УДК 519.6

Е.С. Мангалова, Е.Д. Агафонов

О ПРОБЛЕМЕ ВЫДЕЛЕНИЯ ИНФОРМАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧЕ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Описан подход к решению проблемы выделения информативных признаков в задаче классификации текстовых документов. Задача характеризуется высокой размерностью пространства исходных признаков и сравнительно малым объемом обучающей выборки. Предложен алгоритм формирования подмножеств информативных признаков. С применением алгоритма решена задача классификации медицинских документов.

Ключевые слова: *text mining*, *выделение информативных признаков*, *классификация.*

Разработка медицинских баз данных и предоставление свободного доступа к ним позволяет пользователям осуществлять эффективный поиск документов, содержащих высокоспециализированные медицинские знания. Быстрый рост числа каталогов научных статей и текстовых хранилищ, например таких, как MEDLINE или PubMed Central (PMC), обуславливает необходимость в развитии методов автоматической расстановки тэгов и автоматической классификации текстовых данных. Специалисты часто осуществляют поиск медицинских документов для получения информации о современных средствах диагностики, лекарственных препаратах, возможных осложнениях в результате того или иного лечения и т.п. При этом они используют в запросах специфическую терминологию, которая может быть правильно интерпретирована только с помощью специализированной медицинской онтологии (содержащей специальные для медицины значения терминов), например такой, как Medical Subject Headings (MeSH). Для того чтобы облегчить процесс поиска, все документы базы данных должны быть проиндексированы в соответствии с онтологией. Результаты поиска могут быть сгруппированы в классы документов, которые соответствуют разным разделам (например, «методы диагностики», «симптомы» и т.д.). Такие классы могут быть пересекающимися: один документ может содержать информацию, относящуюся к различным разделам, так статья, посвященная современным способам лечения какого-либо заболевания, может включать в себя как методы диагностики, так и симптомы, информацию о лекарственных препаратах, их противопоказаниях и т.д.

1. Постановка задачи

Постановка задачи классификации сформулирована организаторами JRS 2012 Contest [1]. Имеется некоторое множество научных статей по медицине, которые необходимо отнести к 83 классам. Классы пересекаются, так что статья может быть отнесена сразу к нескольким классам. Ситуация, когда статья не принадлежит ни одному из заданных классов, исключена из рассмотрения. Организаторы конкурса предоставили обучающую выборку – информацию о 10 000 статьях с указанием классов, к которым они отнесены.

Признаками для классификации в рассматриваемой задаче принимают так называемые «степени раскрытия» определенных научных терминов в статьях. Степень раскрытия термина определяется частотой, с которой термин встречается в статье, а также некоторыми дополнительными факторами, которые организаторами конкурса не разглашаются. Выделяют 25 640 ключевых терминов, степени раскрытия которых приведены для каждой статьи в обучающей выборке. Как правило, подавляющее большинство степеней раскрытия для конкретной статьи будет принимать нулевое значение, что соответствует отсутствию термина в статье.

Сформулированную задачу классификации можно отнести к сложным и большим, так как модель принятия решения отсутствует, классические методы классификации неприменимы. Исходный набор признаков классификации значительно превышает объем обучающей выборки: доступные данные в обучающей выборке плохо обусловлены.

Важнейшей задачей на пути решения сформулированной проблемы классификации становится преобразование множества признаков. Требуется перейти к новой системе признаков, значительно уменьшив их количество без потери информативности нового набора.

2. Описание алгоритма

Задача классификации 83 пересекающихся классов может быть сведена к решению 83 задач бинарной классификации (принадлежности или непринадлежности статьи к каждому классу).

