Методы структурного обучения для построения прогностических моделей

Варфоломеева А. А.

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

Научный руководитель к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

Москва,
2013 г.

Цель исследования

Предложить метод прогнозирования структуры суперпозиции регрессионной модели, описывающей предъявленную выборку оптимальным образом.

Проблема

Алгоритмы выбора моделей имеют значительную вычислительную сложность в связи с необходимостью перебора большого числа моделей.

Предложение

Основываясь на собранных прецедентах выбора моделей, адекватно описывающих выборки, построить алгоритм прогноза структуры таких моделей.

Базовые публикации

1. Koza, J. R. Genetic programming // Encyclopedia of Computer Science and Technology, 1998. Vol. 39. P. 29–43.
2. Г.И. Рудой, В.В. Стрижов. Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных // Информатика и её применения, 2013. Том 7, № 1. С. 44–53.
3. Jaakola T., Sontag D. Learning Bayesian Network Structure using LP Relaxations // Proceedings of the 13th International Conference on Artifcial Intelligence and Statistics, 2010. Vol 9(1). P. 358–365.
4. I. Arel, D. C. Rose, T. P. Karnowski. Deep Machine Learning—A New Frontier in Artificial Intelligence Research // IEEE Computational Intelligence Magazine, November 2010. P. 13-19.

Постановка задачи структурного обучения

Дано:

- набор $\mathfrak{D} = \{(\mathbf{D}_k, f_k)\}$;
- $\mathbf{D}_k = (\underset{m\times n}{\mathbf{X}}, \underset{m\times 1}{\mathbf{y}})$;
- $f_k \in \mathcal{F}$ — модель, оптимально приближающая $\mathbf{D}_k$;
- $\mathcal{G}$ — множество порождающих функций;
- $\mathcal{F}$ — множество суперпозиций порождающих функций $g \in \mathcal{G}$:

$$\mathcal{F} = \{f_s \mid \mathbf{f}_s : (\hat{\mathbf{w}}_k, \mathbf{X}) \mapsto \mathbf{y}, s \in \mathbb{N}\}.$$

Требуется:

найти алгоритм $a : \mathbf{D}_k \mapsto f_s$.

Постановка задачи структурного обучения

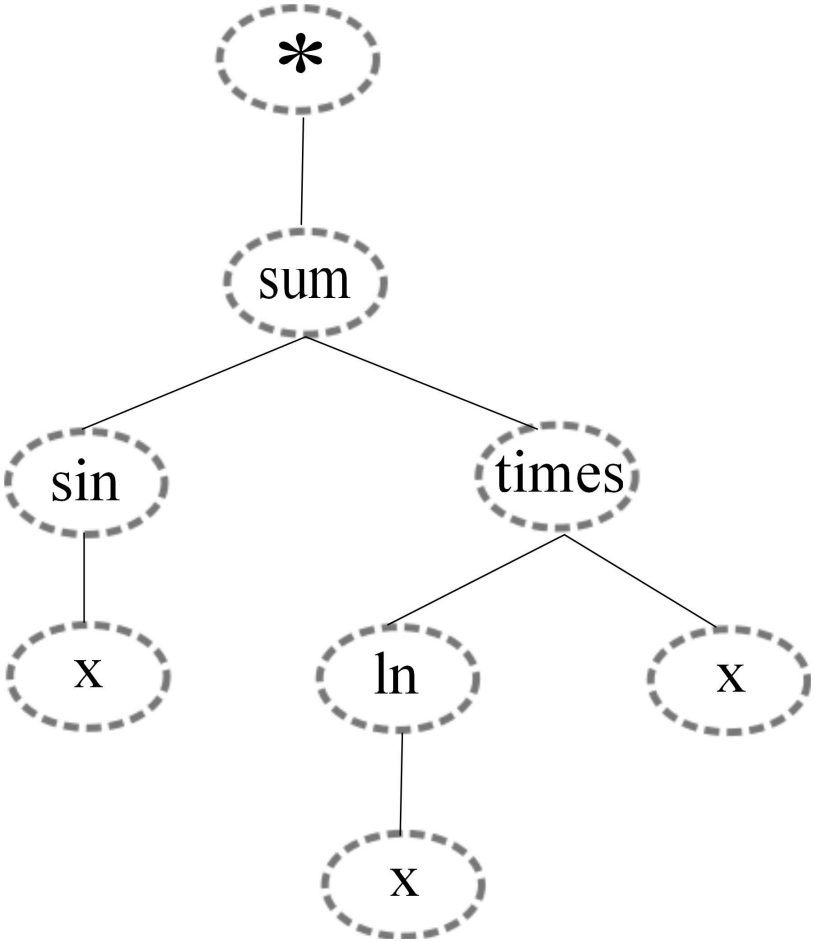
Для множества всех суперпозиций из $\mathcal{F}$ требуется найти такой индекс $\hat{s}$, что функция $f_{\hat{s}}$ среди $f \in \mathcal{F}$ доставляет минимум функции ошибки S:

