Декомпозиция смеси распределений на основе эмпирических данных в задачах текстовой кластеризации

Митяшов А. А.

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

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

Москва,
2016 г.

Цели исследования

Цель

Разделить базу государственных закупок на наборы закупок, соответсвующих одном товару; оценить типичность государственных закупок и их превышение относительно нормальной рыночной цены.

Проблема

Закупки состоят из нестандартизованного текстового описания и набора дополнительных атрибутов. Зачастую данные заполенены некорректно.

Предлагаемый метод решения

Разделить базу государственных закупок на набор кластеров. Близкие госзакупки, представляющие одну форму выпуска одного товара, должны принадлежать одному кластеру.

База государственных закупок

Признаки закупки $\mathbf{x} \in \mathcal{D}$

x^{text}	текстовое описание закупки
x^{meas}	синтетический признак «измерения закупки»
x^{price}	цена закупки
x^{count}	количество товара в закупке
x^{date}	дата совершения закупки
x^{reg}	регион совершения закупки
x^{unit}	единица товара закупки
x^{rcpea}	код ОКПД закупки

Оценка типичности госзакупок

Требуется:

- оценить типичность госзакупки $\mathbf{x} \in \mathcal{D}$;
- оценить рыночную цену товара, представленного в закупке;
- оценить превышение рыночной цены.

Предлагается модель:

$$typicality(\mathbf{x}) = \begin{cases} 1, & \text{если } x^{\text{price}} \leq \text{MarketPrice}; \\ 0, & \text{иначе}, \end{cases}$$

$$excess(\mathbf{x}) = (x^{\text{price}} - \text{MarketPrice})_+.$$

Для получения оценок рыночной цены товара предлагается кластеризовать базу государственных закупок.

Постановка задачи

Требуется кластеризовать базу закупок:

$$a : \mathcal{D} \to \mathcal{C}, \qquad \mathcal{C} = \{C_1, \ldots, C_K\}.$$

Предполагается, что цены госзакупок из одного кластера имеют гамма-распределение:

$$p(x^{\text{price}}) = \Gamma(k_C, \lambda_C), \qquad a(\mathbf{x}) = C.$$

Тогда цены всех госзакупок образуют следующую смесь гамма-распределений:

$$p(x^{\text{price}}) = \sum_{C \in \mathcal{C}} [a(\mathbf{x}) = C] \Gamma(k_C, \lambda_C).$$

Экспертная оценка качества кластеризации

Экспертное оценка качества кластера:

$$g(C) = \left[\frac{\sigma_C}{\mu_C} \leq \theta^{\text{price}}\right] [|C| \geq \text{MinPts}] \left[\max_{\substack{\mathbf{x}_1,\mathbf{x}_2: \\ a(\mathbf{x}_1)=a(\mathbf{x}_2)=C}} dist(\mathbf{x}_1, \mathbf{x}_2) \leq \theta^{\text{dist}}\right],$$

где μ_C, σ_C - средняя цена и стандартное отклонение цены в C, $|C| = |\mathbf{x} : a(\mathbf{x}) = C|$, $\text{MarketPrice}(C) = \mu_C + \sigma_C$ - рыночная цена в C.

Задача кластеризации базы закупок как задача экспертно-интерпертуемой декомпозиции смеси распределений:

$$\begin{cases} \frac{\sum_{C \in \mathcal{C}} g(C)}{|\mathcal{C}|} \to \max\limits_{a:\mathcal{D}\to\mathcal{C}}, \\ L(\mathcal{D}, \mathcal{C}) = -\frac{1}{|\mathcal{D}|} \sum\limits_{\mathbf{x}\in\mathcal{D}} \ln \sum\limits_{C\in\mathcal{C}} [a(\mathbf{x}) = C]\Gamma(k_C, \lambda_C)|_{\mathbf{x}} \to \min\limits_{a:\mathcal{D}\to\mathcal{C}}. \end{cases}$$

Алгоритм решения поставленной задачи

Для уменьшения объемов выборки предлагается использовать тематическое моделирование с аддитивной регуляризацией. В результате получен набор тем $\mathcal{T} = \{T_1, \ldots, T_N\}$. Каждая полученная тема T_i соответствует одной подвыборке. Далее каждая полученная подвыборка кластеризуется с использованием иерархической модификации алгоритма DBScan.

Тематическое моделирование

Для уменьшения объема выборки предлагается использовать тематическое моделирование с аддитивной регуляризацией. Предлагается использовать следующие регуляризаторы

- разреживание матрицы «слова-темы»,
- разреживание матрицы «темы-документы (закупки)»,
- сокращение числа тем,

чтобы добиться того, что каждая закупка соответствовала всего одной теме. Тогда данная закупка будет относится к подвыборке, соответствующей данной теме.

