ПРИКЛАДНАЯ МАТЕМАТИКА И МЕХАНИКА

Вестник Сыктывкарского университета.
Серия 1: Математика. Механика. Информатика.
Выпуск 1 (22). 2017

УДК 519.766.4, 81'32

КЛАССИФИКАЦИЯ ТЕКСТОВ МЕТОДАМИ РАСПОЗНАВАНИЯ ОБРАЗОВ

С. А. Хозяинов

Статья демонстрирует процесс классификации текстов методами распознавания образов. В качестве примера рассмотрена проблема авторства статей, приписываемых А. С. Пушкину. Предложены способы повышения надежности распознающей системы.

Ключевые слова: классификация текстов, методы распознавания образов, атрибуция, А. С. Пушкин.

1. Введение: текст как объект классификации

Построение классификации является важным этапом многих научных исследований. В немалой степени выбор метода решения этой задачи определяется двумя факторами: целью исследования и сложностью объектов классификации. Для предварительной классификации небольшого количества объектов по двум-трем признакам достаточно общенаучных методов. Для создания качественной классификации большого количества сложных объектов требуются значительно более чувствительные и точные методы.

Текст является очень сложным объектом; он многогранен, благодаря чему представляет интерес для разных наук. В любом случае текст — это последовательность знаков письма, в которых зафиксированы некоторое состояние языка, определенная фаза развития общества, взгляды и знания индивидуального или коллективного автора. Поэтому очевидно, что классификация текстов невозможна без их содержательного историко-филологического анализа. В то же время ясно, что формализация методов классификации текстов необходима, если мы хотим иметь возможность исследовать обширный материал в разумные

© Хозяинов С. А., 2017.

сроки. Такие формализованные инструменты исследования пришли в историко-филологические науки извне, из области математических наук.

Пожалуй, одна из наиболее трудных задач при классификации текстов — отыскать баланс между содержательным и формальным подходом к решению задачи (примеры см. в статье [4]). Здесь важно добиться того, чтобы они дополняли друг друга, действовали согласованно: это позволит избежать и субъективности, и бессодержательного формализма, равно отдаляющих классификацию от действительности, получить наиболее полную и точную картину.

2. Постановка задачи и объект исследования

Классификация текстов не является самоцелью филологии — она служит инструментом решения конкретных практических задач.

К их числу принадлежит атрибуция (определение авторства текста) — одна из наиболее сложных филологических проблем. Для ее успешного решения требуется не только разноаспектный лингвистический анализ текста, но и тщательное изучение истории его написания, издания, прочтений и интерпретаций. Точные и недвусмысленные указания на авторство произведения могут содержать письменные документы (частная переписка, критические статьи и пр.). Однако в наиболее сложных случаях, когда таких документов нет (или они существуют, но не содержат нужных сведений), всю информацию приходится извлекать из языка атрибутируемых текстов.

При этом главную роль в исследовании играют инструменты анализа текста, а сама задача, если к ней внимательно присмотреться, сводится к разделению текстов на группы по некоторым достаточно формальным признакам. В их числе, например, наличие (или отсутствие) тех или иных слов, оборотов, синтаксических конструкций, преобладание определенного порядка слов и пр. Если какой-либо текст (или группа текстов) по сумме языковых особенностей подобного рода оказывается близок к некоторому индивидуальному авторскому стилю и при этом определенно далек от других авторских стилей, тогда в общем случае проблема авторства считается решенной.

Этот общий случай решения задачи осложняется одним обстоятельством: атрибуция текста какому-либо автору означает допущение, что наши знания о его творчестве неполны и, следовательно, неполным является описание его индивидуального авторского стиля. Это следует учитывать при определении авторства как на уровне формулирования задач и выводов, так и на уровне методическом. В частности, ясно, что

если мы не имеем однозначных документальных доказательств авторства текста, то нам остается лишь определять *вероятность* того или иного варианта решения проблемы.

В различных вариациях (частных случаях) постановка и решение задачи атрибуции может требовать и других существенных оговорок.

Как можно видеть из этого краткого обзора, атрибуция представляет собой типичную классификационную задачу. На ее примере хорошо видны основные вопросы классификации текстов: 1) какие языковые признаки наиболее важны для классификации текстов? 2) какие методы и при каких условиях можно и, главное, нельзя использовать для классификации текстов?

Особенно внимательное отношение к этим вопросам можно найти в методике определения авторства, разработанной М. А. Марусенко и основанной на идеях теории распознавания образов [3]. Существуют и другие методики атрибуции, тоже использующие формальные методы, но в них не так детально, на наш взгляд, раскрыт филологический аспект проблемы и не так много внимания уделено (или же совсем не уделено никакого внимания) названным выше вопросам [7].

