Алгоритм прогнозирования структуры локально-оптимальных моделей.

Михаил Лепехин

МФТИ ФИВТ

lepehin.mn@phystech.edu

25 апреля 2019 г.

Повышение обобщающей способности ранжирующей модели

Цель

Построить метод для предсказания структуры нелинейной ранжирующей функции на основе генетического алгоритма и сравнить полученные результаты с результатами сообщества TREC.

Проблема

Предсказание структуры нелинейной модели по имеющимся данным - вычислительно сложная задача.

Решение

Кластеризация коллекции текстовых документов и построение суперпозиции ранжирующих функций, каждая из которых обучена на своём кластере.

Существующие методы

Перебор суперпозиций

- P. Goswami, S. Moura, E. Gaussier, M.-R. Amini, F. Maes Exploring the space of ir functions // ECIR'14, 2014, pp. 372–384.

Использование генетического алгоритма

- Fan, Weiguo and Gordon, Michael D. and Pathak, Praveen Personalization of Search Engine Services for Effective Retrieval and Knowledge Management // In Proceedings of the twenty first international conference on Information Retrieval
- A.S. Kulunchakov, V.V. Strijov Study of image retrieval and classification based on adaptive features using genetic algorithm feature selection, Expert Systems with Applications: An International Journal (2017).

Постановка задачи порождения модели

Дано

Коллекция текстовых документов $\mathbf{C}$, состоящая из документов $\{d_i\}_{i=1}^{|\mathbf{C}|}$ и множество поисковых запросов $\mathbf{Q} = \{q_j\}_{j=1}^{|\mathbf{Q}|}$.

Часть документов оценена экспертами. Таким образом задана функция $r(d, q) \to \{0, 1\}$, где оценка 1 ставится в случае релевантности документа d запросу q.

Обозначения

$\mathrm{df}(w, \mathbf{C})$ – число документов $d \in \mathbf{C}$, в которые входит слово w,

$\mathrm{freq}(w, d)$ – число вхождений слова w в документ d,

l_{avg} – среднее число слов в документах коллекции,

$|d|$ – число слов в документе d.

Постановка задачи порождения модели

Рассматриваемые характеристики

$$\text{idf}(w, \mathbf{C}) = \frac{\text{df}(w, \mathbf{C})}{|\mathbf{C}|}$$

$$\text{tf}(w, d, \mathbf{C}) = \text{freq}(w, d) * log\left(1 + \frac{l_{avg}}{|d|}\right)$$

Пусть $f : \mathbb{R}^2 \to \mathbb{R}$ - функция 2 переменных. Тогда её значение на паре (d, q) определяется как сумма её значений на парах (d, w), где $w \in q$ - слово из запроса:

$$f(d, q) := \sum_{w \in q} f(\text{tf}(w, d), \text{idf}(w))$$

Постановка задачи порождения модели

Метрика качества ранжирующей функции

$$\mathrm{MAP}(f, C, \mathbf{Q}) = \frac{1}{|\mathbf{Q}|} \sum_{q \in \mathbf{Q}} \mathrm{AvgP}(f, q),$$

где

$$\mathrm{AvgP}(f, q) = \frac{\sum_{k=1}^{|C_q|} \mathrm{Prec}(k) \times r(q, k)}{\sum_{k=1}^{|C_q|} r(q, k)},$$

$$\mathrm{Prec}(k) = \frac{\sum_{s=1}^{k} r(q, s)}{k}$$

Постановка задачи порождения модели

Пространство исследуемых функций

В качестве математических примитивов $h(x, y)$ будем использовать функции $\sqrt{x}, x + y, x - y, x * y, x/y, log\ x, e^x$. Будем исследовать пространство всех суперпозиций этих примитивов. Обозначим его $\mathcal{F}$.

Оптимизируемая функция

$$f^* = \arg\max_{f \in \mathcal{F}} \mathrm{MAP}(f|\mathbf{C}, \mathbf{Q}) - R(f),$$

где R - регуляризатор, штрафующий за структурную сложность порождаемой суперпозиции.

Постановка задачи на кластерах документов

Разбиение на кластеры

Обозначим L множество всех рассматриваемых слов в документах, $|L| = n$.