Функция расстояния между двумя закупками

- вложенность тектового описания одной закупки в другую:

$$\mathrm{sim}_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{|\mathrm{set}(x_1^{\mathrm{text}}) \cap \mathrm{set}(x_2^{\mathrm{text}})|}{\min(|\mathrm{set}(x_1^{\mathrm{text}})|, |\mathrm{set}(x_2^{\mathrm{text}})|)};$$

- близость по мере Жаккара измерений закупок:

$$\mathrm{sim}_2(\mathbf{x}_1, \mathbf{x}_2) = \frac{|\mathrm{set}(x_1^{\mathrm{meas}}) \cap \mathrm{set}(x_2^{\mathrm{meas}})|}{|\mathrm{set}(x_1^{\mathrm{meas}}) \cup \mathrm{set}(x_2^{\mathrm{meas}})|};$$

- близость цен закупок:

$$\mathrm{sim}_3(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{|x_1^{\mathrm{price}} - x_2^{\mathrm{price}}|}{x_1^{\mathrm{price}} + x_2^{\mathrm{price}}}.$$

Расстояние между двумя закупками:

$$\mathrm{dist}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \sum_{i=1}^{3} \omega_i \mathrm{sim}_i(\mathbf{x}_1, \mathbf{x}_2), \qquad \sum_{i=1}^{3} \omega_i = 1.$$

Иерархическая кластеризация на основе DBScan

Для решения поставленной задачи предлагается использовать иерахическую модификацию алгоритма DBScan.
Предлагается на первой итерации (нулевом уровне иерархии) кластеризовать всю данную выборку. Затем, изменяя параметры алгоритма и веса признаков в метрике, предалагается на следующем уровне иерархии кластеры, полученные на предыдущем уровне иерархии разбивать на новые кластеры.

Иерархическая модификация DBScan

Вход: $\mathcal{X}$ - выборка, MinPts - параметр алгоритма DBScan, θ^{lvl} - максимальный уровень иерархии;

Выход: $\mathcal{C}$ - разбиение выборки $\mathcal{X}$ на кластеры;

$\mathcal{C} := \varnothing$

LOC := $[\{\mathcal{X}, 0\}]$;

пока LOC $\neq \emptyset$

  $\{D, \text{level}\} := \text{LOC.pop}$;

  $\text{dist} = \text{dist}(\omega_1(\text{level}), \omega_2(\text{level}), \omega_3(\text{level}))$ – меняем функцию расстояния;

  $\varepsilon := \text{epsilonSelection}(\text{MinPts}, D, \text{level})$;