Введем следующие обозначения:

x_i^j – степень раскрытия j-го термина в i-й статье ($j = \overline{1,n}$, $i = \overline{1,N}$), степень раскрытия термина может принимать значение от 0 до 999: 0 – термин в статье не встречается, 999 – термин раскрыт в статье полностью;

y_i^k – принадлежность i-й статьи к k-му классу ($i = \overline{1,N}$, $k = \overline{1,m}$)

$$y_i^k = \begin{cases} 1, & i\text{-я статья } \in k\text{-му классу}, \\ 0, & i\text{-я статья } \notin k\text{-му классу}; \end{cases}$$

$\Omega^k = \left\{ i : y_i^k = 1 \right\}$ – множества порядковых номеров статей k-го класса,

$T^k = \left\{ j : x_i^j > 0, i \in \Omega^k \right\}$ – множества терминов, встречающихся в статьях, принадлежащих k-му классу.

В исходном пространстве n признаков классы не являются компактными. Гипотеза компактности состоит в том, что элементы одного и того же класса отражаются в признаковом пространстве в геометрически близкие точки. Если среди признаков имеется много случайных, неинформативных, то элементы класса могут оказаться далекими друг от друга и рассеянными среди элементов других классов [2, с. 29]. Таким образом, требуется выделить такое подмножество исходных признаков, в котором множество элементов класса компактно. Однако применение известных алгоритмов выделения информативных признаков, таких, как метод последовательного добавления признаков, метод последовательного сокращения [2, с. 108], метод случайного поиска с адаптацией [2, с. 110], неприменимы из-за отсутствия большинства терминов в каждой статье. Исключение одно-

го любого признака из множества T^k не приведет к росту качества классификации, так же, как и добавление одно информативного признака к другому.

Идея предлагаемого подхода к выделению информативных признаков состоит в синтезе набора подпространств исходного пространства признаков, таких, что в рамках каждого подпространства среди элементов выборки с ненулевыми значениями признаков возможно выделение компактных групп элементов принадлежащих и не принадлежащих к классу. Для каждого подпространства признаков из исходной формируется своя обучающая выборка, элементы которой не содержат признаков с нулевыми значениями. Объединение полученных подпространств не принесет значимого улучшения качества классификации, так как в рамках каждого подпространства признаков возможно вынесение однозначного решения о принадлежности или не принадлежности элемента с ненулевыми значениями признаков к классу, а после объединения подпространств в одно, за счет уменьшения объема выборки, количество выносимых решений сократится.

Расстояние $\rho_j(x_i, x_s)$ между элементами x_i и x_s в пространстве j-го признака будем определять как

$$\rho_j(x_i, x_s) = \begin{cases} 2\dfrac{\left|x_i^j - x_s^j\right|}{x_i^j + x_s^j}, & x_i^j x_s^j > 0, \\ 1, & x_i^j x_s^j = 0. \end{cases}$$

Расстояние $\rho_j(x_i, x_s)$ может принимать значения от 0 до 1:

$\rho_j(x_i, x_s) = 0$, если степени раскрытия j-го термина в статьях i и s равны,

$\rho_j(x_i, x_s) = 1$, если j-й термин не встречается хотя бы в одной из статей i и s.

Теперь введем меру расстояния $\rho_P(x_i, x_s)$ между элементами x_i и x_s в произвольном подпространстве признаков $P = (p_1, p_2, \ldots, p_r)$, где r – размерность подпространства, $r \in \{1, 2, \ldots, n\}$, $p_v \in \{1, 2, \ldots, n\}$, $v = \overline{1,r}$:

$$\rho_P(x_i, x_s) = \begin{cases} r^{-1}\sum\limits_{v=1}^{r} \rho_{p_v}(x_i, x_s), & \max\limits_v \rho_{p_v}(x_i, x_s) < 1, \\ 1, & \max\limits_v \rho_{p_v}(x_i, x_s) = 1. \end{cases}$$

$\rho_P(x_i, x_s) = 0$, если степени раскрытия всех терминов $p_1, p_2, \ldots, p_r$ в статьях i и s попарно равны,

$\rho_P(x_i, x_s) = 1$, если хотя бы один из терминов $p_1, p_2, \ldots, p_r$ не встречается хотя бы в одной из статей i и s.

