УДК 519.256

Выбор признаков и оптимизация метрики при кластеризации коллекции документов[1]

А. А. Адуенко, А. А. Кузьмин, В. В. Стрижов

Аннотация. Исследуется проблема верификации корректности тематической классификации документов с помощью метрического алгоритма. Предложен алгоритм выбора оптимальной функции расстояния между документами. Исследуется соответствие между полученной кластеризацией документов и их экспертной классификацией. Результаты кластеризации и их соответствие экспертной тематической классификации проиллюстрированы вычислительным экспериментом на реальной коллекции документов.

Ключевые слова: метрическая классификация, метод ближайших соседей, функция близости, выбор признаков, гипотеза компактности.

Введение

Перед программным комитетом конференции с большим числом участников встает задача о корректности отнесения тезисов к заданному набору тем. Для оценки корректности предлагается кластеризовать набор тем, используя принцип ближайшего соседа. Предлагаемый подход к построению тематических моделей называется «жёстким», так как тезис соответствует только одной теме, что согласуется с правилами подготовки конференций. Задача кластеризации документов относится к задачам поиска скрытой неструктурированной информации. Из-за большого числа экспертов и субъективности восприятия ими темы документа оценить качество кластеризации сложно. Требуется сопоставить результаты экспертной и модельной классификации документов.

[1]Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.

На множестве документов предлагается ввести функцию расстояния [1]. Пусть есть некоторое множество слов, каждое из которых хотя бы раз встретилось в одном из документов коллекции. Назовем это множество словарем. В данной работе под документом будем понимать неупорядоченное множество слов из словаря; слова в документе могут повторяться. Документ представлен в виде «мешка слов» [2]. Каждому документу поставим в соответствие вектор, содержащий информацию о словарном составе документа. Размерность этого вектора равна количеству слов в словаре. Расстояние между документами есть расстояние между векторами, соответствующими этим документам.

При составлении словаря выполняется предобработка документов. Слова приводятся к начальной лексической форме (лемматизация), удаляются знаки препинания. Исключаются слова, встречающиеся малое количество раз, а также слова, встречающиеся в большинстве документов (стоп-слова); для этого используется критерий $tf \cdot idf$ (англ. tf — term frequency, idf — inverse document frequency) [3], а также словарь стоп-слов.

Кластеризация документов производится как с помощью метрических алгоритмов кластеризации, например, k-means [4, 5], FOREL [6], C-means [7], STOLP [8], FRiS - STOLP [9], BoostML, DANN [10] и другие [11, 12], так и с помощью вероятностных методов [13], например, с помощью вероятностного латентного семантического анализа (англ. PLSA — Probabilistic Latent Semantic Analysis) [2], или латентного размещения Дирихле (англ. LDA — Latent Dirichlet Allocation) [1]. В данной работе ставится задача метрической кластеризации. Метрические алгоритмы применяются в задачах классификации, когда свойства исследуемых объектов удовлетворяют гипотезе компактности: похожие объекты, как правило, лежат в одном классе.

Центральной задачей кластеризации является задача выбора информативных признаков, которые используются при вычислении расстояний. В этой работе для выбора используется взвешенная метрика — такая метрика, в которой каждому признаку приписывается некоторый вес, пропорциональный степени важности признака в задаче классификации. Путем построения метрического классификатора решается задача выбора оптимальных параметров алгоритма кластеризации: функции близости, включающей взвешенный набор признаков.

В работе также рассматривается проблема оптимизации весов признаков метрики, обеспечивающей наибольшее количество верно классифицирован-

ных элементов контрольной выборки. Для определения весов признаков в метрике вводится функционал качества метрики и находится его максимум. Предложенный метрический алгоритм с выбором признаков иллюстрируется примером тематической классификации тезисов конференции «European Conference on Operational Research, EURO–2012».

1 Постановка задачи

Пусть $W = \{w_1, \ldots, w_n\}$ — заданное множество слов, словарь. Документом d назовем неупорядоченное множество слов из W: $d = \{w_j\}$, где $w_j \in W$ — слово в документе. Пусть $D = \{d\}$ — множество документов, коллекция, а K — заданное число кластеров, на которое требуется разбить множество D. Требуется задать функцию расстояния на множестве документов:

$$\rho(d_i, d_j) : W \times W \longrightarrow \mathbb{R}_+$$

и кластеризовать коллекцию D, максимизируя функцию качества $Q(\rho)$, заданную в (5).