Пусть отображение $V : C \to \mathbb{R}^n$ каждому документу из коллекции C сопоставляет вектор длины n. После применения алгоритма k-means к полученным векторам, образуется множество кластеров $D, |D| = m$.

Для каждого кластера при помощи генетического алгоритма построим семейство ранжирующих функций $F^*_{d_i} = \{f_i^1, \ldots, f_i^n\}$. В каждом семействе i выделим наилучшую по описанной выше метрике ранжирующую функцию $f_i^* \in F_{d_i}$.

Постановка задачи на кластерах документов

Метрика качества на кластерах

Определим

$$f^* = \arg\max_{W \in \mathbb{R}^m} \left(\left(MAP\left(\sum_{i=1}^{m} W_i f_i^* | \mathbf{C}, \mathbf{Q} \right) \right) - \sum_{i=1}^{m} R(f_i^*) \right)$$

Веса W_i находятся при помощи линейной регрессии.

Метод решения

Базовый алгоритм

Используется генетический алгоритм со следующими процедурами:

- мутация — замена произвольной вершины на заново сгенерированную.
- скрещивание (crossover) — обмен местами двух произвольных вершин деревьев.

Регуляризация

$$R(f) = |f|^2,$$

где $|f|$ - число вершин в дереве функции f.

Цель эксперимента

Цель эксперимента

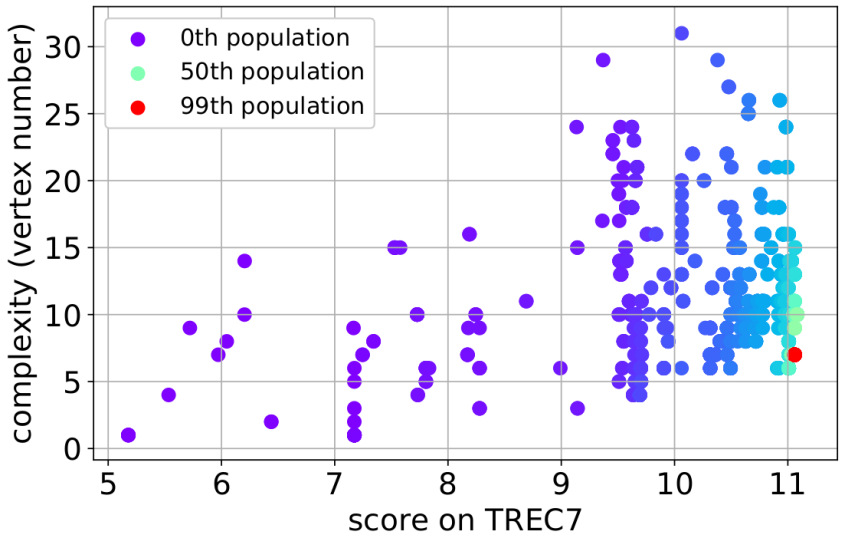
Проверить работоспособность метода. Улучшить результаты по сравнению с работами сообщества TREC.

Данные

Коллекция TREC (датасеты 5-8).
https://trec.nist.gov/data.html

Результаты базового эксперимента

Зависимость сложности модели от значения целевого критерия.



Результаты базового эксперимента

Результаты при сравнении на корпусах TREC-5, TREC-6, TREC-7.

Superposition	TREC-5	TREC-6	TREC-7
Функции сообщества			
f_1	8.785	13.715	10.038
f_2	8.908	13.615	9.905
f_3	8.908	13.615	9.905
Найденные наилучшие функции			
h_5^*	**9.537**	13.762	10.584
h_6^*	8.903	**13.967**	10.771
h_7^*	8.526	13.424	**11.060**