Для искомого набора пространств признаков $D^k = \{D_l^k, l = \overline{1, L^k}\}$, где $D_l^k = \left(d_{1,l}^k, d_{2,l}^k, \ldots, d_{r,l}^k\right)$, L^k – количество подпространств, требуется выполнение условий

$$\begin{aligned} &\forall i \in \Omega^k : \rho_{D_l^k}\left(x_i, x_{s,i,k,1}^*\right) < \rho_{D_l^k}\left(x_i, x_{s,i,k,0}^*\right) \cup \rho_{D_l^k}\left(x_i, x_{s,i,k,1}^*\right) = \rho_{D_l^k}\left(x_i, x_{s,i,k,0}^*\right) = 1, \\ &\forall i \notin \Omega^k : \rho_{D_l^k}\left(x_i, x_{s,i,k,1}^*\right) > \rho_{D_l^k}\left(x_i, x_{s,i,k,0}^*\right) \cup \rho_{D_l^k}\left(x_i, x_{s,i,k,1}^*\right) = \rho_{D_l^k}\left(x_i, x_{s,i,k,0}^*\right) = 1, \end{aligned} \quad (1)$$

где $$x^*_{s,i,k,0} = \arg\min_{x_s}\left\{\rho_{D_l^k}(x_i, x_s) : x_s \notin \Omega^k, x_i \neq x_s\right\},$$

$$x^*_{s,i,k,1} = \arg\min_{x_s}\left\{\rho_{D_l^k}(x_i, x_s) : x_s \in \Omega^k, x_i \neq x_s\right\}.$$

Иными словами, в каждом из искомых пространств признаков D_l^k ближайшим к элементу, принадлежащему *k*-му классу, должен быть элемент *k*-го класса, а к элементу, не принадлежащему этому классу, – элемент также ему не принадлежащий.

Определим итеративную процедуру построения множества пространств признаков $D^k = \left\{D_l^k, l = \overline{1, L^k}\right\}$:

1. $r = 1$, $L^k = 0$, пространство поиска *r*-го термина в комбинации $G_r^k = T^k$;

2. Формирование множества подпространств признаков $D_l^k = \left(d_{1,l}^k, d_{2,l}^k, ..., d_{r,l}^k\right)$, где $d_{v,l}^k \in G_v^k$, $v = \overline{1, r}$, $l = \overline{L^k + 1, L^k + R_r^k}$, R_r^k – количество множеств признаков размерности r, таких, что выполняются условия (1);

3. Сокращение пространства поиска *r*-го термина:

$$G_r^k = G_r^k \setminus \bigcup_{l=L^k+1}^{L^k+R_r^k} D_l^k,\ L^k = L^k + R_r^k,\ r = r + 1,\ G_r^k = G_{r-1}^k;$$

4. Классификация в подпространствах $D^k = \left\{D_l^k, l = \overline{1, L^k}\right\}$:

$$\hat{y}_i^k = \max_l\left\{y_q^k : x_q = \arg\min_{x_s}\rho_{D_l^k}(x_i, x_s)\right\}.$$

5. Вычисление ошибки классификации:

$$E_r = N^{-1}\sum_{i=1}^{N} I\left(\hat{y}_i^k \neq y_i^k\right). \qquad (2)$$

6. Если $E_r < \varepsilon$ или $E_{r-1} - E_r < \delta$, или $G_r^k = \varnothing$, множество подпространств признаков $D^k = \left\{D_l^k, l = \overline{1, L^k}\right\}$ сформировано, иначе к шагу 2.

Формирование наборов информативных признаков происходит последовательно: если D_l^k удовлетворяет условиям (1), то D_l^k не может быть подпространством любого другого набора признаков. Таким образом, с одной стороны, с ростом размерности искомых пространств *r* возрастает количество комбинаций признаков, с другой стороны, сокращается область поиска таких комбинаций G^k.

Специфика задачи классификации текстовых документов такова, что большинство статей содержат от 80 до 130 терминов (см. рис. 1), большинство терминов встречаются менее чем в 100 статьях (см. рис. 2).

