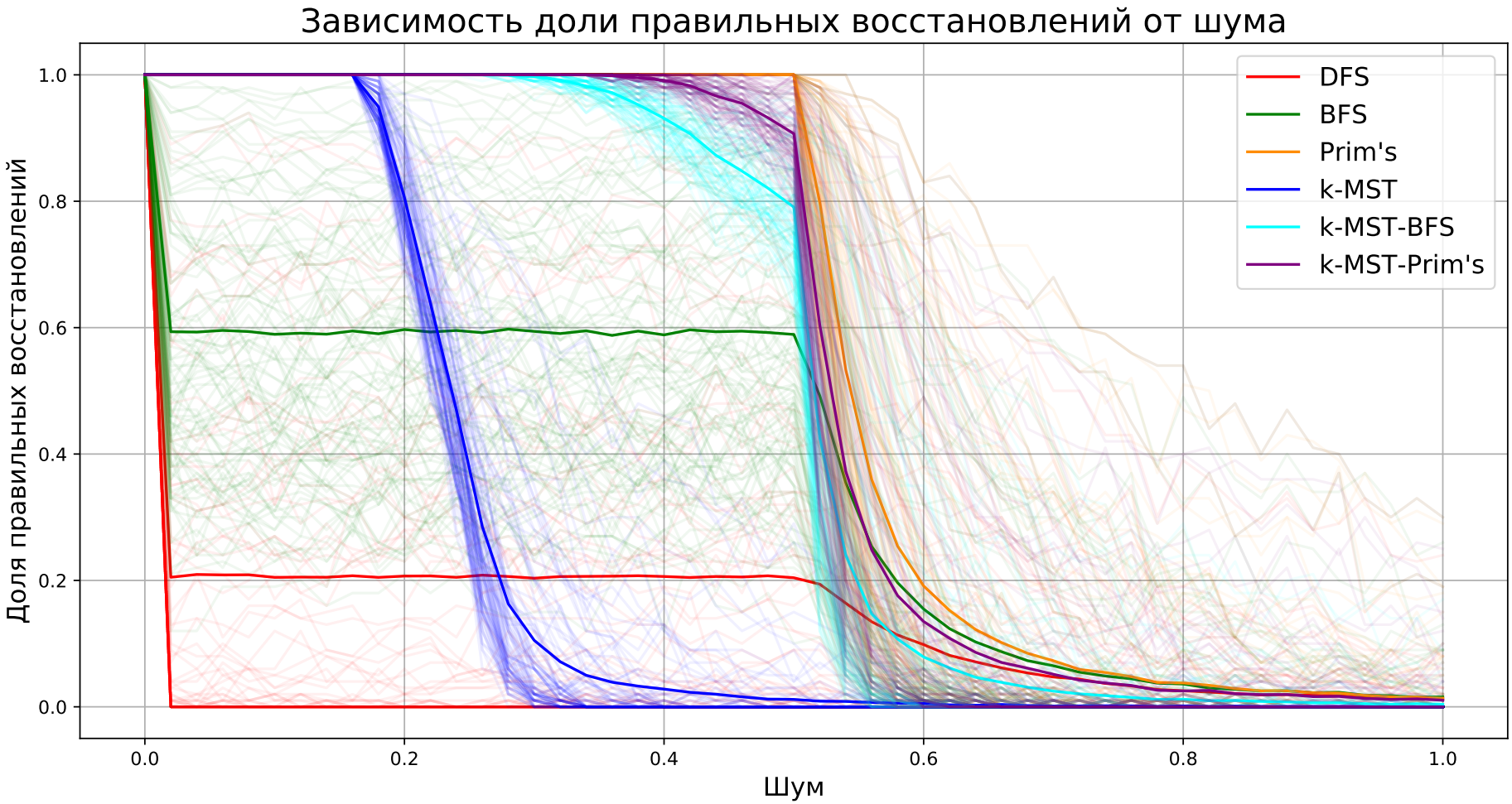
Время работы (500000 запусков)

base	*time*, *s*		k-MST	*time*, *s*
DFS	186		*None*	1272
BFS	181		*BFS*	1221
Прим	571		Прим	1745