Применение методов распознавания образов в задачах атрибуции складывается из четырех основных моментов: 1) построения математических моделей априорных классов и распознаваемых объектов (эти понятия вводятся ниже); 2) определения метрики сходства и различия объектов в признаковом пространстве (выбор функции расстояния между объектами); 3) построения решающего правила распознающего алгоритма; 4) проверки качества классификации. В конкретном эксперименте по атрибуции содержание пункта 1 определяется методами получения априорной информации и ее оценки, а способы реализации пунктов 2—4 задают математический облик системы распознавания текстов.

Рассмотрим реализацию первых трех пунктов этой методики на примере решения проблемы авторства статей, приписываемых А. С. Пушкину. При этом постараемся описать наиболее характерные стороны классификации, для чего рассмотрим атрибуцию десяти текстов (см. табл. 1, где N — число предложений; полное описание эксперимента см. в книге [6]).

3. Построение математических моделей текстов

На этом этапе эксперимента уже завершена стадия историко-филологического анализа, определен список текстов, подлежащих классификации. Этот список состоит не менее чем из двух групп: 1) корпуса

Таблица 1

Статьи, приписываемые А. С. Пушкину

Код	Текст	Выходные данные	N
T02	Письмена Вавилонские	Литературная Газета, 1830, № 1, с. 7	9
T04	Краткая всеобщая география	Литературная Газета, 1830, № 3, с. 22—23	8
T15	Невский альманах на 1830 год	Литературная Газета, 1830, № 12, с. 96	9
T21	«Все благоразумные люди предвидели. . . »	Литературная Газета, 1830, № 23, с. 186	5
T23	«С некоторых пор, Журналисты. . . »	Литературная Газета, 1830, № 36, с. 293	13
T26	«В газете: Le Furet. . . »	Литературная Газета, 1830, № 45, с. 72	2
T29a	«Горестно видеть. . . »	Литературная Газета, 1830, № 53, с. 139—140	16
T41	Французская академия	Современник, 1836, т. II, с. 14—52	15
T43	Письмо к издателю	Современник, 1836, т. III, с. 321—329	85
T45	Несколько слов о Современнике	Северная пчела, 1836, № 86, с. 341—344	34

текстов со спорным авторством; 2) по меньшей мере двух корпусов текстов писателей, претендующих на авторство текстов группы 1. Тексты группы 1 называются *распознаваемыми объектами* (РО), корпусы текстов группы 2 — *априорными классами* (АК).

В нашем эксперименте были изучены пять АК мощностью k текстов и объемом n предложений: Ω_1 (П. А. Вяземский), $k = 18$, $n = 946$; Ω_2 (Н. В. Гоголь), $k = 29$, $n = 613$; Ω_3 (А. А. Дельвиг), $k = 36$, $n = 591$; Ω_4 (А. С. Пушкин), $k = 43$, $n = 1001$; Ω_5 (О. М. Сомов), $k = 9$, $n = 296$. Корпус текстов со спорным авторством составили 46 текстов общим объемом 650 предложений. Объект T29 был разбит на две части (T29а и T29б) по причине его внутренней неоднородности (в этой статье мы рассматриваем атрибуцию первой его части — T29а). Благодаря этому число РО увеличилось до 47.

Покажем на примерах, из чего состоит процедура построения математических моделей АК и РО.

3.1. Определение *априорного словаря параметров* (АСП). Эта часть процедуры является ключевой. Неправильный выбор параметров для исходного описания текстов способен обессмыслить всю процедуру классификации. В АСП мы включили 49 параметров синтаксического и морфологического уровней, которые позволяют измерить те характеристики текстов, которые адекватно описывают особенности индивидуального стиля писателя (например, количество сочиненных и подчиненных предложений, количество знаменательных и служебных слов, количество существительных, прилагательных, наречий и т. п.). Проблемы формирования АСП подробно описаны М. А. Марусенко [3, с. 66—75].

3.2. Создание первичных описаний АК на языке АСП. Итогом этого этапа являются матрицы данных, каждая из которых отражает результаты лингвистического анализа текстов того или иного АК и обладает размерностью Nxn, где N — число параметров, а n — число предложений. В нашем случае было получено пять матриц данных размерностью 49x100 каждая.

3.3. Свертывание исходного параметрического пространства, или определение *словаря* так называемых *информативных параметров* (ИП) — параметров, позволяющих наилучшим образом различать АК.

Методы реализации этого этапа заслуживают обсуждения в отдельной статье. Здесь же мы остановимся на одном из методов — т. н. схеме Бонгарда [1].