Анализ рис. 1 и 2 показывает, что рост количества комбинаций признаков с ростом количества признаков в подпространстве *r* ограничен. Это обусловлено тем, что вероятность того, что комбинация терминов встречается во многих статьях, мала.

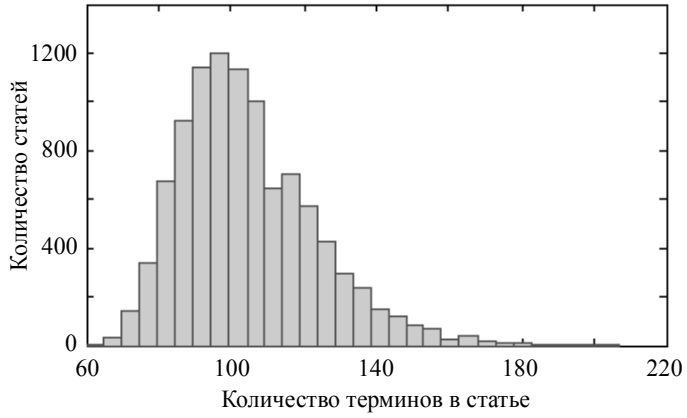

Рис. 1. Гистограмма распределения количества терминов в статье

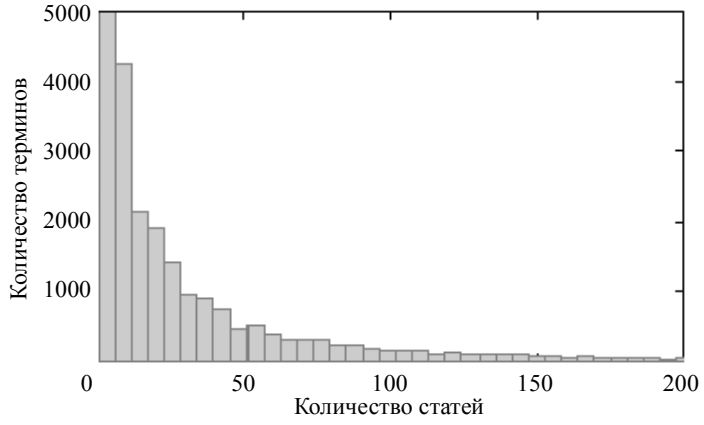

Рис. 2. Гистограмма распределения количества статей, содержащих одинаковые термины

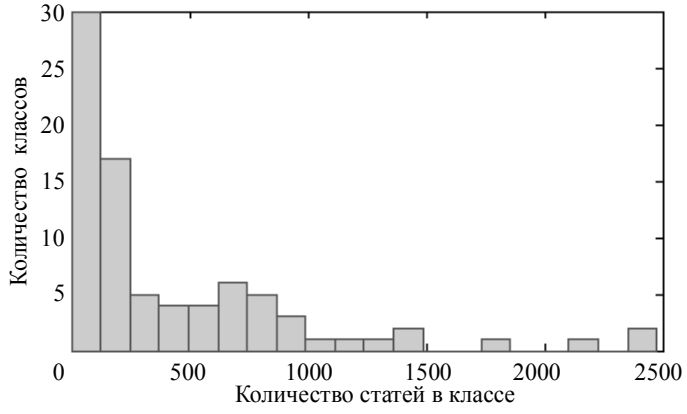

Рис. 3. Гистограмма распределения статей в классах

Конфигурации наборов информативных признаков непосредственно зависят от количества элементов, принадлежащих *k*-му классу. При этом элементы обучающей выборки распределены по классам неравномерно: гистограмма распределения статей в классах $|\Omega^k|$ приведена на рис. 3. Следует отметить, что для некоторых малочисленных классов пространства признаков $D^k = \left\{ D_l^k, l = \overline{1, L^k} \right\}$, удовлетворяющие условиям (1), могут не существовать.

Применим процедуру выделения подмножеств информативных признаков для классов с различным количеством представителей. Вначале рассмотрим самый многочисленный класс (k = 40, $|\Omega^k| = 2475$).

