ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 43 за 2022 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

М.Ю. Воронина, А.А. Кислицын, Ю.Н. Орлов

# Построение двухфакторных паттернов в задаче классификации текстов



***Рекомендуемая форма библиографической ссылки:*** Воронина М.Ю., Кислицын А.А., Орлов Ю.Н. Построение двухфакторных паттернов в задаче классификации текстов // Препринты ИПМ им. М.В.Келдыша. 2022. № 43. 24 с. https://doi.org/10.20948/prepr-2022-43
https://library.keldysh.ru/preprint.asp?id=2022-43

**О р д е н а  Л е н и н а**
**ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ**
**имени М.В.Келдыша**
**Р о с с и й с к о й  а к а д е м и и  н а у к**

**М.Ю. Воронина, А.А. Кислицын, Ю.Н. Орлов**

# Построение двухфакторных паттернов в задаче классификации текстов

**Москва — 2022**

**Воронина М.Ю., Кислицын А.А., Орлов Ю.Н.**

Построение двухфакторных паттернов в задаче классификации текстов

Построены двухфакторные паттерны эмпирических распределений частот биграмм для машинной классификации текстов по авторам и тематике. Атрибуты текста распознаются методом ближайшего соседа применительно к эталонным распределениям. Близость между распределениями понимается в смысле нормы в L1. Пара «автор-тема» неизвестного текста определяется как такая, к эталонному распределению которой тестируемый текст находится ближе всего. Анализируется проблема распознавания автора безотносительно темы текста и темы безотносительно автора. Исследуются также возможности укрупнения и детализации классификационных признаков.

***Ключевые слова:*** машинная классификация, текст, распределение биграмм, спектральный портрет, кластеризация

**Voronina M.Yu., Kislitsyn A.A., Orlov Yu.N.**

Two-factor patterns construction in problems of texts classification

Two-factor patterns of empirical distributions of bigram frequencies for machine classification of texts by authors and subject are constructed. Text attributes are recognized by the nearest neighbor method in relation to reference distributions. The proximity between distributions is understood in the sense of the norm in L1. The "author-topic" pair of an unknown text is defined as a nearest neighbor pattern. The problem of recognizing the author regardless of the topic of the text and the topic regardless of the author is analyzed. The possibilities of enlarging and detailing classification features are also being investigated.

***Keywords:*** machine classification, text, bigram distribution, spectral portrait, clustering

Работа выполнена при поддержке гранта РФФИ, проект № 19-29-01174

## Содержание

**Введение**

Во многих задачах машинной классификации актуальным является вопрос об уровне детализации информации, то есть о числе классов. В идеале желательно, чтобы информация была бы представлена в виде совокупности базисных классов, попарно не пересекающихся между собой и позволяющих однозначно идентифицировать структуру изучаемой системы. На практике же, когда нет априорных знаний о системе, такая задача не всегда может быть даже корректно сформулирована, поскольку проблема отчасти лежит в области философии: объект характеризуется «формой», то есть наблюдаемыми измеримыми параметрами, а также и «содержанием», которое гипотетически трактуется как определенная система, порождающая эти параметры. В зависимости от типа задач между «формой» и «содержанием» есть большая или меньшая корреляция, которая анализируется на стадии «обучения» алгоритма. На стадии же применения алгоритма «содержание» недоступно и должно быть определено по «форме». В этом смысле идентификация имеет два аспекта. Первый – надо распознать саму форму. Второй – сопоставить форме наиболее вероятное содержание. Тогда объект X считается распознанным как объект A, если некоторый набор определяемых параметров подходит под соответствующую классификацию.

Иногда задачи распознавания таковы, что оба аспекта не четко разделены. Например, картина условно неизвестного художника X распознана как картина художника A. Предположим, что картину написал художник B, имитируя манеру художника A. Сам же художник B собственных картин в манере художника A не писал. Какой ответ при распознавании авторства картины следует тогда считать правильным: опознание действительного автора данной конкретной картины или того, к кому это произведение ближе всего по некоторому перечню индикаторов? Это – типичная проблема двухфакторной идентификации.