Представим документ с номером i в виде вектора

$$\mathbf{x}_i = [c(d_i, w_1), \ldots, c(d_i, w_j), \ldots, c(d_i, w_n)]^\mathsf{T}, \tag{1}$$

где $c(d_i, w_j)$ — число вхождений слова w_j в документ d_i. Опишем множество документов и их тематику следующим образом. Задана выборка — множество m пар

$$\{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}, \tag{2}$$

состоящих из вектора признаков (документа) $\mathbf{x} \in \mathbb{R}^n$ и метки класса (темы документа) y, принимающего значения из множества $\{1, 2, \ldots, K\}$, где K — общее число тем.

Введем на множестве объектов взвешенную метрику ρ и найдём параметры, при которых она лучше всего подходит для сравнения объектов. Расстояние же между документами вводим как расстояние между векторами, им соответствующими:

$$\rho(d_i, d_j) = \rho(\mathbf{w})(\mathbf{x}, \mathbf{x}').$$

Рассмотрим набор взвешенных метрик Минковского с фиксированным параметром $\mu \geq 1$:

$$\rho(\mathbf{x}, \mathbf{x}') = \sqrt[\mu]{\sum_{j=1}^{n} w_j |x_j - x'_j|^\mu}, \quad \text{где} \sum_{j=1}^{n} w_j = 1, \quad w_1, \ldots, w_n \geq 0. \tag{3}$$

Функция расстояния ρ является метрикой:

1) $\rho(\mathbf{x}, \mathbf{x}') = 0 \Leftrightarrow \mathbf{x} = \mathbf{x}'$, т. к. все разности вида $x_k - x'_k$ равны 0;

2) $\rho(\mathbf{x}, \mathbf{x}') = \rho(\mathbf{x}', \mathbf{x})$, т. к. $|x_i - x'_i| = |x'_i - x_i|$;

3) $\rho(\mathbf{x}, \mathbf{x}') \leqslant \rho(\mathbf{x}, \mathbf{x}'') + \rho(\mathbf{x}'', \mathbf{x}')$ для произвольного $\mu \geq 1$ см. [14].

В качестве функции расстояния $\rho(\mathbf{x}, \mathbf{x}')$ между векторами $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$ можно рассмотреть разные частные случаи (3), в частности расстояние городских кварталов при $\mu = 1$, евклидово расстояние при $\mu = 2$ или расстояние Чебышёва $\rho(\mathbf{x}, \mathbf{x}') = \max\limits_{j=1,\ldots,n} |x_j - x'_j|$ при $\mu = \infty$.

Чтобы оценить веса признаков w_j, введем функцию благонадежности метрики $Q(\rho)$. Введём вспомогательную функцию близости $\delta(\mathbf{x}_i)$ данного объекта обучающей выборки $\mathbf{x}_i$, где $i \in \mathcal{I}$, к объектам своего класса:

$$\delta(\mathbf{x}_i) = \frac{\overline{r}(\mathbf{x}_i) - r(\mathbf{x}_i)}{\overline{r}(\mathbf{x}_i) + r(\mathbf{x}_i)}, \tag{4}$$

где $r(\mathbf{x}_i)$ — расстояние от $\mathbf{x}_i$ до k ближайших объектов того же класса, $\overline{r}(\mathbf{x}_i)$ — расстояние от $\mathbf{x}_i$ до k ближайших объектов из других классов. Определим эти расстояния.

Для произвольного i зададим разбиение множества индексов $\mathcal{I}$ выборки $\mathcal{I} = \mathcal{P} \sqcup \mathcal{N}$ следующим образом:

$$\mathcal{P}(j) = \{j \in \mathcal{I} | y_j = y_i\},$$

$$\mathcal{N}(i) = \{j \in \mathcal{I} | y_j \neq y_i\},$$

то есть $\mathcal{P}(i)$ — индексы объектов того же класса, что и $\mathbf{x}_i$, а $\mathcal{N}(i)$ — индексы объектов из других классов.