$$\hat{s} = \arg \min_{s \in \{1,\ldots,|\mathcal{F}|\}} S(f_s \mid \hat{\mathbf{w}}_k, \mathbf{D}_k),$$

где $\hat{\mathbf{w}}_k$ — оптимальный вектор параметров модели f_s для каждой $f \in \mathcal{F}$ при фиксированной $\mathbf{D}$:

$$\hat{\mathbf{w}}_k = \arg \min_{\mathbf{w} \in \mathbb{W}_s} S(\mathbf{w} \mid f_s, \mathbf{D}_k).$$

Правила построения дерева Γ_f суперпозиции f



Дерево Γ_f

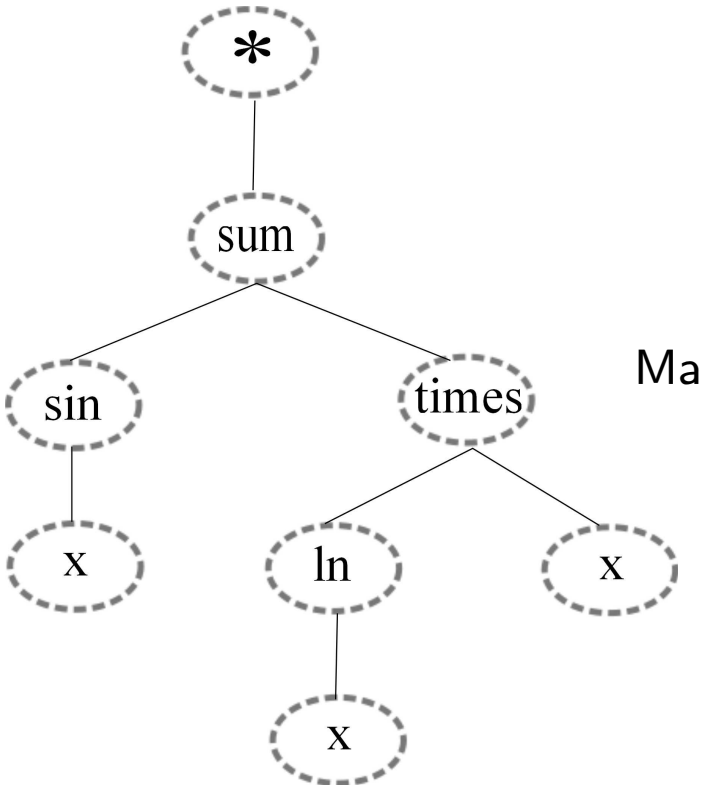
1. Корень дерева - $*$;
2. $V_i \mapsto g_r$;
3. $\text{val}(V_j) = v(g_{r(i)})$;
4. $\text{dom}(g_{r(i)}) \supset \text{cod}(g_{r(j)})$;
5. аргументы g_r упорядочены;
6. x_i — листья Γ_f.

$$f = \sin(x) + (\ln x)x;$$

Правила построения дерева Γ_f суперпозиции f

1. корнем дерева является специальный символ “ * ”, имеющий одну дочернюю вершину;
2. в остальных вершинах V_i дерева Γ_f находятся элементарные функции из набора $\mathcal{G}$;
3. число дочерних вершин V_j у некоторой вершины V_i равно арности соответствующей функции g_r: $v = v(g_r)$;
4. область определения функции дочерней вершины V_j содержит область значений функции родительской вершины V_i: $\mathrm{dom}(g_{r(i)}) \supset \mathrm{cod}(g_{r(j)})$;
5. порядок смежных некоторой вершине V_i вершин соответствует порядку аргументов соответствующей функции $g_r, r = r(i)$;
6. в листьях дерева Γ_f находятся свободные переменные x_i.