На первом ее этапе из АСП выделяется подмножество параметров, релевантных для различения одной или более пар классов. Важным вопросом здесь является выбор классифицирующей функции. В качестве таковой мы применили t-критерий Стьюдента в приближении Уэлча с уровнем значимости $\alpha = 0,05$ и числом степеней свободы $f = 198$ (критическое значение — 1,973):

$$t = \frac{|\overline{x}_1 - \overline{x}_2|}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{0,5}}. \qquad (1)$$

Классификация АК позволила выделить подмножество из 29 параметров, самый мощный из которых оказался релевантен для различения семи пар АК. Статистические характеристики АК и результаты вычисления t-критерия здесь приведем не в полном объеме, а выборочно (см. табл. 2—3) — только по тем параметрам, которые в итоге оказались информативными (см. далее).

Таблица 2

Статистические характеристики АК по ИП

ИП	Ω_1		Ω_2		Ω_3		Ω_4		Ω_5	
	$\bar{x}_i$	σ_i	$\bar{x}_i$	σ_i	$\bar{x}_i$	σ_i	$\bar{x}_i$	σ_i	$\bar{x}_i$	σ_i
X09	0,650	0,857	0,650	0,821	0,510	0,689	0,330	0,551	0,520	0,759
X18	5,030	3,836	3,890	2,930	4,260	3,296	3,330	3,039	4,670	3,671
X24	1,100	1,235	1,510	1,411	1,010	1,185	0,790	0,957	1,300	1,124
X32	1,660	1,451	1,740	1,368	1,500	1,219	1,320	0,909	1,990	1,878

Таблица 3

Значения t-критерия Стьюдента для пар АК по ИП

ИП	Ω_1/Ω_2	Ω_1/Ω_3	Ω_1/Ω_4	Ω_1/Ω_5	Ω_2/Ω_3	Ω_2/Ω_4	Ω_2/Ω_5	Ω_3/Ω_4	Ω_3/Ω_5	Ω_4/Ω_5
X09	0,000	1,273	3,141	1,136	1,306	3,236	1,163	2,040	0,098	2,026
X18	2,362	1,522	3,474	0,678	0,839	1,327	1,661	2,074	0,831	2,812
X24	2,187	0,526	1,984	1,198	2,714	4,223	1,164	1,444	1,776	3,455
X32	0,401	0,844	1,986	1,391	1,310	2,557	1,076	1,184	2,189	3,211

На втором этапе схемы Бонгарда производится свертывание параметрического пространства на подмножестве релевантных параметров, т. е. определяется собственно состав словаря ИП. Содержанием процедуры является обработка корреляционной матрицы связей параметров $R = \{\bar{\rho}_{jk}\}_{j,k=1}^{n}$ размерностью nxn, где n — число параметров из АСП, а выборочные коэффициенты корреляции $\bar{\rho}_{jk}$ представляют собой косинус угла a_{jk} в N-мерном пространстве (объектов) между векторами x^j и x^k:

$$\bar{\rho}_{jk} = cos\alpha_{jk}. \tag{2}$$

Значения $\bar{\rho}_{jk}$ лежат в интервале $-1 \leq \bar{\rho}_{jk} \leq 1$. Матрица симметрична, т. е. $\bar{\rho}_{jk} = \bar{\rho}_{kj}, j, k = \overline{1, n}$, а каждая ячейка ее главной диагонали содержит единицу — значение корреляции j-го параметра с самим собой.

Выбор метода обработки корреляционной матрицы также представляет собой отдельный вопрос. Опишем пример применения так называемого критерия эффективности E_i, который определяется как отношение средней внегрупповой корреляции i-го релевантного параметра $\bar{r}_i^{n-m}$ к средней внутригрупповой корреляции этого же параметра $\bar{r}_i^{m}$

[2, с. 10—11]:

$$\bar{r}_i^m = \frac{\sum_{j=1}^{m} |r_{ij}| - 1}{m - 1}, \tag{3}$$

$$\bar{r}_i^{n-m} = \frac{\left(\sum_{j=1}^{n} |r_{ij}| - 1\right) - \left(\sum_{j=1}^{m} |r_{ij}| - 1\right)}{n - m - 1}, \tag{4}$$

$$E_i = \frac{\bar{r}_i^{n-m}}{\bar{r}_i^m}, \tag{5}$$

где m — число релевантных параметров (29), r_{ij} — коэффициент корреляции i и j-го параметров в матрице, n — число всех параметров из АСП (49).

Эффективным считается параметр, для которого значение E_i больше единицы. Кроме того, чтобы словарь ИП был достаточно мощным, следует выбирать из числа параметров, различающих наибольшее количество пар АК. В нашем случае первое условие не выполняется совсем, но в группе параметров, различающих более четырех пар авторов (лучший показатель мощности), существуют параметры, обладающие наибольшим значением E_i. Исходя из того, что словарь ИП не должен быть большим, чтобы сохранять невысокую внутригрупповую корреляцию между ИП, мы сформулировали очевидное для нашей ситуации правило отбора ИП: $E_i > 0,6$ (см. табл. 4 и 5, где N — количество различаемых пар авторов). Этому решающему правилу удовлетворили 4 параметра, уже представленные выше в табл. 2 и 3.