В настоящей работе исследуется задача одновременной классификации произведений по авторам и по жанрам на примере корпуса литературных текстов на русском языке. На стадии «обучения» составляются эталонные распределения, по близости к которым и будет проводиться идентификация. Анализируется возможность объединения построенных двухфакторных эталонов по одному из факторов так, что новый однофакторный эталон также обладает идентифицирующим свойством. Вопрос о том, на какое количество однозначно идентифицируемых подклассов возможно разделить множество объектов, принадлежащих данному классу, является актуальной задачей как в методическом, так и в алгоритмическом планах.

Проблема неопределенности числа классов отмечается во многих исследованиях по кластерному анализу (см., напр. [1-5]). В зависимости от области приложения разные авторы используют различные методы оптимизации разбиения. Так, в работах [1, 2] используется метод

имитационного моделирования с последующим применением алгоритма k-means для классификации множества некоторых объектов как многомерных векторов. В работе [3] оптимальное количество кластеров ищется исходя из принципа минимального среднеквадратичного отклонения по получающимся в итоге фазовым состояниям системы. Различные метрические критерии устойчивости разбиения рассматриваются в [4, 5]. Центроидные и нецентроидные методы обсуждаются в [6], где строится также классификация указанных методов классификации. В [7] одним из авторов настоящей работы строится оптимальное равномерное разбиение гистограммы при аппроксимации выборочной плотности функции распределения временных рядов, в том числе нестационарных.

В целом многочисленность методов и критериев кластеризации свидетельствует о том, что пока еще нет универсального подхода к решению этой задачи. В частности, при идентификации атрибутов текстов на естественных языках (авторства, тематики, языка, жанра, эпохи написания) метод анализа существенно зависит от выбора атрибутов. Для определения жанра и эпохи эффективны семантические методы [8], а для определения автора и тематического направления – методы на основе анализа *n*-грамм и построения спектральных портретов [9-11].

Объектами, изучаемыми в настоящей работе, являются литературные тексты, написанные профессиональными писателями на русском языке или переведенными на русский язык, возможно, разными переводчиками. Это исследование основано на статистическом эксперименте с корпусом текстов на русском языке, результаты которого применительно к однофакторной классификации были описаны в [12]. В работе [12] изучалась эффективность метода биграмм и триграмм для идентификации автора текста. В продолжение этой работы мы изучим возможность двухфакторной идентификации и кластеризации обучающего корпуса по двум атрибутам – автору и тематике.

Идея статистического распознавания атрибутов текстов состоит в том, что предположительно существует некоторая генеральная совокупность частот символов, которая отвечает данному автору и данному тематическому направлению, а каждый конечный текст реализуется как выборка из этой совокупности.

Близость между текстами понимается как близость между эмпирическими распределениями частот (ЭРЧ) в смысле нормы в L1. Атрибутом неизвестного текста назначается тот, к эталону которого тестируемый текст находится ближе всего. Подчеркнем, что мы используем не прямой метод ближайшего соседа, когда атрибут неизвестного текста назначается по атрибуту ближайшего к нему текста, а метод ближайшего эталона, построенного как средневзвешенное распределение экспертно отобранных текстов. Основной исследуемый вопрос состоит в том, можно ли распознать автора, пишущего в разных жанрах, с помощью эталона биграмм. Это практический аспект работы. Теоретическая проблема – выбор одного из двух направлений распознавания: по дереву признаков или по их совокупности. В данном случае приоритет за вторым вариантом. Выяснилось, что если

сначала размещать тексты в некоторые крупные классы (жанры), куда входят многие авторы, а потом внутри крупного класса пытаться распознать мелкую структуру в виде отдельных авторов, то ошибка такого метода в разы превосходит ошибку распознавания с помощью двумерного эталона, сразу настроенного на двухфакторную особенность текста.

## 1. Паттерны биграмм для двухфакторной идентификации

Пусть $D^i_{a,t}(j)$ – распределение частот символа $j$ в $i$-ом тексте автора $a$, написанном на тему (в жанре) $t$. Под символом будем подразумевать $n$-грамму (букву, пару букв, тройку и т.п.). Пусть также $N^i_{a,t}$ есть количество указанных символов в данном $i$-м тексте. Двухфакторным паттерном (или эталоном) для идентификации автора и темы текста будем называть взвешенное эмпирическое распределение частот по всей совокупности произведений данного автора, принадлежащих обучающему корпусу текстов, экспертно отнесенных к определенной тематической категории:

$$F_{a,t}(j) = \frac{1}{N_{a,t}} \sum_i N^i_{a,t} D^i_{a,t}(j), \quad N_{a,t} = \sum_i N^i_{a,t}\,. \tag{1.1}$$

На данном этапе считаем, что существует библиотека текстов, относительно которых достоверно известны автор (один) и тема (одна). Суммирование по $i$ в формуле (1.1) проводится от 1 до $n_{a,t}$ – количества различных полных текстов (статей или книг) автора $a$ по теме $t$.