Ограничение на построение Z_f



$$f = \sin(x) + (\ln x)x$$

Матрица связей Z_f дерева Γ_f

	sum	times	ln	sin	x
$*$	1	0	0	0	0
sum	0	1	1	0	0
times	0	0	0	1	1
ln	0	0	0	0	1
sin	0	0	0	0	1

Матрица вероятностей связей P_f дерева Γ_f

	sum	times	ln	sin	x
$*$	0.7	0.1	0.1	0.1	0.2
sum	0.2	0.7	0.8	0.1	0.2
times	0.1	0.3	0	0.8	0.8
ln	0.2	0.1	0.3	0.1	0.9
sin	0.1	0.2	0.1	0	0.8

$\mathcal{M}$ — множество матриц, соответствующих суперпозициям из $\mathcal{F}$.

Алгоритм прогнозирования структуры суперпозиции

$a : \mathbf{D}_k \mapsto f_s$.

Заданы:

прецеденты $\mathbf{D}_s$, состоящие из пар $(\mathbf{X}, \mathbf{y})$.

Требуется:

найти матрицу вероятностей связи P_s;
найти $Z_{f_s} = \arg \max\limits_{Z \in \mathcal{M}} \sum_{i,j} P_{ij} \times Z_{i,j}$.

Построение дерева $\hat{\Gamma}_f$

Задана матрица P, состоящая из блока $P'_{l+1\times l}$:

$$P'_{ij} = p(g_i \to g_j);$$

и блока $P''_{l+1\times n}$:

$$P''_{ik} = p(g_i \to x_k).$$

Требуется: построить матрицу Z_f дерева $\hat{\Gamma}_f$.

Вершина i открыта, если
$(i \leq l)\&(\exists j : (j,i) = 1)\&(\nexists k : (i,k) = 1)$.

Процедура построения дерева $\hat{\Gamma}_f$

Задано K — максимально допустимая сложность суперпозиции.

- Объявляем вершину дереву открытой: $i = 1$.
- Пока количество единиц в матрице не превышает K, повторяем:
  1. выбираем $c_j = \max\limits_{j=1,\ldots,l} P_{ij}$ для всех открытых i;
  2. достраиваем матрицу: $j^* = \arg\max\limits_j c_j, \quad (i, j^*) = 1$;
  3. добавляем j^* к списку открытых вершин, если $(i, j^*) \in P'$;
- если количество единиц превышает K, ставим в соответствие открытым вершинам независимые переменные: $k^* = \arg\max\limits_k P''_{ik}, \quad (i, k^*) = 1$ для всех i-открытых.

Цель эксперимента

Верифицировать предложенную процедуру прогноза суперпозиций.

Задана выборка $\mathfrak{D} = \{(\mathbf{D}_s, f_s)\}$ и алгоритм a:$\mathbf{D} \mapsto \hat{\Gamma}$.
Выполняется процедура LOO следующего вида:

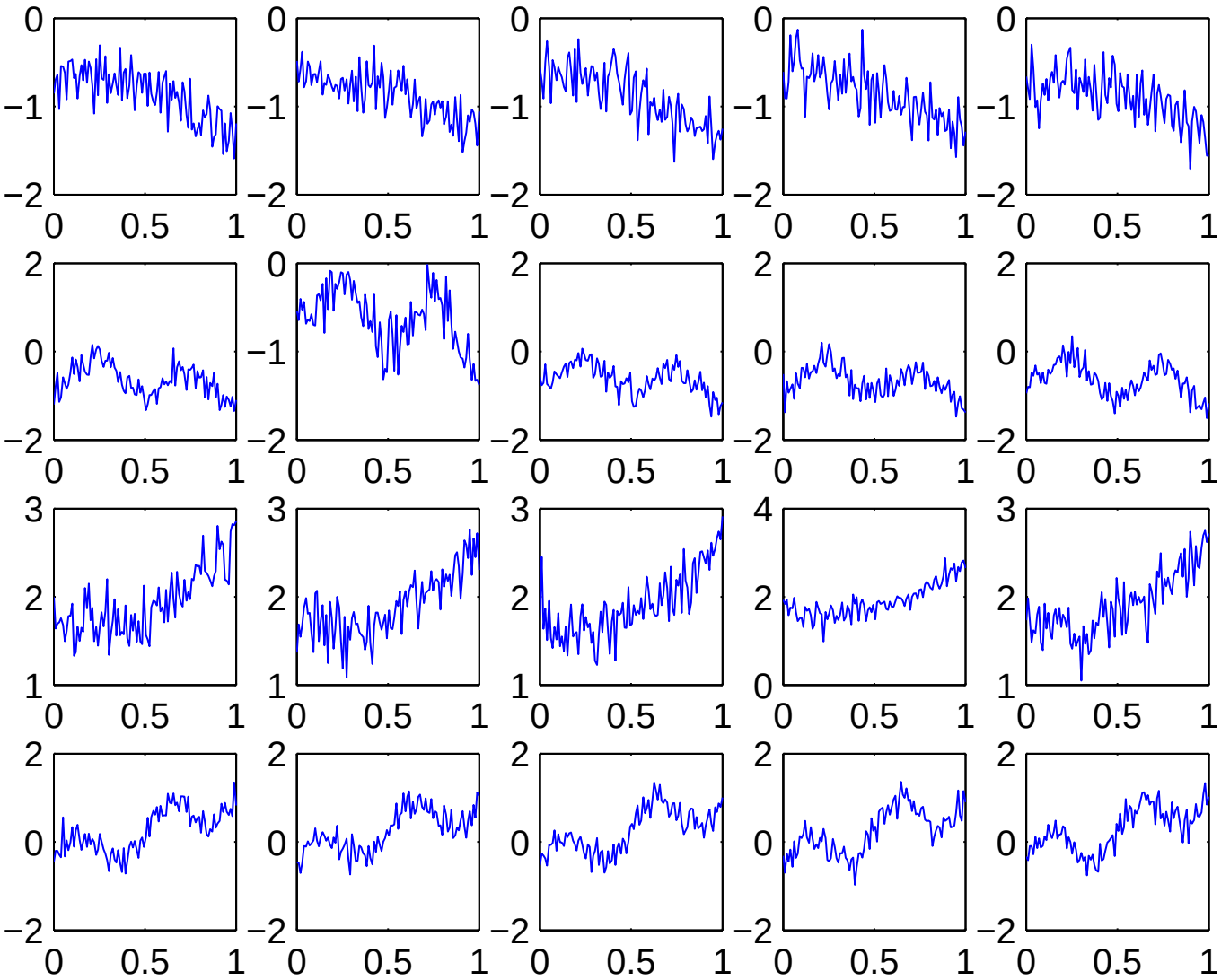
1. Оптимизируются параметры a по обучающей подвыборке $\mathfrak{D} \setminus \{\mathbf{D}_k\}$.
2. Вычисляется $\hat{\Gamma}_k = a(\mathbf{D}_k)$.
3. По $\hat{\Gamma}_k$ строится модель $\hat{f}_s$.
4. Настраиваются параметры $\hat{\mathbf{w}}_k$ модели $\hat{f}_s$.
5. Вычисляется $S(\hat{\mathbf{w}}_k, \hat{f}_s, f_s) = \| \mathbf{y} - f(\mathbf{w}_k, \mathbf{X}) \|_2$.