Определим расстояние до k ближайших соседей как

$$r(\mathbf{x}_i) = \sum_{p=1}^{k} \rho_p(\mathbf{x}_i, \mathbf{x}_j), \quad \text{где} \quad j \in \mathcal{P}(i),$$
$$\overline{r}(\mathbf{x}_i) = \sum_{\overline{p}=1}^{k} \rho_{\overline{p}}(\mathbf{x}_i, \mathbf{x}_j), \quad \text{где} \quad j \in \mathcal{N}(i),$$

а $p = p(j)$ и $\overline{p} = \overline{p}(j)$ — индексы сортировки по возрастанию элементов множеств

$$\{\rho_p(\mathbf{x}_i, \mathbf{x}_j) | j \in \mathcal{P}(i)\} \quad \text{и} \quad \{\rho_{\overline{p}}(\mathbf{x}_i, \mathbf{x}_j) | j \in \mathcal{N}(i)\}$$

соответственно.

Функция $\delta(\mathbf{x}_i)$ обладает следующим свойством:

$$\delta(\mathbf{x}_i) \approx \begin{cases} -1, & \text{объект } \mathbf{x}_i \text{ близок к объектам чужого класса;} \\ 0, & \text{объект } \mathbf{x}_i \text{ пограничный;} \\ +1, & \text{объект } \mathbf{x}_i \text{ близок к объектам своего класса.} \end{cases}$$

В предположении гипотезы компактности наилучшей (то есть такой, при которой все объекты лежат близко к объектам своего класса) метрикой будет та, у которой $d(\mathbf{x}_i) \approx 1$.

Функция благонадежности $Q(\rho)$ метрики ρ имеет вид

$$Q(\rho) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \delta(\mathbf{x}_i), \tag{5}$$

где ρ и $\delta(\mathbf{x}_i)$ определены в (3) и (4). Задача выбора оптимальной метрики сводится к нахождению максимума функции благонадежности $Q(\rho)$ на множестве $\mathcal{M}$ взвешенных метрик вида (3):

$$\widehat{\rho} = \arg\max_{\rho \in \mathcal{M}} Q(\rho).$$

2 Описание алгоритма выбора метрики

Множество метрик $\mathcal{M}$ задано различными наборами признаков — элементов вектора $\mathbf{x}$, входящих в метрики и их весовыми коэффициентами $\mathbf{w}$. Обозначим взвешенную метрику

$$\rho_{\mathcal{A}}(\mathbf{w}) \;=\; \rho(\mathbf{w}_{\mathcal{A}})(\mathbf{x}_{\mathcal{A}}, \mathbf{x}'_{\mathcal{A}}),$$

где $\mathcal{A}$ — набор индексов признаков, входящих в вектор $\mathbf{x}$, а набор параметров $\mathbf{w}$ метрики $\rho_{\mathcal{A}}$ есть набор весовых коэффициентов слагаемых с индексами j, входящих в линейную комбинацию (3).

Алгоритм основан на последовательном добавлении признаков. На каждом шаге алгоритма в набор $\mathcal{A}$ добавляется j-й признак с весом w_j. Индекс признака и его вес определяются из условия локальной максимальности функции благонадежности $Q(\rho)$. Рассмотрим множество индексов признаков $\mathcal{J} = \{1, 2, \ldots, n\}$ и набор $\mathcal{A} \subseteq \mathcal{J}$. На первом шаге алгоритма $\mathcal{A} = \varnothing$. Итеративно повторяются два шага.

1. Найдем такой признак $\widehat{j}$, который доставляет функции благонадежности $Q(\rho)$ максимум:

$$\widehat{j} = \arg\max_{j \in \mathcal{J} \setminus \mathcal{A}} Q\left(\rho_{\mathcal{A} \cup \{j\}}\left(\widehat{\mathbf{w}}\right) | D_{\mathcal{L}}\right), \quad \text{где} \tag{6}$$

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}_+^{|\mathcal{A}|+1}} Q\left(\rho_{\mathcal{A} \cup \{j\}}\left(\mathbf{w}\right) | D_{\mathcal{T}}\right), \quad \text{при условии} \quad \|\mathbf{w}\|_1 = 1. \tag{7}$$

   Функция $Q(\rho)$ определена в (4). Обозначение $Q(\rho|D_{\mathcal{L}})$ означает, что при решении задачи используется только те элементы выборки D, индексы которой принадлежат множеству $\mathcal{L}$. При этом множество индексов $\mathcal{I}$ всей выборки D разбивается на обучающее $\mathcal{L}$ и контрольной $\mathcal{T}$ множество, $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$,случайным образом на каждом шаге алгоритма.