Таблица 4

Показатели корреляции и эффективности параметров

N	Параметр	$\bar{r}_i^m$	$\bar{r}_i^{n-m}$	E_i
7	X28	0,4163	0,2199	0,5281
5	X24	0,2620	0,1987	**0,7585**
	X42	0,4332	0,2465	0,5690
4	X09	0,3571	0,2261	**0,6330**
	X18	0,5159	0,3276	**0,6351**
	X26	0,4738	0,2802	0,5914
	X32	0,3960	0,2533	**0,6395**
	X53	0,4921	0,2822	0,5734
	X54	0,5069	0,2686	0,5300
	X55	0,5033	0,2699	0,5362

Таблица 5

Словарь ИП (N — мощность параметра)

Параметр	$\bar{r}_i^m$	$\bar{r}_i^{n-m}$	E_i	N
X09 (число подчиненных предложений первой степени)	0,3571	0,2261	0,6330	4
X18 (число служебных слов)	0,5159	0,3276	0,6351	4
X24 (число наречий)	0,2620	0,1987	0,7585	5
X32 (число подлежащих)	0,3960	0,2533	0,6395	4

3.4. Определение координат эталонов АК и РО — т. е. статистических характеристик этих АК и РО. На этом этапе сначала решается задача определения оптимального объема выборки. Для этого используется относительная стандартная ошибка оценки [2, с. 14—15]:

$$V_{\bar{x}} = \frac{V}{\sqrt{n}}\sqrt{1-f}, \tag{6}$$

где $V = \frac{\sigma}{x}$ — коэффициент разброса значений признака; $f = \frac{n}{N}$ — доля отбора (n — объем выборки, N — объем совокупности (класса)).

Так, по данным предварительной выборки из АК Ω_4 (А. С. Пушкин) были получены следующие значения $V_{\bar{x}}$: 0,16 (X09); 0,09 (X18); 0,11 (X24); 0,07 (X32). Как видно, ошибка оценки для X09 оказалась больше, чем для других параметров. То же самое наблюдалось при вычислении ошибки оценки для других АК. Определив заранее приемлемую величину $V_{\bar{x}} = 0,05$, мы вычислили необходимый объем выборки n:

$$n = \frac{N}{1 + \left(\frac{V_{\bar{x}}}{V}\right)^2 N}. \tag{7}$$

Для каждого АК итоговый объем выборки был установлен по значению n для ИП X09, так как по всем классам оно оказалось больше, чем для других ИП. Это позволило измерить ИП X09 с необходимым уровнем точности, а все остальные ИП с заведомо избыточным уровнем точности (см. табл. 6, где N — общий объем АК, n — итоговый объем выборок при $V_{\bar{x}} = 0,05$).

Далее методом случайного отбора были сделаны выборки. АК Ω_5 и все РО подверглись сплошному обследованию, так как их объемы оказались невелики. Координаты эталонов АК Ω_1, Ω_4 и интересующих нас РО приведены в табл. 7 (где N — итоговое количество исследованных предложений).

Таблица 6

Объем выборки для эталонов АК

Класс	N	$\bar{x}_{X09}$	σ_{X09}	$n_{V_{\bar{x}=0,05}}$
Ω_1 (П. А. Вяземский)	946	0,650	0,857	400
Ω_2 (Н. В. Гоголь)	613	0,650	0,821	312
Ω_3 (А. А. Дельвиг)	591	0,510	0,689	326
Ω_4 (А. С. Пушкин)	1001	0,330	0,551	527
Ω_5 (О. М. Сомов)	296	0,520	0,759	219

4. Моделирование системы классификации текстов

После того как координаты АК и РО определены, остается построить математическую модель системы распознавания текстов. Система предполагает реализацию *детерминированного* и *вероятностного алгоритмов распознавания* (ДАР и ВАР).

4.1. ДАР классифицирует РО в многомерном параметрическом пространстве (в нашем случае 4-мерном). В качестве классифицирующей функции снова изберем t-критерий Стьюдента с уровнем значимости $\alpha = 0,05$. Решение о принадлежности РО X_i некоторому АК принимается лишь тогда, когда наблюдаемое значение критерия (t_j) не превышает пороговое $(t_\alpha = 1,96)$ в каждом из n измерений параметрического пространства, поэтому решающее правило имеет вид:

$$\begin{cases} X_i \in \Omega_A, t_j^{A,X_i} \le t_\alpha, j = \overline{1,n}, \\ X_i \in \Omega_B, t_j^{B,X_i} \le t_\alpha, j = \overline{1,n}. \end{cases} \tag{8}$$

Использование $t_\alpha = 1,96$ здесь ориентировано на объем эталона самого большого класса (Ω_4 — 527 предложений) и повышение точности классификации. Поясним утверждение. Известно, что чем больше число степеней свободы f, тем ниже пороговое значение (а для $p = 0,95$ при f, стремящемся к бесконечности, пороговое значение, уменьшаясь, стремится к 1,96). Если мы хотим по некоторому параметру найти статистически существенные различия между двумя выборками, повышение порога для одного и того же значения f будет означать снижение вероятности ошибки. Следовательно, когда мы хотим выявить отсутствие таких различий (решаем обратную задачу), для уменьшения вероятности ошибки следует снижать порог. Тогда количество положительных решений по решающему правилу (8) уменьшится, т. е. снизится вероятность ложного срабатывания распознающей системы.