Формализуем задачу одновременной идентификации автора и темы неизвестного текста.

Пусть построены двухфакторные паттерны (1.1). Тогда идентификация автора и темы некоторого неизвестного, т.е. отсутствующего в библиотеке (обучающем корпусе), текста проводится на основе принципа ближайшего эталона. Это означает, что авторство $\alpha$ и тема $\tau$ неизвестного текста с ЭРЧ $D_{\alpha,\tau}(j)$ определяются как такая пара $(a,t)$, для которой

$$l(a,t) = \sum_j \left| F_{a,t}(j) - D_{\alpha,\tau}(j) \right| = \min\,. \tag{1.2}$$

Минимум ищется по всем парам $(a,t)$ параметров эталонов.

Предположим, что авторы и темы таковы, что построенные эталоны позволяют точно проводить классификацию текстов на основе (1.2). Возникает вопрос: могут ли литературные авторы (т.е. профессиональные писатели) быть идентифицированы по всей совокупности своих произведений без выделения специфических тем? С одной стороны, как показывают примеры, приведенные в [9, 13], это вполне возможно. С другой стороны, если среди произведений автора есть как литературные тексты, так и узкопрофессиональные научные работы, то объединение тех и других в один авторский эталон может привести к снижению точности идентификации текстов обоих типов. В настоящей работе мы формализуем условие возможности указанной кластеризации тем автора, а

также авторов темы так, что тексты сформированного кластера определяются безошибочно.

Если взвешенно просуммировать эталонные ЭРЧ (1.1) по различным темам, то получается совокупный эталон писателя по всем его произведениям:

$$f_a(j) = \frac{1}{K_a}\sum_t N_{a,t} F_{a,t}(j), \quad K_a = \sum_t N_{a,t}\,. \tag{1.3}$$

Проблема, которую мы изучаем, состоит в следующем. Будет ли после укрупнения тематик автора отдельный его текст $D_{\alpha,\tau}(j)$ на некоторую конкретную тему наиболее близок именно к укрупненной ЭРЧ (1.3) безотносительно темы произведения? То есть сохранится ли идентифицирующее свойство автора на укрупненной ЭРЧ:

$$L(a) = \sum_j \left| f_a(j) - D_{\alpha,\tau}(j) \right| = \min\,. \tag{1.4}$$

Аналогично можно построить совокупный по писателям эталон темы:

$$g_t(j) = \frac{1}{M_t}\sum_a N_{a,t} F_{a,t}(j), \quad M_t = \sum_a N_{a,t}\,. \tag{1.5}$$

Для этой статистики возникает задача определения тематической принадлежности того же самого произведения $D_{a,\tau}(j)$ по формуле минимизации следующего функционала:

$$\Lambda(t) = \sum_j \left| g_t(j) - D_{\alpha,\tau}(j) \right| = \min\,. \tag{1.6}$$

Представляется почти очевидным, что если две темы достаточно далеко отстоят одна от другой в смысле расстояния в L1 для двух сравниваемых авторов, причем авторы по каждой из тем отделяются один от другого, то при объединении обеих тем для каждого из авторов получим ЭРЧ, которое может потерять распознающее свойство относительно авторства текста, написанного, естественно, только по одной теме. Тем не менее желательно сформулировать условия, при которых авторский или тематический однофакторные паттерны (1.3) и (1.5), построенные на основе двухфакторных паттернов, сохраняют правильную различающую способность на том же множестве текстов.

В теоретико-вероятностном смысле основной аспект изучаемой задачи может быть сформулирован в следующем виде. Имеется некоторое множество случайных векторов, компоненты которых имеют заданные распределения и принадлежат единичному симплексу размерности $J$. Пусть эти векторы могут быть разбиты на непересекающиеся $K$ множеств так, что каждый вектор из любого данного множества в определенной норме находится ближе к среднему вектору своего множества, чем к среднему вектору другого множества. Выберем некоторое количество $k$, $1 < k < K$, этих множеств и объединим их в одно множество. Какова вероятность того, что новая совокупность из $K - k + 1$ множеств будет обладать тем же свойством близости своего элемента к среднему вектору множества, что и первоначальная совокупность?