2. Добавим найденный индекс $\widehat{j}$ в множество $\mathcal{A}$ и повторим предыдущий пункт.

Алгоритм повторяется до уменьшения значения функции качества $Q(\rho|D_{\mathcal{L}})$ на контрольной выборке. Задача (6) решается перебором аргумента за $|\mathcal{J} \setminus \mathcal{A}|$ шагов, а задача (7) решается методом градиентного спуска с ограничениями; её решение описано в [15].

Так как на каждой итерации вышеописанного алгоритма в силу изменения вектора весов $\mathbf{w}$, метрики $\rho_{\mathcal{A}}(\mathbf{w})$, происходит изменение не только значения функции благонадежности Q, но и векторов $\mathbf{x}$ на которых определяются средние расстояния r и $\bar{r}$, то после каждого шага минимизации требуется пересчитывать все функции $\delta(\mathbf{x}_i)$.

Для того, чтобы существенно снизить сложность вычисления функции благонадежности Q, рассмотрим следующий метод, использующий взвешенную метрику ρ при $\mu = 1$ в (3). В качестве функционала качества метрики

введем среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{i<j}[y_i = y_j]} \to \min, \tag{8}$$

которое с учетом параметра μ метрики приобретает вид

$$F_0 = \frac{\sum_{i<j}[y_i = y_j] \cdot \left(\mathbf{v}^\intercal |\mathbf{x}_i - \mathbf{x}_j|\right)}{\sum_{i<j}[y_i = y_j]} \to \min \quad \text{при условии} \quad \|\mathbf{v}\|_1 = 1, \mathbf{v} \geq \mathbf{0}.$$

F_0 является линейной функцией относительно весов $\mathbf{v}$. Ее требуется минимизировать на множестве, которое является выпуклым. У этой задачи существует решение, причем минимум реализуется в вершине n-мерного симплекса. В каждой вершине этого симплекса все координаты равны нулю, кроме одной, которая равна единице. Таким образом будет отобран единственный признак, что не отвечает требованиям задачи. Поэтому предлагается при построении метрики $\rho(\mathbf{w})(\mathbf{x}, \mathbf{x}')$ вес w_j каждого признака сделать обратно пропорциональным соответствующему элементу вектора $\mathbf{v}$.

3 Прогнозирование тематики документа

Класс объекта $\mathbf{x}_i$ прогнозируется при помощи процедуры голосования. Для объекта $\mathbf{x}_i$ выбирается множество индексов $\mathcal{K}(\mathbf{x}_i)$, состоящее из индексов k ближайших к $\mathbf{x}_i$ объектов,

$$\mathcal{K}(i) = \{\rho_p(\mathbf{x}_i, \mathbf{x}_j) | j \in \mathcal{I}, p \leq k\},$$

где $p = p\{j\}$ — индекс сортировки множества по возрастанию. Для всех классов q выберем из множества $\mathcal{K}(i)$ подмножество индексов $\mathcal{K}_q(i)$ объектов из q-го класса:

$$\mathcal{K}_q(i) = \{j \in \mathcal{K}(\mathbf{x}_i) | y_j = q\}.$$

Для каждого класса q определим его вклад в классификацию объекта $\mathbf{x}_i$ следующим образом. Чем ближе объекты q-го класса лежат к объекту $\mathbf{x}_i$, тем больший вклад в классификацию они вносят.

Определим класс объекта $\mathbf{x}_i$ как тот класс, который вносит наибольший вклад:

$$\widehat{y_i} = q(\mathbf{x}_i) = \underset{q\in\{1,\ldots,K\}}{\arg\max} \sum_{j\in\mathcal{K}_q(i)} \frac{[y_j = q]}{\rho(\mathbf{x}_i, \mathbf{x}_j)},$$

где индикаторная функция $[\cdot]$ определена как

$$[y_j = q] = \begin{cases} 1, & \text{если } y_j = q; \\ 0, & \text{если } y_j \neq q. \end{cases} \tag{9}$$