Таблица 7

Координаты эталонов АК Ω_1 и Ω_4 и ряда РО

Объект	N	ИП	$\bar{x}_i$	σ_i	Объект	N	ИП	$\bar{x}_i$	σ_i
Ω_1	400	X09	0,663	0,846	T23	13	X09	0,615	0,625
		X18	5,153	3,927			X18	2,923	2,165
		X24	1,305	1,448			X24	0,769	0,799
		X32	1,758	1,444			X32	1,385	0,487
Ω_4	527	X09	0,455	0,701	T26	2	X09	1,500	0,500
		X18	3,833	3,242			X18	2,500	0,500
		X24	1,070	1,326			X24	0,500	0,500
		X32	1,499	1,118			X32	2,000	0,000
T02	9	X09	0,444	0,497	T29a	16	X09	0,250	0,433
		X18	3,778	2,200			X18	5,750	5,379
		X24	0,111	0,314			X24	1,188	0,882
		X32	1,889	0,875			X32	2,000	1,768
T04	8	X09	0,500	0,500	T41	15	X09	0,733	1,482
		X18	3,375	1,728			X18	4,267	4,171
		X24	0,875	0,599			X24	0,933	1,806
		X32	1,250	0,433			X32	1,867	1,204
T15	9	X09	0,333	0,667	T43	85	X09	0,529	0,776
		X18	3,111	2,961			X18	3,776	2,763
		X24	1,222	1,397			X24	1,000	1,328
		X32	1,222	0,416			X32	1,482	0,876
T21	5	X09	1,000	0,894	T45	34	X09	1,029	1,403
		X18	8,000	10,040			X18	6,618	5,941
		X24	2,200	1,939			X24	1,765	1,848
		X32	1,800	1,470			X32	1,824	1,740

Классификация посредством ДАР носит итерационный характер. После определения принадлежности РО некоторому АК возникает модификация этого класса мощностью j, включающая в себя и исходный АК, и РО. Координаты эталонов таких классов вычисляются по формулам [2, с. 13]:

$$\bar{x}_i = \frac{(\bar{x}_1 n_1 + \bar{x}_2 n_2 +, ..., +\bar{x}_j n_j)}{\sum_{k=1}^{j} n_k}, \tag{9}$$

$$\sigma_i = \sqrt{\frac{\sum_{k=1}^{j} \sigma_k^2 n_k}{\sum_{k=1}^{j} n_k}}. \tag{10}$$

Так, по итогам первой итерации ДАР пять РО (Т04, Т15, Т23, Т29а, Т43) были отнесены к АК Ω_4, в результате чего возникла модификация этого класса Ω_4^1 с координатами (ср. с данными табл. 7): $j_{\Omega_4^1} = 6$ (один АК и пять РО); $n = 658$; $\bar{x}_{X09} = 0,462$, $\sigma_{X09} = 0,702$; $\bar{x}_{X18} = 3,839$, $\sigma_{X18} = 3,218$; $\bar{x}_{X24} = 1,058$, $\sigma_{X24} = 1,303$; $\bar{x}_{X32} = 1,500$, $\sigma_{X32} = 1,089$. Сходные изменения произошли с АК Ω_1 и Ω_2, причем класс Ω_1 дополнялся дважды, а одним из РО, отнесенных к нему, стал Т45.

4.2. Работа ВАР предполагает измерение таксономического расстояния между оставшимися РО и видоизмененными классами и последующую оценку полученных результатов.

Поскольку признаки, характеризующие АК и РО, измерены с помощью разных шкал (см. п. 3.1), в качестве функции расстояния между объектами целесообразно использовать взвешенное евклидово расстояние (ЕР):

$$d(a,b) = \left[\sum_{j=1}^{n} \omega_j \left(x_{aj} - x_{bj}\right)^2\right]^{0,5}, \tag{11}$$