Решение с использованием кластеризации

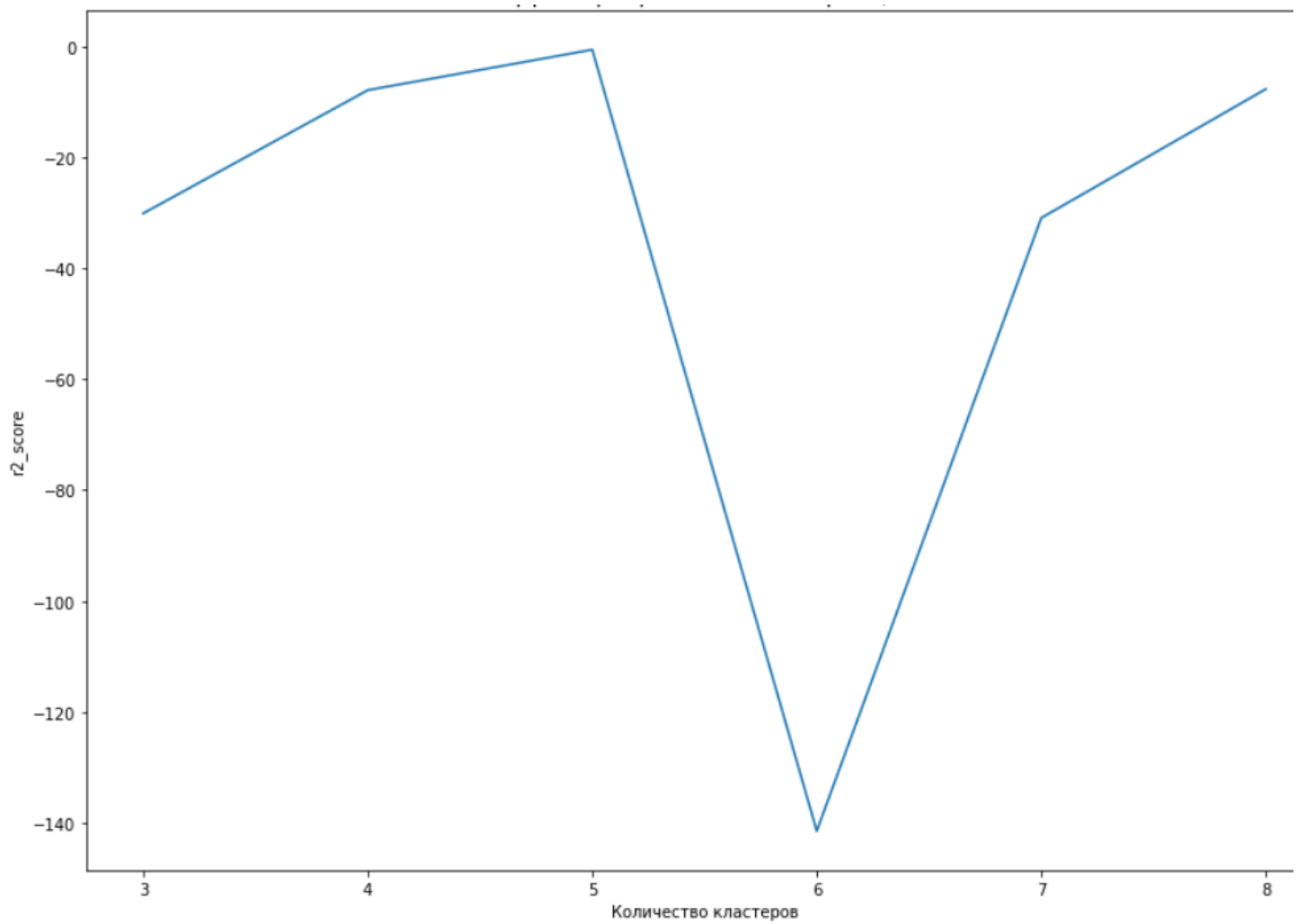
Преобразование текстовых документов в векторы