4 Кластеризация набора документов

Предлагается, используя взвешенную функцию расстояния $\rho_{\mathcal{A}}(\widehat{\mathbf{w}})(\mathbf{x}, \mathbf{x}')$ с заданным набором весов $\widehat{\mathbf{w}}$ и набором признаков $\mathcal{A}$ провести кластеризацию коллекции с помощью алгоритма k-means. Число кластеров K считаем фиксированным согласно постановке задачи. Задаем начальное приближение положений центров кластеров $\mathbf{u}_q, q \in \{1, \ldots, K\}$. Затем для каждого элемента $\mathbf{x}_i$ находим ближайший к нему центр $\mathbf{u}_q$ относим его к кластеру с номером $y_i = q$:

$$y_i = \underset{q\in\{1,\ldots,K\}}{\arg\min}\ \rho(\mathbf{x}_i, \mathbf{u}_q).$$

Осуществляем пересчет положений центров, помещая их в центр масс соответствующих кластеров:

$$\mathbf{u}_q = \frac{\sum_{i\in\mathcal{I}} [y_i = q]\mathbf{x}_i}{\sum_{i\in\mathcal{I}} [y_i = q]}.$$

Алгоритм останавливается, когда кластеризация y_i элементов $\mathbf{x}_j$ стабилизируется.

Для оценки качества кластеризации используем функцию среднего внутрикластерного расстояния F_0, определенную в (8), и функцию среднего межкластерного расстояния

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{i<j}[y_i \neq y_j]} \to \max.$$

Чтобы учесть и внутрикластерное и межкластерное расстояние введем функцию

$$F = \frac{F_0}{F_1} \to \min.$$

5 Оценка соответствия экспертной и вычисленной тематической классификации

Для оценки соответствия рассмотрим два набора: набор $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$ экспертных оценок тем (2), $y_i \in \{1, \ldots, K\}$, где K — число тем, и набор $\widehat{\mathbf{y}}$ вычисленных тем документов коллекции. Рассмотрим произвольную пару документов d_1, d_2. Для оценки соответствия классификаций сосчитаем общее число расхождений $\chi(\mathbf{y}, \widehat{\mathbf{y}})$ в парах по всем документам и поделим на максимальное возможное расхождение.

Наибольшее число расхождений зададим формулой

$$Y = \frac{m(m-1)}{2} - \sum_{i=1}^{K} \frac{M_i(M_i - K)}{2K},$$

где M_i — число документов $\widehat{\mathbf{y}}$ в каждом вычисленном тематическом классе. Это соответствует, например, случаю, когда элементы каждого класса поровну распределяются по всем тематическим кластерам.

Зададим функцию соответствия $\chi(\mathbf{y}, \widehat{\mathbf{y}})$ как

$$\chi(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{Y} \sum_{i<j} [y_i \neq \widehat{y}_j], \quad i, j \in \{1, \ldots, m\}, \tag{10}$$

где индикаторная функция определена в (9). Чем ближе полученная величина к единице, тем сильнее расхождение. Если полученная величина равна нулю, то тематические классификации совпадают.

6 Вычислительный эксперимент

Для иллюстрации работы и оценки качества алгоритма проведен эксперимент кластеризации тезисов научной конференции «European Conference on Operational Research, EURO–2012» на 26 кластеров. Количество кластеров выбрано в согласии с количеством научных областей (англ. main areas) конференции. Каждому тезису конференции был поставлен в соответствие вектор $\mathbf{x}$ в пространстве $\mathbb{R}^n$, где n — количество слов в словаре W. Компоненты вектора заданы одним из трех способов:

А) $0, 1$ в зависимости от того, встречается ли термин в документе;

B) $tf \cdot idf$ терминов;

C) количество раз, которое термин встречается в документе.

Здесь критерий

$$tf(d_i, w_j) = \frac{c(d_i, w_j)}{\sum_{l=1}^{|d|} c(d_i, w_l)},$$

где $c(d_i, w_j)$ — количество раз, которое термин w_j встречается в документе d_i, а критерий

