Методы структурного обучения в задаче обобщения структур ранжирующих моделей

Варфоломеева Анна

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

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

Москва,
2015 г.

Цель работы

Задача

Предложить метод прогнозирования структуры суперпозиции ранжирующей функции, описывающей предъявленную выборку оптимальным образом. Выборка состоит из пар “описание объекта – соответствующее ему оптимальная суперпозиция”.

Исследуемая проблема

Требуется построить ранжирующую модель для коллекции документов. Модель присваивает каждому документу ранг согласно поступающему запросу. Предлагается алгоритм прогнозирования структуры модели, как альтернатива алгоритму перебора суперпозиций элементарных функций.

Метод

Используется метод структурного обучения. Метод восстанавливает скрытую структуру, заданную на исходных данных.

Текущее состояние области

1. Jaakola T., Sontag D. Learning Bayesian Network Structure using LP Relaxations, 2010.
2. Koza, J. R. Genetic programming, 1998.
3. Г.И. Рудой, В.В. Стрижов. Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных, 2013.
4. Parantapa Goswami. Exploring the Space of IR Functions, 2014.
5. Clinchant Stephane, Gaussier Eric. Information-based Models for Ad Hoc IR, 2010.

Поиск глобальной модели

- коллекция документов $\mathbb{C} = \{d_i\}_{i=1}^{N}$;
- множество запросов к этой коллекции $\mathbb{Q} = \{q_j\}_{j=1}^{|Q|}$;
- покрытие коллекции $\mathbb{C}$ на подколлекции $C_k, \bigcup C_k = \mathbb{C}$;
- $f_k \in \mathcal{G}$ — оптимальная ранжирующая функция на подколлекции C_k для множества запросов $\mathbb{Q}$:

$$f_k : (C_k \times \mathbb{Q}, f_k) \mapsto \{0, 1\};$$

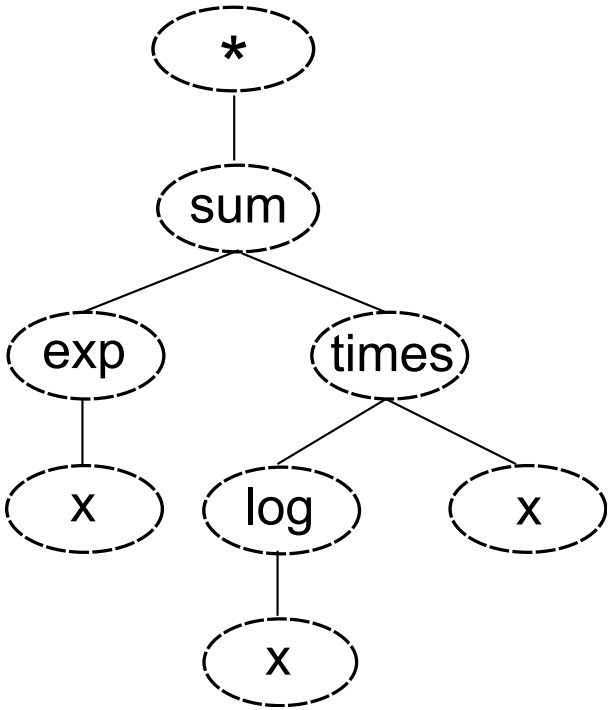
- множество порождающих функций $\mathcal{G}$;
- множество правил порождения суперпозиции $\mathcal{G} = \{g \to B(g, g)|U(g)|S\}$,где $B = \{+, -, *, /\}$, $U = \{sqrt, log, exp\}$, $S = \{x, y\}$

Требуется:

найти глобальную ранжирующую функцию

$$f^* = \arg\max_{f \in \mathcal{G}}(MAP(f, \mathbb{C}, \mathbb{Q})).$$

Правила построения дерева Γ_f суперпозиции f



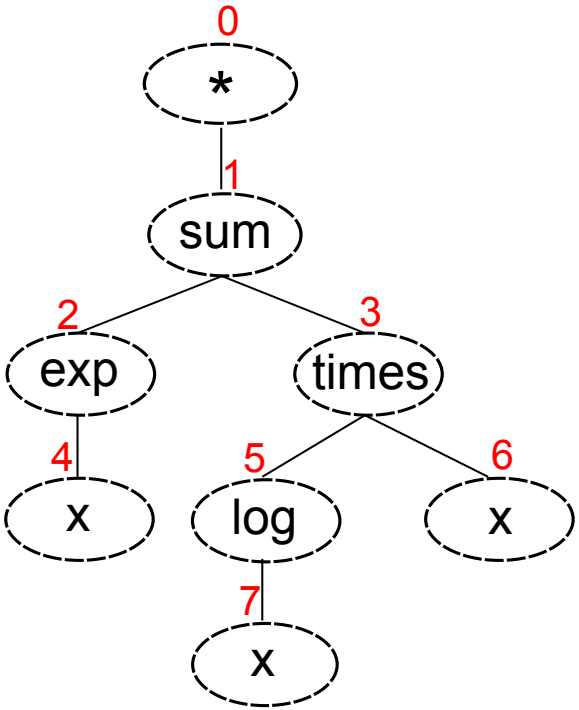
$\mathcal{G} = \{\exp(\cdot), \log(\cdot), \text{sum}(\cdot,\cdot), \text{times}(\cdot,\cdot)\}$.

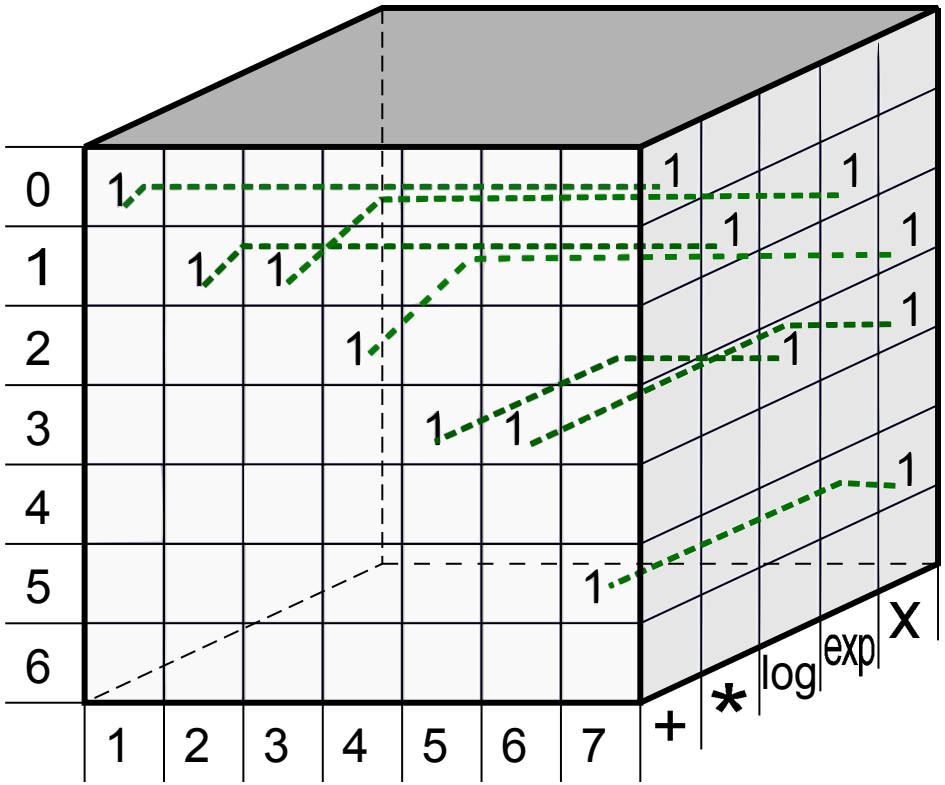
$f = \exp(x) + (\log x)x$

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.

Ограничение на построение матрицы $\mathbf{Z}_f$ суперпозиции f





$$f = \exp(x) + (\log x)x$$

Трехиндексная матрица связей Z_f дерева Γ_f

- вершины дерева пронумерованы;
- первые два индекса — номера вершин в ребре;
- третий индекс — выбранная элементарная функция на конце ребра.

Прогнозирование структуры суперпозиции

Имеется обучающая выборка состоящая из документов C_k, множества запросов к ним $\mathbb{Q}$, и соответствующих ранжирующих функций f_k: $(C_k \times \mathbb{Q}, f_k)$.
Найти алгоритм прогнозирования

$$a : C_s \times \mathbb{Q} \mapsto f_s.$$

Этапы алгоритма прогнозирования:

1. найти матрицу вероятностей P_k;
2. найти $Z_{f_s} = \arg \max_{Z \in \mathcal{M}} \sum_{i,j,k} P_{ijk} \times Z_{ijk}$.