1) One Hot Encoding
2) Doc2vec
3) Latent semantic analysis

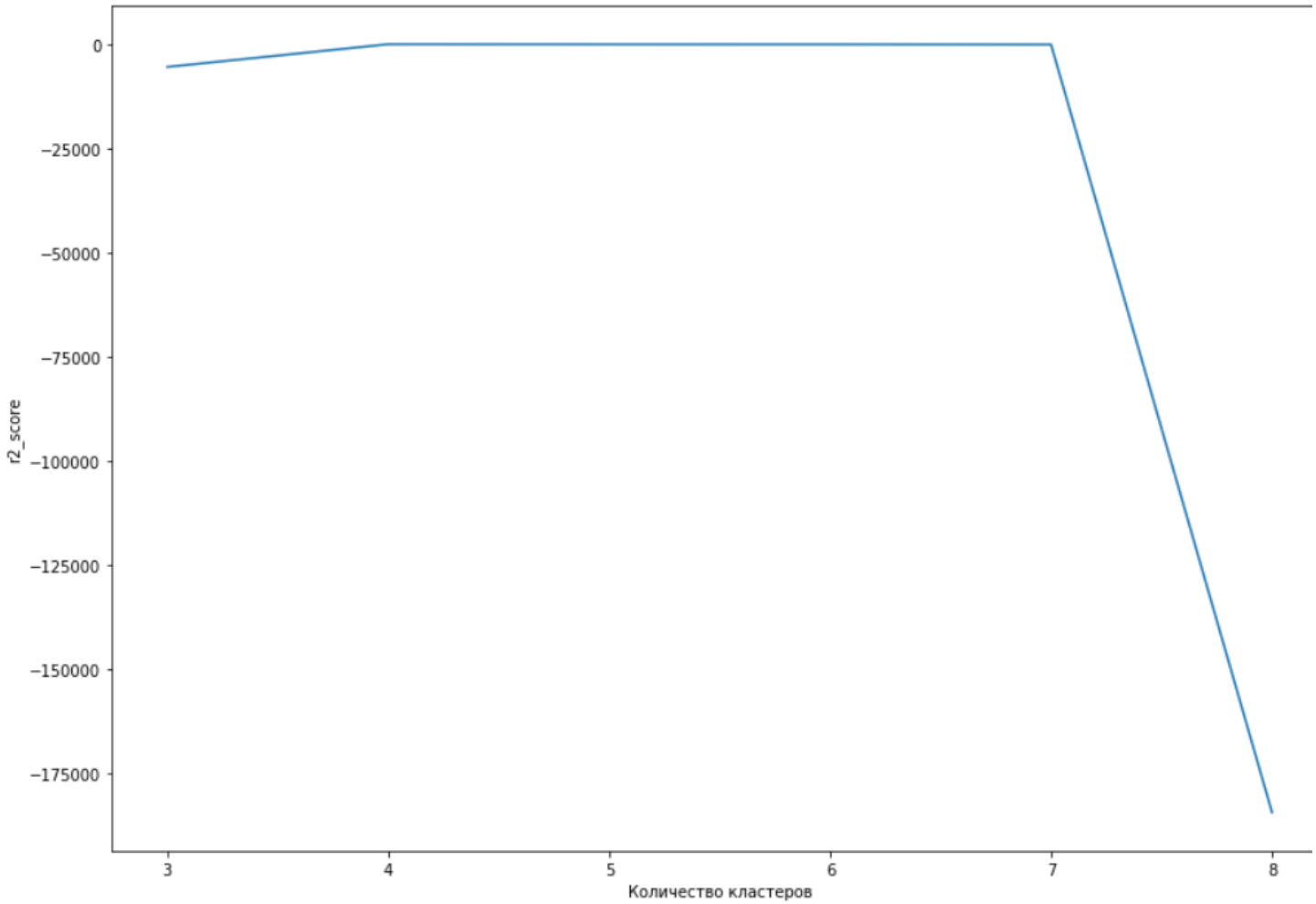
Метод построения суперпозиции моделей

Строится таблица X размера $m \times m$ следующего вида. В i строке, j столбце записывается значение MAP ранжирующей фунции f_j^* на i-м кластере. В качестве вектора $y \in \mathbb{R}^m$ берётся вектор $(1, ..., 1)^\mathsf{T}$, состоящий из одних единиц. Далее по полученным X и y обучается линейная регрессия.

Результаты эксперимента 1



Результаты эксперимента 2



Заключение

- Показана работоспособность метода
- Для каждого корпуса была получена функция наилучшим образом ранжирующая документы для данного запроса
- Показано, что при оптимальном выборе числа кластеров значение критерия качества резко повышается.

Планируется

- Протестировать различные методы построения суперпозиции по кластерам на данных из TREC.
- Обосновать использование кластеризации: проверить гипотезу о липшицевости отображения из метрического пространства кластеров документов в пространство моделей.