где n — размерность евклидова пространства E_n, a и b — две точки в нем с координатами $a(x_{a1}, x_{a2}, ..., x_{an})$, $b(x_{b1}, x_{b2}, ..., x_{bn})$, а $\omega_j = 1/\sigma_j$ — весовой коэффициент j-й переменной, причем $\omega_j = 1$, если все признаки выражены в общих единицах измерения. Последнее условие можно выполнить путем стандартизации исходной матрицы данных с помощью среднеквадратичного отклонения, что равнозначно приведению всех параметров к единой шкале. В нашем случае исходная матрица данных Z образована координатами эталонов распознаваемых объектов и априорных (или видоизмененных по результатам ВАР) классов. Элементы стандартизованной матрицы данных X вычисляются по формуле [3, с. 76]:

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{\sigma_j}, i = \overline{1,N}, j = \overline{1,n}, \tag{12}$$

где z_{ij} — стандартизуемый элемент исходной матрицы данных Z, $\bar{z}_j$ — среднее значение j-го параметра матрицы Z, σ_j — квадратичное отклонение j-го параметра матрицы Z, $N = 43$ (число объектов в матрице Z: пять АК и 38 РО, оставшихся после ДАР), а $n = 4$ (число параметров в матрице Z).

Итак, на основе стандартизованной матрицы X определим взвешенные ЕР при $\omega_j = 1$. Далее по матрице взвешенных ЕР найдем вероят-

ности принадлежности РО каждому из классов [3, с. 58]:

$$P_{ji} = \frac{1}{d_{ji}} \left(\sum_{k=1}^{n} \frac{1}{d_{jk}} \right)^{-1}, \tag{13}$$

где d_{ji} — расстояние между j-м РО и i-м классом, d_{jk} — расстояние между j-м РО и k-м классом, n — число классов. Фрагменты матрицы X, матриц взвешенных ЕР $d(X_j, \Omega_i)$ и вероятностей принадлежности РО классам $P(X_j \in \Omega_i)$ приведены в табл. 8—9[1]. В этих таблицах мы уже не видим те РО, которые были включены в структуру классов на предыдущем этапе (см. п. 4.1).

Таблица 8

Фрагмент стандартизованной матрицы данных X

Объекты	Параметр			
	X09	X18	X24	X32
T02	-0,407	-0,339	-1,770	0,425
T21	1,491	2,601	2,475	0,227
T26	3,199	-1,229	-0,979	0,671
T41	0,579	0,001	-0,099	0,376

Отметим, что ВАР не чувствителен к свойствам распределений, определяющим условия применимости критерия Стьюдента (нормальность распределений и равенство их дисперсий), и может быть использован в качестве теста результатов ДАР. Для проведения теста мы заново построили стандартизованную матрицу данных, на этот раз включив в нее все объекты (общим числом 52: пять АК и 47 РО) — в том числе и те, что были распознаны на этапе работы ДАР. В двух случаях тест показал результаты, отличные от итогов ДАР, и в трех — от итогов

[1]В таблицах 8, 9 и 11, а также в пункте 5 статьи приводятся числовые данные, которые отличаются от данных, используемых в публикациях [5] и [6]. Причина различий: в указанных источниках описаны вычисления, сделанные по ошибочным исходным данным в части объекта T26. В текущей статье приведены исправленные исходные данные и результаты вычислений. Следствие указанных различий в данных: изменение решения ВАР по объекту T26 (см. табл. 9 и 10): по результатам публикаций [5] и [6] объект был атрибутирован А. А. Дельвигу (Ω_3) — в текущей статье с учетом исправлений объект атрибутируется Н. В. Гоголю (Ω_2). В остальном решения ВАР при классификации объектов не различаются.

В табл. 9 в связи с округлением значений вероятностей до трех десятичных разрядов их сумма не везде строго равна 1 без округления.

ВАР, основанного на итогах ДАР и учитывающего изменения координат эталонов классов (см. табл. 10, где ТА — тестируемый алгоритм).

Таблица 9

Фрагменты матриц взвешенных ЕР и вероятностей (X_j — объект)

X_j	$d(X_j, \Omega_i)$					$P(X_j \in \Omega_i)$				
	Ω_1^2	Ω_2^1	Ω_3	Ω_4^1	Ω_5	Ω_1^2	Ω_2^1	Ω_3	Ω_4^1	Ω_5
Т02	2,886	3,364	2,108	2,110	2,591	0,175	0,150	0,240	0,240	0,195
Т21	2,722	2,962	3,670	4,195	3,324	0,242	0,223	0,180	0,157	0,198
Т26	3,830	3,769	3,914	3,994	4,029	0,204	0,207	0,200	0,196	0,194
Т41	1,161	1,530	—	1,294	—	0,376	0,286	—	0,338	—

Таблица 10

Тест результатов ДАР и ВАР посредством ВАР

X_j	$P(X_j \in \Omega_i)$					ТА	Прежний результат
	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5		
Т02	0,185	0,148	0,239	0,234	0,194	ВАР	Нет решения
Т21	0,227	0,233	0,180	0,159	0,201	ВАР	Ω_1^2 (П. А. Вяземский)
Т26	0,206	0,206	0,200	0,193	0,194	ВАР	Ω_2^1 (Н. В. Гоголь)
Т29а	0,202	0,146	0,223	0,175	0,255	ДАР	Ω_4 (А. С. Пушкин)
Т45	0,385	0,387	—	0,228	—	ДАР	Ω_1 (П. А. Вяземский)

После определения искомых вероятностей требуется интерпретировать полученный результат. Обозначив i-й РО через X, естественным представляется записать решающее правило принадлежности этого РО j-му классу следующим образом:

$$X_i \in \Omega_j, P(X_i \in \Omega_j) > P(X_i \in \Omega_k), k = \overline{1,n}, k \neq j, \tag{14}$$

где n — число классов.