Построение дерева $\hat{\Gamma}_f$

Задана матрица P, состоящая из блока $P'_{s \times s \times l}$:
P'_{ijk} — вероятность того, что на конце ребра (i, j) в дереве находится функция g_k
и блока $P''_{s \times s \times n}$:
P''_{ijk} — вероятность того, что на конце ребра (i, j) в дереве находится переменная x_k.
Назовем вершину i открытой, если ее арность не равна нулю, но у нее нет дочерних вершин.

Требуется:

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

Способы построения:

1. жадный алгоритм с сохранением нескольких лучших вариантов;
2. метод динамического программирования.

Модификация жадного алгоритма для построения дерева $\hat{\Gamma}_f$

Задано K — максимально допустимая сложность суперпозиции.
Задано R — поддерживаемое число оптимальных суперпозиций.

- Объявляем корень дерева открытой вершиной.
- Пока количество единиц в матрице не превышает K, повторяем:
  1. сортируем по убыванию $P^r_{i^r jk}$ для каждой открытой i^r;
  2. выбираем R лучших $P^r_{i^r j_l k_l}, l = 1, \ldots R$;
  3. достраиваем матрицы-кандидаты Z^r_l: $Z^r[i^r, j_l, k_l] = 1$;
  4. добавляем j_l к списку открытых вершин для матрицы Z^r, если $(i^r, j_l, k_l) \in P'$;
- если количество единиц превышает T, используем только независимые переменные: $(j^*, k^*) = \arg\max_k P''_{ijk}$, $(i, j^*, k^*) = 1$ для всех i-открытых.

Метод динамического программирования для построения дерева $\hat{\Gamma}_f$

Дано недостроенное дерево Γ с матрицей Z, и процедура вызывается в одной из его открытых вершин $FindOptTree(j^*, k^*)$;

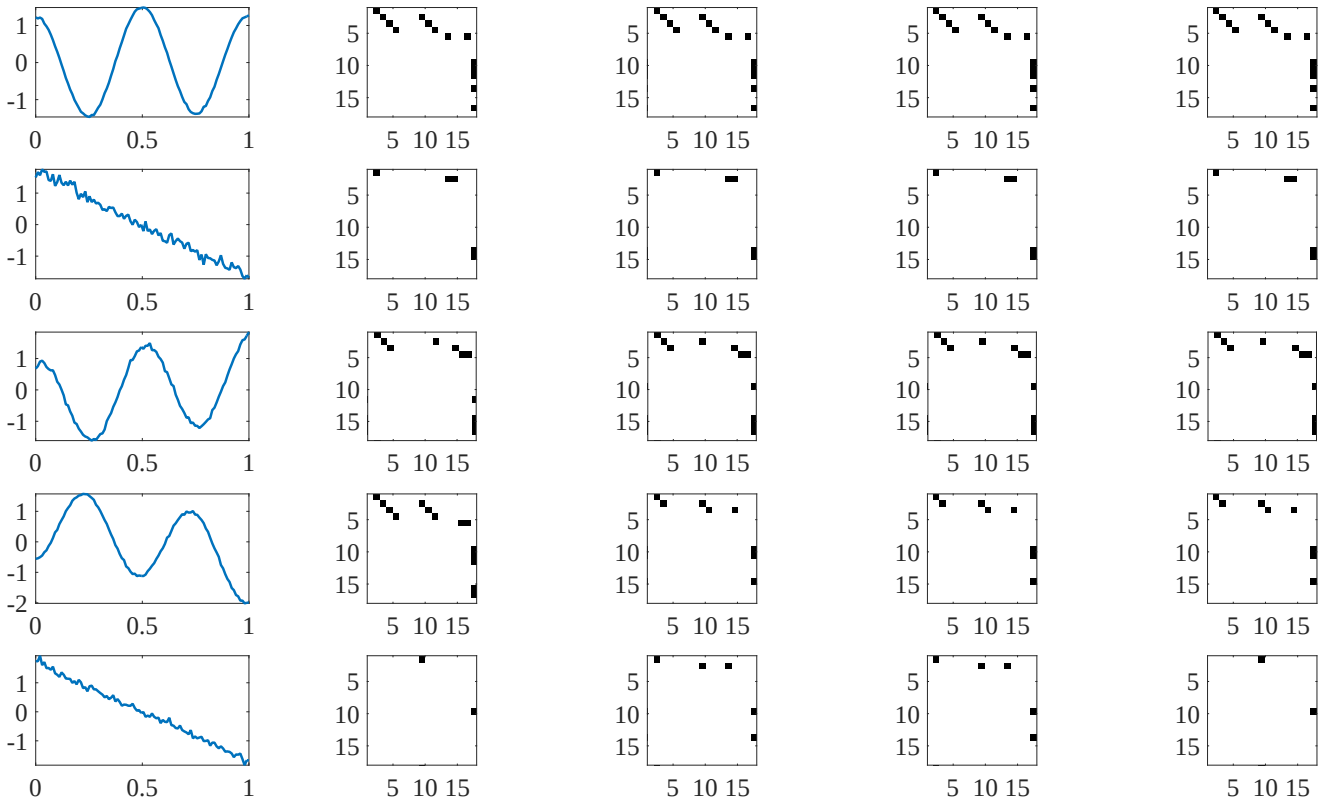
- если k^* — независимая переменная, то процедура завершена, возвращаем $(p_{j^*} = 1, Z = Z)$;
- если k^* — элементарная функция, то:
- для всех возможных переходов из j^* $(j^*, j, k), j = j^* + 1, \ldots, s, k = 1, \ldots, l + n$
  1. считаем $(p_j, Z_j) = FindOptTree(j, k)$;
  2. возвращаем новую вероятность поддерева