  $\tilde{\mathcal{C}} := \text{DBScan}(D, \varepsilon, \text{MinPts})$

  если level $< \theta^{\text{lvl}}$ **то**

    для всех $C \in \tilde{\mathcal{C}}$

      $\text{LOC.append}(\{C, \text{level} + 1\})$

  иначе

    для всех $C \in \tilde{\mathcal{C}}$

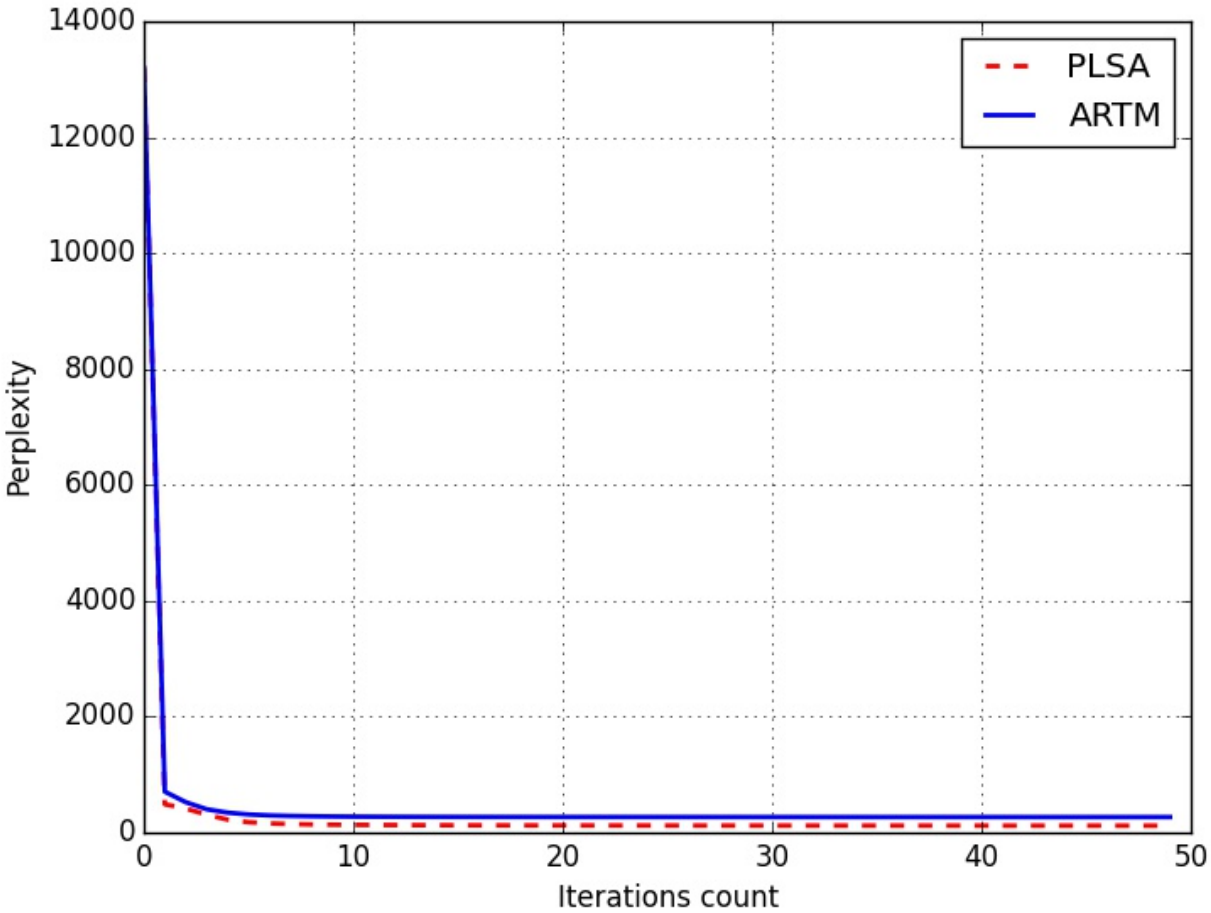
      $\mathcal{C}.\text{append}(C)$

Вычислительный эксперимент

Для вычислительного эксперимента использовались закупки из следующих ОКПД:

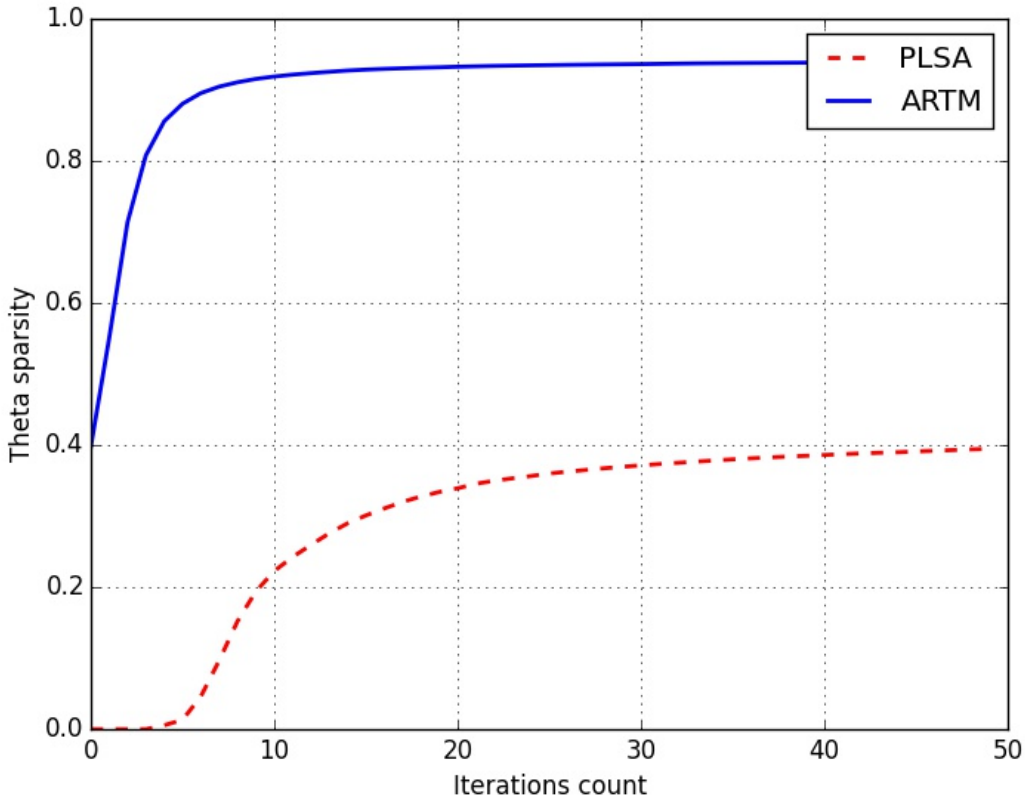
- 33.10.15.131: Инструменты режущие и ударные с острой (режущей) кромкой однолезвийные;
- 21.20.10.239: Препараты для лечения нервной системы;
- 24.42.13.779: Средства противораковые;
- 24.42.13.796: Средства противовирусные;
- 36.63.21.110: Ручки шариковые;
- 33.10.15.121: Шприцы-инъекторы медицинские с инъекционными иглами и без них;
- 30.01.24.110: Части и принадлежности копировально-множительных машин.
  Размер выборки - более 400 тыс. закупок, количество уникальных слов, встречающихся не менее 5 раз - 8 тыс.