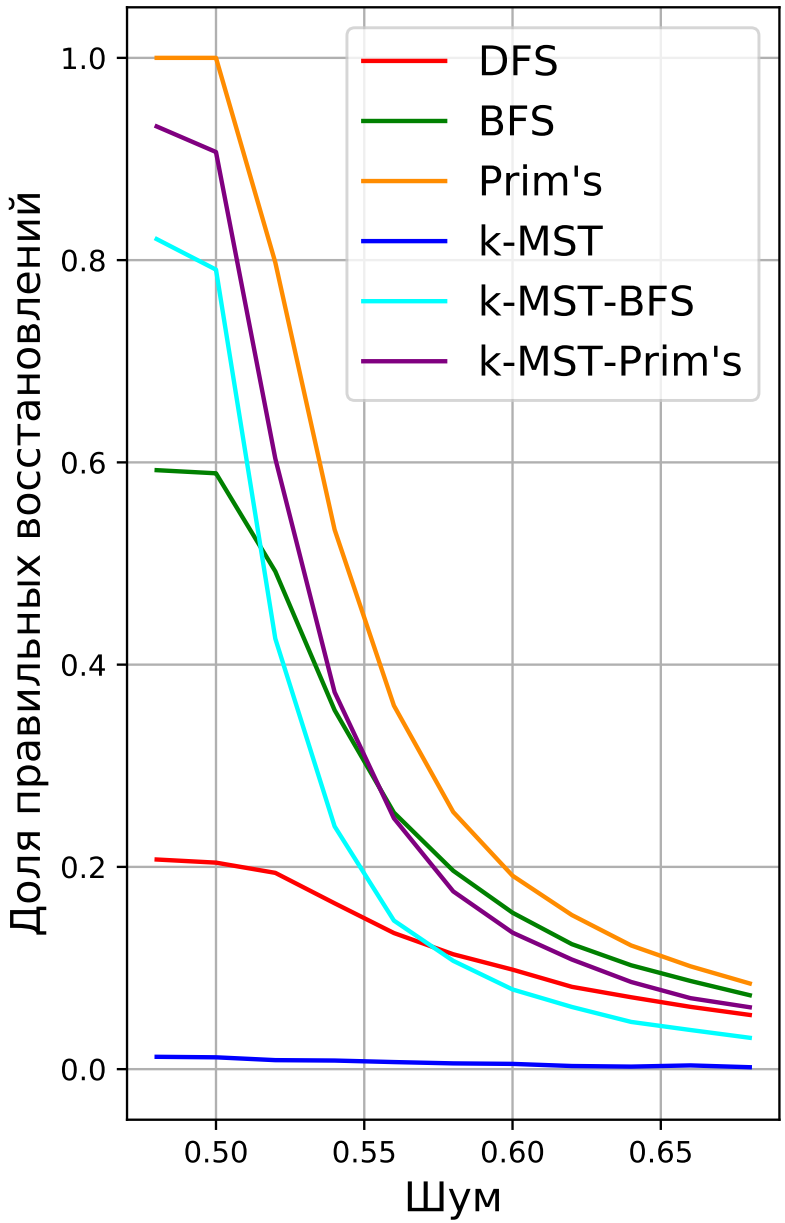
Качество при шуме ~ 0.5

Шум	.50	.52	.54	.56	.58
DFS	.50	.43	.32	.25	.20
BFS	.83	.68	.48	.35	.27
Прим	**1.0**	**.83**	**.59**	**.42**	**.32**
None	.01	.01	.01	.01	.01
BFS	.85	.48	.27	.19	.14
Прим	.95	.65	.42	.30	.22

Результаты экспериментов, большие арности



Результаты экспериментов, большие арности



Время работы (500000 запусков)

base	*time*, *s*		k-MST	*time*, *s*
DFS	201		*None*	1061
BFS	197		*BFS*	1029
Прим	685		Прим	1581

Качество при шуме ~ 0.5

Шум	.50	.52	.54	.56	.58
DFS	.20	.19	.16	.13	.11
BFS	.59	.49	.36	.25	.20
Прим	**1.0**	**.80**	**.53**	**.36**	**.25**
None	.01	.01	.01	.01	.01
BFS	.79	.43	.24	.15	.11
Прим	.91	.60	.37	.25	.18

Заключение

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