$$p_{j^*} = \max_{j=j^*+1,\ldots,s,k=1,\ldots,l+n} P_(j^*, j, k) * p_j$$

  3. достраиваем возвращаемую матрицу $Z = Z_j, \quad Z(j^*, j', k') = 1$, где

$$(j', k') = \arg\max_{j=j^*+1,\ldots,s,k=1,\ldots,l+n} P_(j^*, j, k) * p_j.$$

Тестирование методов прогнозирования Γ_f на синтетических данных



Метод динамического программирования показывает на

Прогнозирование ранжирующей функции на коллекции аннотаций EURO 2010

Цель эксперимента

Зная локальные ранжирующие функции для подколлекций, получить новую ранжирующую функцию для всей коллекции документов, проверить ее качество с помощью функционала MAP.

- 1663 документа в коллекции;
- 24 подколлекции документов;
- 1663 запроса;
- экспертно заданные ранги релевантности;
- множество порождающих функций $\mathcal{G} = \{g \to B(g,g)|U(g)|S\}$,где $B = \{+,-,*,/\}$, $U = \{sqrt, log, exp\}$, $S = \{x, y\}$;
- локальные ранжирующие функции: $\log \frac{x}{y} + \sqrt{y}$, $\sqrt{x} - x$, $\log \frac{x}{y*y}$, $\log \frac{x}{y} + \sqrt{x}$, $\log \sqrt{\sqrt{\frac{x}{y}}}$, $x + y - x^2$, $\log \frac{x}{y} + y$, ...

Прогнозирование ранжирующей функции на коллекции аннотаций EURO 2010

Номер	Символьная запись	MAP
1	$\frac{\log x + \log y}{\exp x}$	0.321
2	$\sqrt{x} - x$	0.308
3	$\sqrt{x} - \exp x + \sqrt{y} - \log(x)$	0.306
4	$\log \frac{x}{y}$	0.303
5	$\log \frac{x}{y*y}$	0.302
6	$x + y - x^2$	0.297
7	$\log \frac{x}{y}$	0.291

Спрогнозированные с помощью структурного обучения функции.

Заключение

- Поставлена и решена задача прогнозирования структуры ранжирующей функции для коллекции документов.
- Предложено описание допустимых суперпозиций, удовлетворяющее необходимым ограничениям.
- Предложены алгоритмы построения допустимой суперпозиции по вероятностной матрице прогноза.
- Разработан алгоритм прогнозирования структуры ранжирующей функции.
- На синтетических данных протестированы предложенные алгоритмы построения суперпозиции по вероятностной матрице прогноза.
- Предлагаемый метод прогнозирования был протестирован на выборке аннотаций к докладам конференции EURO 2010 и показал адекватные результаты.

Публикации

1. *Варфоломеева А.А.* Локальные методы прогнозирования с выбором метрики // Машинное обучение и анализ данных, 2012. Т.1, Вып. 3. Стр. 367–375.
2. *Варфоломеева А.А., Стрижов В.В.* Алгоритм разметки библиографических списков методами структурного обучения // Информационные технологии, 2014, Вып. 7. Стр. 11–15.
3. *Бочкарев А., Варфоломеева А.А.* Структурное обучение при порождении моделей. // Машинное обучение и анализ данных, 2015.