Сравнение ARTM и PLSA



Сравнение значения перплексии для PLSA и ARTM. Из графика видно, что ухудшение перплексии при использовании ARTM незначительно.

Сравнение ARTM и PLSA



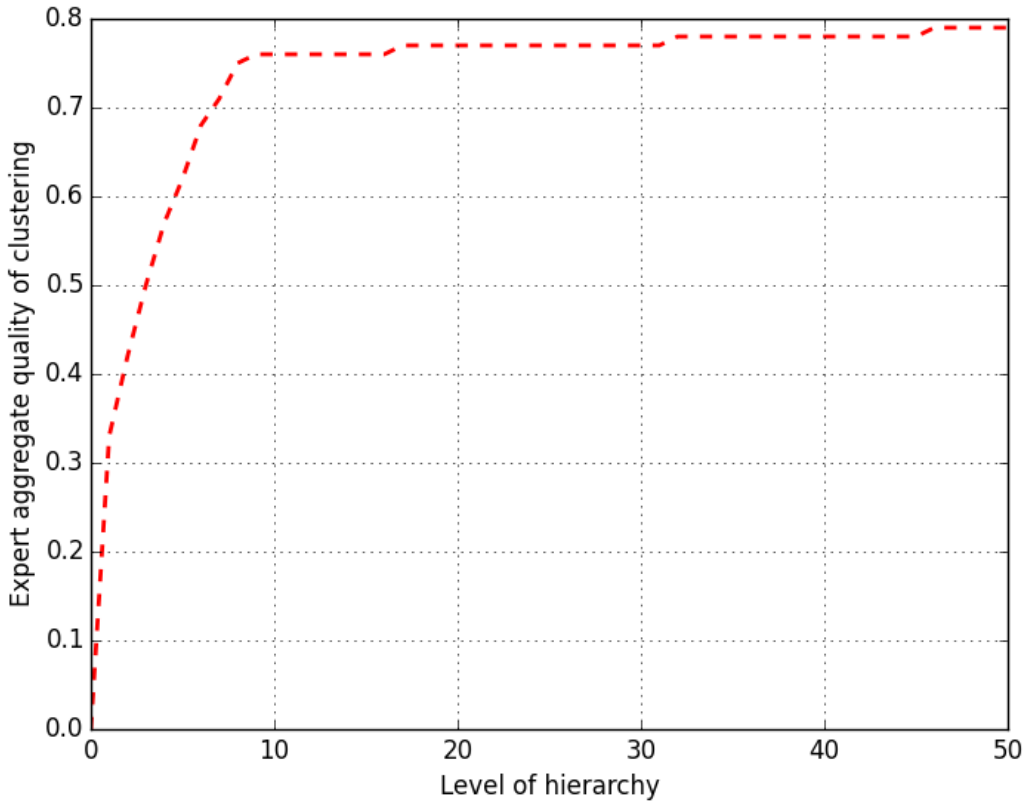
Сравнение разреженности матрицы «темы-документы» для PLSA и ARTM с указанными ранее регуляризаторами. Как видно из графика, разреженности матрицы «темы-документы» недостаточно для однозначного отнесения закупки к теме.

Сравнение алгоритмов кластеризации

Алгоритм	Отрицательное лог-правдоподобие	Экспертное качество кластеров
Hierarchical DBScan	1,14	0,79
DBScan	2,11	0,41
K-Means	1,52	0,75
Spectral Clustering	1,27	0,82
Affinity Propagation	1,87	0,61
Agglomerative Clustering	1,09	0,49
EM-алгоритм	0,93	0,21
Topic Modeling	5,27	0

Также в ходе вычислительного эксперимента была принята гипотеза о гамма-распределении цен в кластере.

Оценка качества кластеризации



На графике показана зависимость экспертного агрегированного качества кластеров от максимально допустимого уровня иерархии. Как видно из графика, 10 уровней достаточно.

Заключение

В ходе данного исследования были разработаны

- алгоритм кластеризации базы государственных закупок;
- способ оценки рыночной цены товара;
- способ оценки типичности государственных закупок и их превышения рыночной цены.

Результаты данной работы использованы при разработке системы анализа государственных закупок «Антирутина-44».