Ясно, что при такой формулировке правила пороговое значение вероятности принадлежности РО к некоторому классу априорно установлено быть не может. В аспекте определения межтекстовых расстояний более существенна не сама вероятность, а отношение $p1/p2$, или «оценка значимости вероятностей» [8, с. 15]. Атрибутируя текст пяти авторам, получим значения вероятностей пяти рангов. Так, по данным табл. 10,

для текста Т02 оценка значимости вероятности первого ранга (0,239) по отношению к вероятности второго (0,234) примерно равна 1,021. Применение всего набора оценок значимости вероятностей первого ранга затрудняет определение веса каждого решения ВАР. Для этой цели удобно применять среднее значение набора таких оценок по каждому решению ВАР, которое можно назвать *коэффициентом значимости принятого решения*:

$$\bar{k}_{P(X_i \in \Omega_j)} = \frac{1}{n-1} \sum \frac{P(X_i \in \Omega_j)}{P(X_i \in \Omega_k)}, X_i \in \Omega_j, k = \overline{1,n}, k \neq j, \qquad (15)$$

где $P(X_i \in \Omega_j)$ — вероятность первого ранга, n — число классов. Значение коэффициента тем выше, чем определеннее принятое решение.

5. Результаты работы распознающего автомата

Всего классификации были подвергнуты 46 текстов (47 объектов — см. выше п. 3). С учетом результатов теста работы ДАР с помощью ВАР были получены следующие результаты:

1. Классификацию пяти объектов следует признать неоконченной (см. табл. 10).

2. С вероятностью больше 0,95 А. С. Пушкину атрибутируется 4 текста (см. табл. 11), П. А. Вяземскому — 2 текста, Н. В. Гоголю — 1 текст.

3. С невысокой долей вероятности (от 0,210 до 0,568) определены авторы 35 объектов (за исключением Т02, Т21 и Т26, классификацию которых мы признали незавершенной). Для 10 объектов получен достаточно высокий коэффициент значимости принятого решения ($2,067 \leq \bar{k}_i \leq 5,859$), а для 25 — низкий ($1,069 \leq \bar{k}_i \leq 1,862$). К числу последних принадлежит и рассмотренный нами текст Т41 с коэффициентом значимости принятого решения 1,214 (см табл. 11). Однако надо заметить, что деление оценок на «высокие» и «низкие» произвольно — нет объективного инструмента для разделения этих оценок на зоны значимости.

6. Заключение

В заключение назовем задачи, решение которых, как нам видится, должно повысить надежность распознающей системы:

1. Поиск критерия, наиболее эффективного для классификации текстов.

2. Поиск оптимального метода анализа корреляционной матрицы связей параметров. Избранный метод определяет состав информативных параметров, следовательно, влияет на конечный результат.

Таблица 11

Общие результаты атрибуции по итогам работы ДАР и ВАР

Текст	Класс	$\bar{k}_i$
T04	Ω_4 (А. С. Пушкин)	Д
T15	Ω_4 (А. С. Пушкин)	Д
T23	Ω_4 (А. С. Пушкин)	Д
T41	Ω_1 (П. А. Вяземский)	1,214
T43	Ω_4 (А. С. Пушкин)	Д

3. Поиск такого инструмента, который позволил бы для каждой конкретной классификационной задачи определять *оптимальное количество* информативных параметров.

4. Определение порогового значения $\bar{k}_i$ при интерпретации результатов измерения межтекстового расстояния.

Требуется ряд экспериментов по классификации текстов, групповая принадлежность которых заранее известна и не подвергается сомнению. Важным условием является тесная стилистическая близость этих текстов, их «похожесть» друг на друга по разным параметрам языка и стиля. Такая постановка задачи позволит проверить, насколько эффективно и уверенно система атрибуции различает те малозаметные нюансы языка текстов, которые играют критически важную роль в решении сложных проблем авторства.

Список литературы

1. **Бонгард М. М.** Проблема узнавания. М.: Наука, 1967. 320 с.

2. В поисках потерянного автора: Этюды атрибуции / М. А. Марусенко, Б. Л. Бессонов, Л. М. Богданова и др. СПб.: Филол. ф-т С.-Петерб. гос. ун-та, 2001. 216 с.