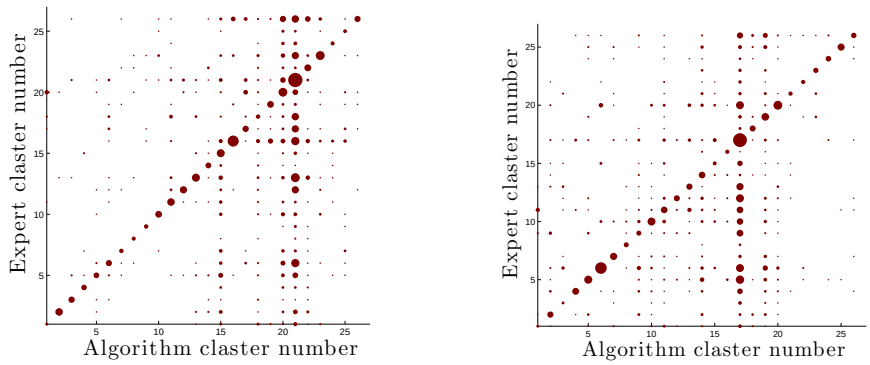
$$idf(w_j) = \log \frac{|D|}{|\{d_i : d_i \ni w_j\}|}.$$

Критерий $tf(d, w_j)$ показывает, насколько типичен термин w_j для документа d_i, а критерий $idf(w_j)$ показывает, насколько термин w_j является типичным для всей коллекции документов D. В вычислительном эксперименте из исходного словаря W отсеивались слова w со слишком большим idf (очень редкие слова, возможно, опечатки) и со слишком маленьким (то есть слова, не выделяющие ни один документ из коллекции остальных).

Отбор слов проводился, чтобы получить не очень крупный словарь W. Это требование обусловлено тем, что в каждом тезисе не более 600 знаков, то есть порядка 50–100 слов, многие из которых являются стоп-словами, например, “and”, “or”, “not”. Отсев стоп-слов проводился фильтрацией заданного списка стоп-слов и с помощью отсева по критерию $tf \cdot idf$. Исключались слова, которые встречались более, чем в 50 и менее, чем в 10 документах из коллекции в 1342 документа.

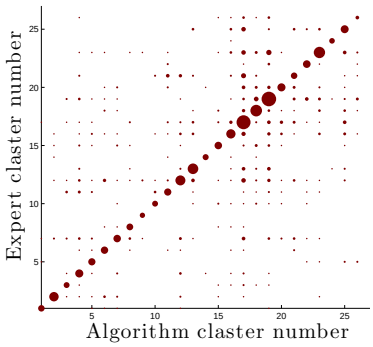
Для кластеризации коллекции документов была выбрана метрика, оптимальная в смысле (8). Начальное положение центров кластеров задавалось согласно экспертной классификации: вычислялось среднее значение по всем документам каждого класса. Для визуализации и оценки расхождений между построенной и экспертной моделью был использован следующий метод, см. рис 1: строилось расхождение между экспертной и полученной моделью. По оси абсцисс и ординат откладывались номера классов от 1 до K, затем все тезисы откладывались на плоскости, причем координатой по оси абсцисс тезиса являлся номер вычисленного класса, полученный предлагаемым алгоритмом, а по оси ординат — номер класса, к которому отнес эксперт данный тезис.

Номера классов упорядочены следующим образом. По исходным центрам кластеров $\mathbf{u}_q$, $q \in \{1, \ldots, K\}$, с помощью метрики $\rho_{\mathcal{A}}(\widehat{\mathbf{w}})(\mathbf{x}, \mathbf{x}')$ вычислялось

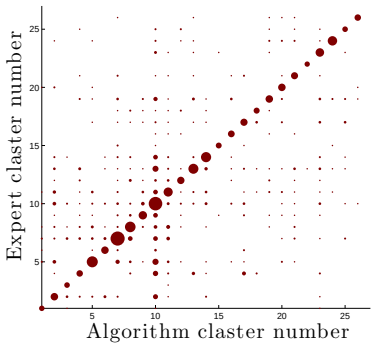


(a) Признаки — $tf \cdot idf$.

(b) Признаки — число повторов слов.



(c) Булевы признаки с отбором признаков.



(d) Булевы признаки без отбора признаков.

Рис. 1: Сравнение экспертной классификации и получившейся кластеризации.