Порождение синтетической выборки

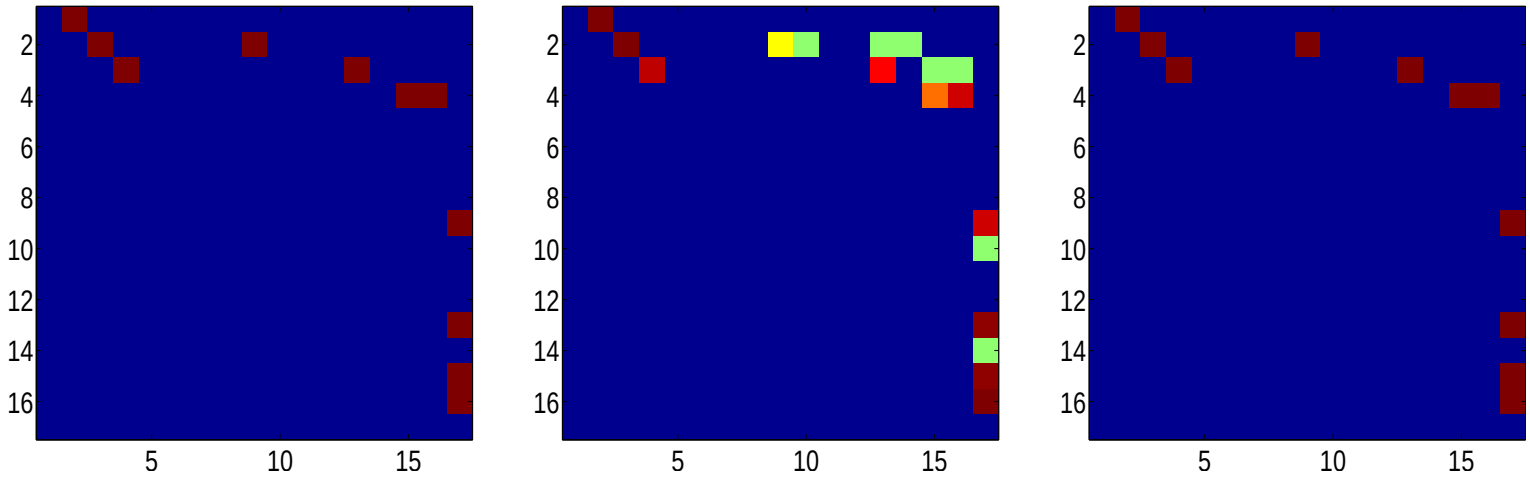
1. Фиксируем модель f_s из множества $\mathcal{F}$ и параметры $\mathbf{w}_s \in \mathbb{W}_s$;
2. задаем значения $\mathbf{X}$;
3. вычисляем $f(\mathbf{w}_s, \mathbf{X})$;
4. фиксируем τ_f, $|\tau_f| < \epsilon$;
5. вычисляем $\mathbf{y} = f(\mathbf{w}_s, \mathbf{X}) + \tau_f$;
6. повторяем r раз для каждой модели $f \in \mathcal{F}$

Получаем множество прецедентов: $\mathbf{D} = (\underset{m \times n}{\mathbf{X}}, \underset{m \times 1}{\mathbf{y}})$ и соответствующие им модели f.