3. **Марусенко М. А.** Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд-во ЛГУ, 1990. 168 с.

4. **Родионова Е. С., Хозяинов С. А., Митрофанова О. А.** Корпусы текстов в исследованиях по атрибуции литературных произведений // *Труды международной конференции «Корпусная линг-*

вистика — 2008». СПб.: С.-Петербургский гос. университет, Факультет филологии и искусств, 2008. С. 338—349.

5. **Хозяинов С. А.** Атрибуция публицистики, приписываемой А. С. Пушкину // *Прикладная и математическая лингвистика : материалы секции XXXVII Международной филологической конференции, 11-15 марта 2008 г., Санкт-Петербург / отв. ред. Т. Г. Скребцова. СПб.: Ф-т филологии и искусств СПбГУ, 2008. С. 20—30.*

6. **Хозяинов С. А.** Атрибуция публицистики, приписываемой А. С. Пушкину. Решение проблемы авторства методами распознавания образов / LAP LAMBERT Academic Publishing. Saarbrücken, 2012. 252 с.

7. **Хозяинов С. А.** Некоторые проблемы и методы квантитативно-структурного изучения авторских стилей // *Известия Российского государственного педагогического университета им. А. И. Герцена. 2008. № 28 (63). С. 378—383.*

8. **Якубайтис Т. А., Скляревич А. Н.** Вероятностная атрибуция типа текста по нескольким морфологическим признакам. Рига: ИЭВТ, 1982. 53 с.

СГУ им. Питирима Сорокина *Поступила 25.02.2017*

Summary

Khozyainov S. A. Text classification using methods of pattern recognition

This paper illustrates the text classification process using methods of pattern recognition. The problem of authorship of social and political essays attributed to A. S. Puskin is considered as an example. Means of increasing the reliability of the recognition system are suggested.
Keywords: text classification, methods of pattern recognition, authorship attribution, A. S. Puskin.

References

1. **Bongard M. M.** *Problema uznavaniya* (Recognition Problem), Moscow: Nauka, 1967, 320 p.

2. **Marusenko M. A., Bessonov B. L., Bogdanova L. M., Anikin M. A., Miasojedova N. E.** *V poiskakh poteryannogo avtora: Etyudy atributsii* (In search of the lost author. Studies in attribution), St. Petersburg: Faculty of Philology, Saint Petersburg University, 2001, 216 p.

3. **Marusenko M. A.** *Atributsiya anonimnykh i psevdonimnykh literaturnykh proizvedenii metodami raspoznavaniya obrazov* (Attribution of anonymous and pseudonymous literary works using methods of pattern recognition), Leningrad: Leningrad University, 1990, 168 p.

4. **Rodionova E., Khozyainov S., Mitrofanova O.** Text corpora in attribution of literary works, *Proceedings of the International Conference «Corpus Linguistics — 2008»*, St. Petersburg: St. Petersburg State University, Faculty of Philology and Arts, 2008, pp. 338—349.

5. **Khozyainov S. A.** Atributsiya publitsistiki, pripisyvaemoi A. S. Pushkinu (Attribution of social and political essays attributed to A. S. Puskin), *Prikladnaya i matematicheskaya lingvistika: Materialy sektsii XXXVII Mezhdunarodnoi filologicheskoi konferentsii*, 11—15 marta 2008 g., Sankt-Peterburg (Applied and mathematical Linguistics: Materials of the section XXXVII International philological conference, March, 11—15, St. Petersburg), St. Petersburg, 2008, pp. 20—30.

6. **Khozyainov S. A.** *Atributsiya publitsistiki, pripisyvaemoi A. S. Pushkinu. Reshenie problemy avtorstva metodami raspoznavaniya obrazov* (Attribution of social and political essays attributed to A. S. Puskin. Autorship attribution using methods of pattern recognition), LAP LAMBERT Academic Publishing, Saarbrücken, 2012, 252 p.

7. **Khozyainov S.** Some problems and methods of quantitative and structural research of authors' styles, *Izvestiya RGPU im. A. I. Gertsena*, № 28 (63), St. Petersburg, 2008, pp. 378—383.

8. **Yakubaitis T. A., Sklyarevich A. N.** *Veroyatnostnaya atributsiya tipa teksta po neskol'kim morfologicheskim priznakam* (Probability attribution of text type on the several morphological markings), Riga, 1982, 53 p.

Для цитирования: Хозяинов С. А. Классификация текстов методами распознавания образов // *Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика. 2017. Вып. 1 (22). С. 3–20.*

For citation: Khozyainov S. A. Text classification using methods of pattern recognition, *Bulletin of Syktyvkar University, Series 1: Mathematics. Mechanics. Informatics,* 2017, №1 (22), pp. 3–20.