Таблица 1: Значение функции ошибки для разных способов построения набора признаков

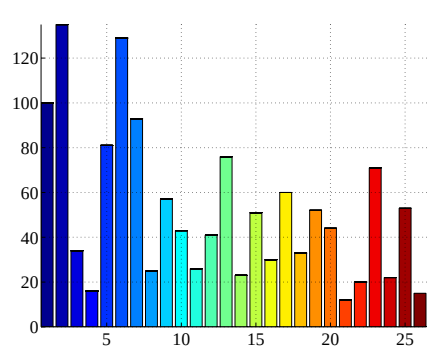
Способ построения признака	Появление термина в документе, на оптимальном наборе $\mathcal{A}$ A)	То же, на всем словаре W A)	Критерий $tf \cdot idf$ B)	Число появлений термина C)
Значение функции ошибки χ	0.073	0.079	0.142	0.154

матрица парных расстояний. Затем выполнялась процедура метрического шкалирования. Вычислялась первая главная компонента матрицы расстояний, классы (представленные исходно как элементы неупорядоченного множества) упорядочивались в соответствии с порядком проекций строк матрицы парных расстояний (строка соответствует классу) на первую главную компоненту. Таким образом, на рис. 1, чем дальше тезис находится от прямой $x = y$ на данной плоскости, тем сильнее отличаются тематики кластеров, к которым он был отнесен предлагаемым алгоритмом и экспертом.

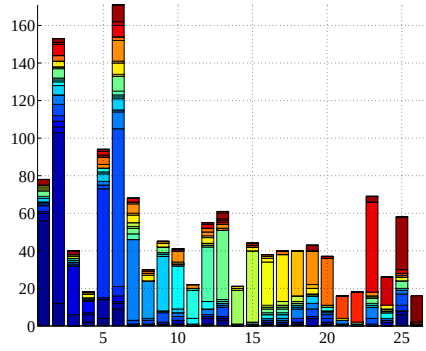
Были проведены эксперименты с использованием различных способов задания признаков; результаты приведены на рис. 1 (a)–(d). Координатой каждого круга по оси абсцисс являлся номер класса, к которому алгоритм отнес данный документ, по оси ординат — класс, назначенный экспертом. Радиус круга определялся количеством документов, попавших в эту точку. Наибольшему кругу соответствует количество документов, равное примерно 80. Как видно из табл. 1, показывающей оценку соответствия экспертной и вычисленной тематической классификации (10), наиболее адекватным способом задания вектора признаков способ A), т. е. факт присутствия термина в документе.

Перераспределение документов и сравнение с их начальным распределением показаны на рис. 2. Каждый столбец соответствует классу; каждому документу присваивался цвет экспертного класса, заданного экспертом. Высота части столбца одного цвета показывала количество документов, экспертно отнесенных к кластеру с данным цветом, которое алгоритм отнес к кластеру, соответствующему номеру столбца. Рис. 2 (a) построен по экспертной кластеризации, поэтому каждый столбец имеет документы только одного цвета, одновременно являющегося цветом класса. Из рис. 2 (b)–(d) видно, что большая часть документов остается в кассах, назначенных экспертами.

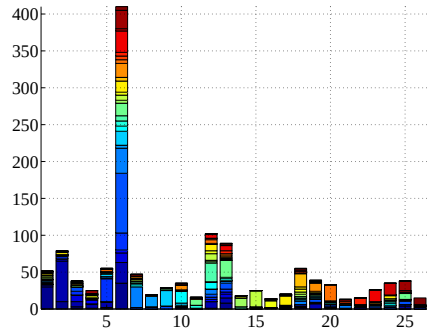
Для сравнения работы алгоритма с отбором признаков и без него изображены гистограммы на рис. 2. Можно заметить, что дополнительный отбор признаков позволяет удалить шумовые слова, например, «matrix», «compute», «activity», которые не удалились при нормализации документов, хотя они не несут особой информации о принадлежности тезиса к определенной теме. Наличие их приводит к появлению кластеров, включающих документы чужих кластеров, как например кластер номер 6 на рис. 2 (c), (d).



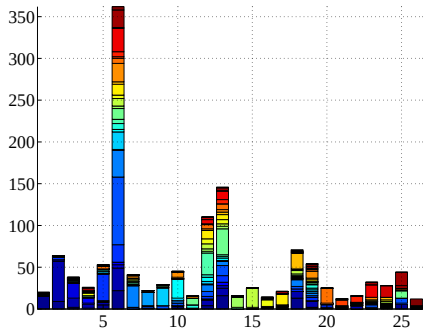
(a) Экспертная классификация.



(b) Булевы признаки с отбором признаков.