На рис. 4 приведены зависимости количества подпространств признаков D^k, количества признаков, входящих хотя бы в одно подпространство множества D^k и ошибки классификации от *r* ($E_0 = |\Omega^k|/N$), гистограмма распределения количества подпространств, содержащих одинаковые термины.

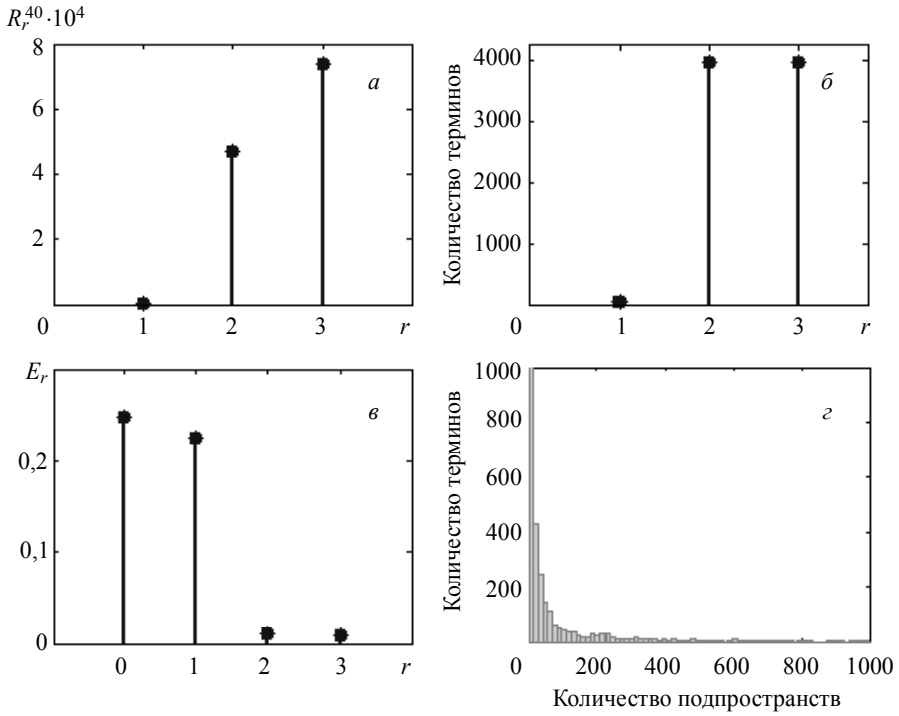

Рис. 4. Зависимость количества подпространств признаков от размерности подпространства *r* (*а*); зависимость количества признаков, входящих в подпространства, от размерности подпространства *r* (*б*); ошибка классификации (*в*); гистограмма распределения количества подпространств, содержащих тот или иной термин

Применим алгоритм выделения информативных признаков для класса, находящегося на 42-м месте по количеству представителей (см. рис. 5). Выбор этого класса (*k* = 29, $|\Omega^k| = 201$) обусловлен тем, что количество представителей для него – среднее по всем классам.