Вид выборки

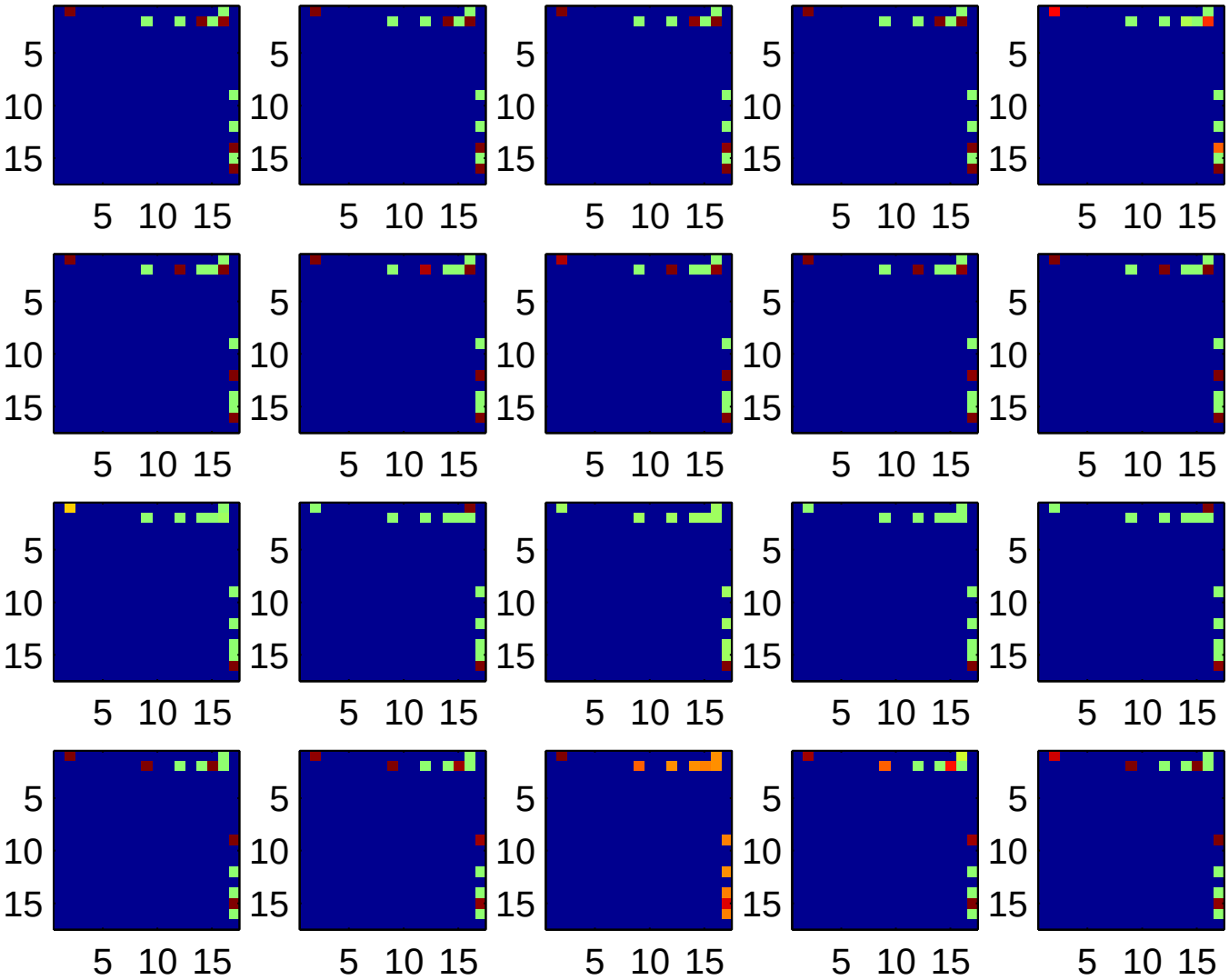


Исходная и спрогнозированная суперпозиция

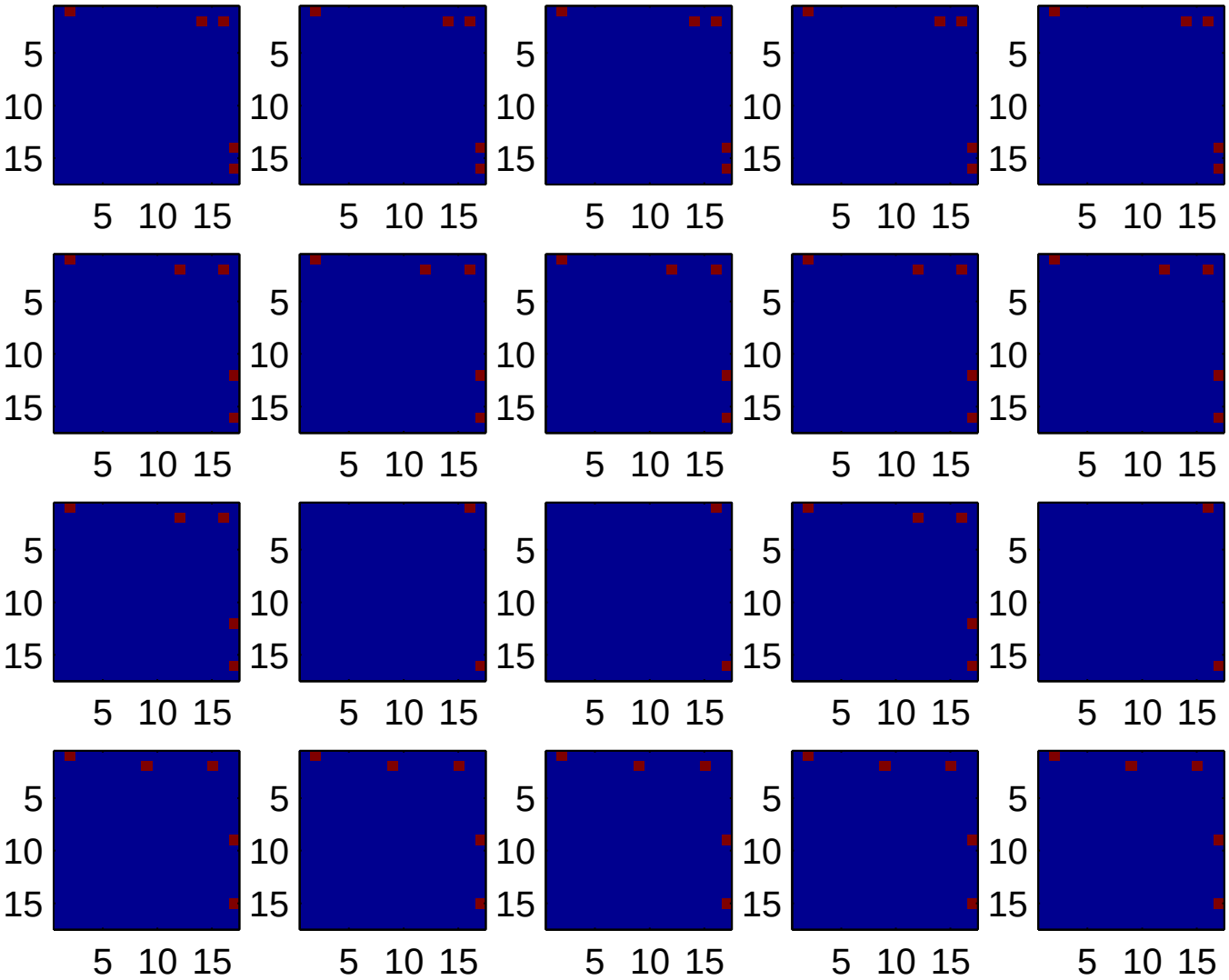


$$f = w_1 \cos(x) + w_2 \sin(x) + w_3 x + w_4 \ln(x + 1).$$

Полученные матрицы вероятностей P_f



Построенные деревья Γ_f



Зависимость ошибки от шума и параметров модели

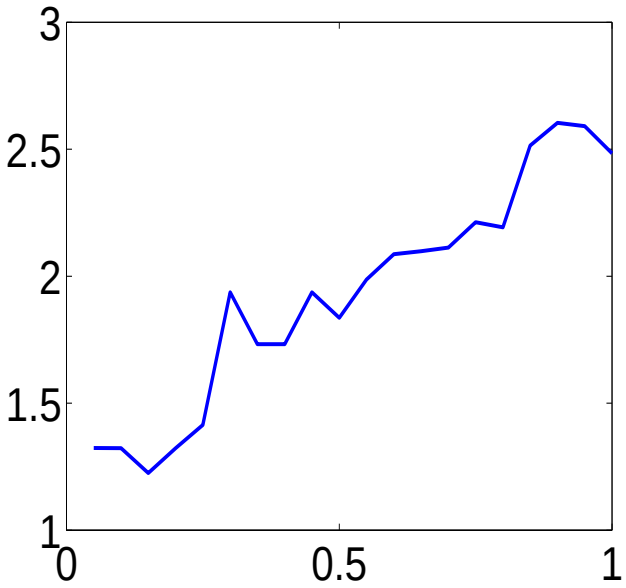


Рис.: Зависимость S от размера ϵ

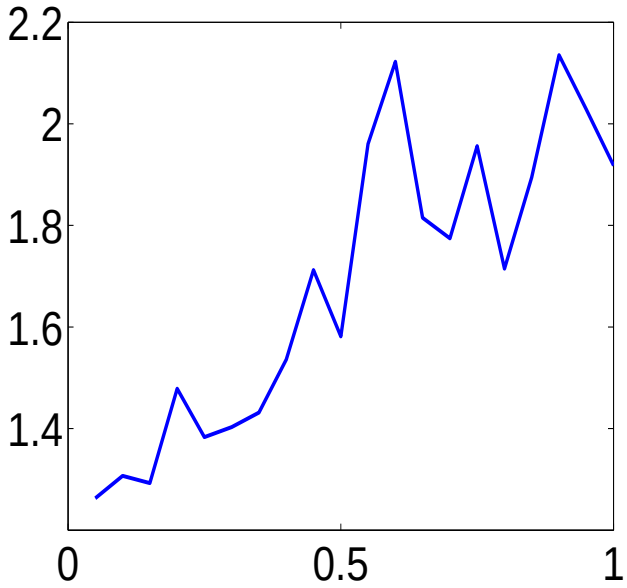


Рис.: Зависимость S от размера $\delta \mathbf{w}$

Публикации по теме

Варфоломеева А.А. Локальные методы прогнозирования с выбором метрики // Машинное обучение и анализ данных, 2012. Т.1, Вып. 3. Стр. 367–375.

Варфоломеева А.А., Стрижов В.В. Выбор признаков при разметке библиографических списков методами структурного обучения // Вестник СПбГУ, подано в редакцию.

Заключение

- Поставлена и решена задача прогнозирования структуры суперпозиций регрессионных моделей.
- Предложено описание допустимых суперпозиций, удовлетворяющее необходимым ограничениям.
- Предложен алгоритм построения допустимой суперпозиции по вероятностной матрице прогноза.
- Разработан алгоритм прогнозирования структуры регрессионной модели.
- На синтетических данных предложенный алгоритм показал адекватные результаты.