(c) Признаки: число повторов слов.



(d) Признаки: $tf \cdot idf$.

Рис. 2: Распределение документов по темам для экспертной классификации и для построенных кластеризаций.

Заключение

В данной работе решается задача метрической кластеризации коллекции документов. Для вычисления расстояний используется метрика, оптимальная относительно введенной функции благонадежности. Получено точное решение задачи оптимизации метрики. Предложена количественная характеристика рассогласования классификаций. Вычислительный эксперимент показал, что в целом вычисленная тематическая классификация соответствует экспертной классификации тезисов конференции EURO-2012, но есть и заметные отклонения от классификации, что может свидетельствовать о значительном числе междисциплинарных статей.

Список литературы

[1] *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation. // Journal of Machine Learning Research, 2003. Vol. 3. Pp. 993-1022.

[2] *Hofmann T.* Probabilistic latent semantic indexing. // Proceedings of the 22nd annual interanational ACM SIGIR conference on research and development in information retrieval. New York: ACM, 1999. Pp. 50–57.

[3] *Manning C. D., Raghavan P., Schütze H.* Introduction to information retrieval. // Cambridge: Camdridge University Press, 2008.

[4] *Hartigan J. A., Wong M. A.* Algorithm as 136: A k-means clustering algorithm. // Applied statistics, 1978. Vol. 28. Pp. 100–108.

[5] *Loohach R., Garg K.* Effect of distance functions on simple k-means clustering problem. // International Journal of Computer Applications, 2012. Vol. 49. No. 6.

[6] *Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С.* Алгоритмы обнаружения эмпирических закономерностей. // Новосибирск: Наука, 1985.

[7] *Pal N. R.,Bezdek J. C.* On cluster validity for the fuzzy c-means model. // IEEE Transactions on Fuzzy Systems, 1995. Vol. 3(3). Pp. 370–379.

[8] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. // Новосибирск: Издательство И.М., 1999.

[9] *Борисова И. А.* Использование fris-функции для построения решающего правила и выбора признаков (задача комбинированного типа dx). //Новосибирск. Знания. Онтологии. Теории. Материалы Всероссийской Конференции, 2007. Т. 1. С. 37–44.

[10] *Tibshirani R., Hastie T.* Discriminative adaptive nearest neighbor classification. // IEEE transactions on pattern analysis and machine intelligence, 1996.

[11] *Peng J., Gunopulos D., Domenciconi C.* An adaptive metric machine for pattern classification. // NIPS, 2000.

[12] *Zagoruiko N. G.* Methods of recognition based on the function of rival similarity. // Pattern recognition and image analysis, 2008. Vol. 18(1). Pp. 1–6.

[13] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models. // Frontiers of computer science in China, 2010. Vol. 4(2). Pp. 280–301.

[14] *Константинов Р. В.* Функциональный анализ. Курс лекций. Долгопрудный: МФТИ, 2009.

[15] *Boyd S., Vandenberghe L.* Convex Optimization. Cambridge University Press, 2004.

Поступило 06.10.2012.

Адуенко Александр Александрович (aduenko1@gmail.com), студент, Московский физико-технический институт.

Кузьмин Арсентий Александрович (senatormipt@gmail.com), студент, Московский физико-технический институт.

Стрижов Вадим Викторович (strijov@ccas.ru, http://strijov.com), к.ф.-м.н., н.с., Вычислительный центр Российской академии наук.

Feature selection and metrics' optimisation when clustering documents' collection.

A. Aduenko, A. Kuzmin, V. Strijov

Abstract. This paper deals with the problem of verification of correctness of a thematic clustering of texts with the help of metric algorithms. The algorithm of selection the optimal distance function for texts is proposed. Correspondence between received texts' clustering and their expert classification is studied. The results of clusterisation and their correspondence to expert thematic classification are illustrated in the computing experiment on the real text collection.

Keywords: metric classification, nearest neighbors method, similarity function, monotonous function, feature selection, compactness hypothesis.

Aduenko Alexandr (aduenko1@gmail.com) student, Moscow Institute of Physics and Technology.

Kuzmin Arsentii (senatormipt@gmail.com) student, Moscow Institute of Physics and Technology.

Strijov Vadim (strijov@ccas.ru, http://strijov.com), PhD, researcher, Computing Center of the Russian Academy of Sciences.