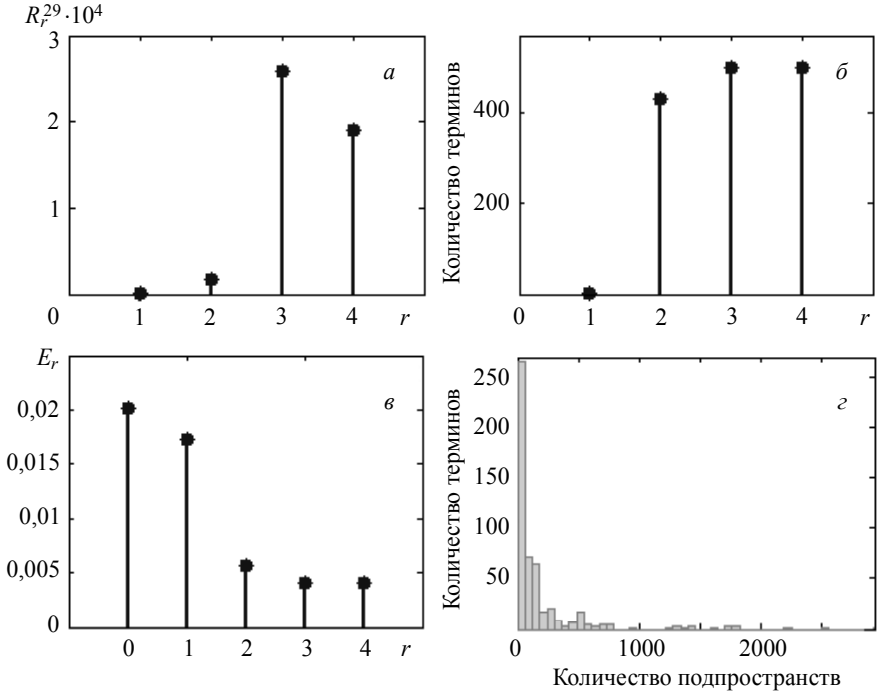

Рис. 5. Зависимость количества подпространств признаков от размерности подпространства r (*а*); зависимость количества признаков, входящих в подпространства, от размерности подпространства r (*б*); ошибка классификации (*в*); гистограмма распределения количества подпространств, содержащих тот или иной термин (*г*)

Приведенные графики (рис. 4, 5) иллюстрируют закономерности функционирования алгоритма классификации. Сходимость ошибки к установившемуся значению фактически наблюдается для значения $r = 2$.

Табл. 1 содержит результаты сравнения ошибок классификации (2), вычисленных по методу скользящего экзамена, для следующих методов:

1 – предложенного алгоритма синтеза информативных признаков с последующей классификацией,

2 – метода ближайших соседей, учитывающего все n признаков,

3 – алгоритма Random Forest [3, с. 5], учитывающего все n признаков.

Сравнение методов классификации по критерию качества

k	$\lvert\Omega^k\rvert$	Ошибка классификации		
		1	2	3
29	201	0,0044	0.0183	0.0188
40	2475	0,054	0.1869	0.198

Заключение

В работе предложен подход к синтезу системы информативных признаков в задаче классификации текстовых документов. Он предусматривает формирование подпространств признаков в соответствии с предположением о компактности

объектов классификации в каждом из подпространств. Подпространства признаков, входящие в систему, выбирались исходя из выполнения условий (1). Описанный подход позволяет значительно улучшить качество классификации по сравнению с методом ближайших соседей, учитывающим все n признаков, Random Forest. Дальнейшее улучшение качества будет требовать корректировки метрики, критерия принадлежности к классу.

ЛИТЕРАТУРА

1. JRS 2012 Data Mining Competition: Topical Classification of Biomedical Research Papers. [электронный ресурс], URL: tunedit.org/challenge/JRS12Contest. (дата обращения: 15.04.2012).
2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.
3. *Breiman L.* Random forests // Machine Learning. 2001. V. 45 (1). P. 5–32.

Мангалова Екатерина Сергеевна
Сибирский государственный аэрокосмический университет
Агафонов Евгений Дмитриевич
Сибирский федеральный университет
E-mail: e.s.mangalova@hotmail.com, agafonov@gmx.de Поступила в редакцию 5 мая2012 г.

Mangalova Ekaterina S., *Agafonov Evgeny D.* (Siberian State Aerospace University, Siberian Federal University). **On features selection approach for text mining problem.**

Keywords: text mining, *feature selection*, classification.

One approach of classification features selection for the text mining problem is proposed in the paper The initial system of features is defined in a high-order space, at the same time learning data set is relatively small. Classes form vast intersected system. One algorithm of features subsets generation is proposed in the paper. It is based upon compactness hypothesis: in every resulting features subset the nearest element to the one that belongs to the k's class, should also belong to the k's class, and the nearest element to the one that doesn't belong to the k's class, shouldn't belong to the k's class. Using the algorithm a medical documents classification problem, offered by JRS 2012 Contest team, has been solved. By its classification accuracy the proposed approach exceeds the nearest neighbors method and the Random Forest algorithm.