Решение этой задачи позволит сформулировать условие на отделимость тематических направлений внутри творчества одного автора в рамках метода

сравнения эмпирических распределений *n*-грамм. Подчеркнем, что в идеале автор в целом (т.е. по всей совокупности своих произведений) должен отделяться как от отдельных тематических направлений других авторов, так и от их усредненных паттернов по соответствующей совокупности произведений.

## 2. Распределение расстояний от текста до авторского паттерна

Рассмотрим задачу сравнения расстояний между текстами и эталонами для авторов, пишущих в рамках нескольких заданных тем. Пусть для каждого автора и темы построены двухфакторные паттерны $F_{a,t}(j)$. Получим формулу распределения расстояний между отдельным текстом и эталоном в предположении, что известны распределения $\psi^{j}_{aa,tt}(x)$ отклонения $x$ эмпирической частоты $j$-го символа в отдельном произведении автора $a$ на тему $t$ от эталонной вероятности того же символа для этого же автора и той же темы, и аналогично распределения $\psi^{j}_{ab,tt}(x)$, $\psi^{j}_{aa,t\tau}(x)$, $\psi^{j}_{ab,t\tau}(x)$. Эмпирически эти распределения строятся по статистикам отклонений $x^{i}_{j} = D^{i}_{\alpha,\tau}(j) - F_{a,t}(j)$ частот символа $j$ в $i$-ом произведении каждого автора. Величинам же $\left|x^{i}_{j}\right|$ отвечают распределения

$$\varphi^{j}_{ab,t\tau}(x) = \psi^{j}_{ab,t\tau}(x) + \psi^{j}_{ab,t\tau}(-x). \qquad (2.1)$$

Расстояние между некоторым текстом с распределением $D_{\alpha,\tau}(j)$ и произвольным эталоном $F_{a,t}(j)$ определяется формулой (1.2). Поскольку же оба распределения нормированы на единицу, то

$$l(a,t) = \sum_{j=1}^{J}\left|F_{a,t}(j) - D_{\alpha,\tau}(j)\right| = \sum_{j=1}^{J-1}\left|F_{a,t}(j) - D_{\alpha,\tau}(j)\right| + \left|\sum_{j=1}^{J-1}\left(F_{a,t}(j) - D_{\alpha,\tau}(j)\right)\right|. \qquad (2.2)$$

Как известно, если даны плотности распределений $Y(u)$ и $Z(v)$ величин $u$ и $v$, то плотность распределения $\rho(l)$ величины $l = u + v$ дается формулой

$$\rho(l) = \int Y(u)Z(l-u)du. \qquad (2.3)$$

Интеграл в (2.3) символизирует численное нахождение свертки по имеющимся эмпирическим данным для входящих в нее распределений.

Получим формулу для распределения $Y(u)$ первой суммы в (2.2). Опуская временно для краткости индексы авторства и темы текстов, введем величины накопленных отклонений и их модулей для распределений символов $j$:

$$z_k = \sum_{j=1}^{k} x_j, \;\; y_k = \sum_{j=1}^{k}\left|x_j\right|. \qquad (2.4)$$

Поскольку тогда $y_{J-1} = y_{J-2} + \left|x_{J-1}\right|$, то в силу (2.1) и (2.3) распределение $Y_{J-1}\left(y_{J-1}\right)$ дается формулой

$$Y_{J-1}(y_{J-1}) = \int_0^{y_{J-1}} Y_{J-2}(s)\varphi_{J-1}(y_{J-1}-s)ds. \qquad (2.5)$$

В результате получаем, что плотность распределения суммы модулей отклонений имеет вид $J-2$-кратной свертки:

$$Y_{J-1}(y_{J-1}) = \int_0^{y_{J-1}} \varphi_{J-1}(y_{J-1}-s_{J-2}) \int_0^{s_{J-2}} \varphi_{J-2}(s_{J-2}-s_{J-3})... \qquad (2.6)$$

$$\int_0^{s_3} \varphi_3(s_3-s_2) \int_0^{s_2} \varphi_2(s_2-s_1)\varphi_1(s_1)ds_1 ds_2 ... ds_{J-2}.$$

Аналогично плотность распределения $V(z_{J-1})$ величины накопленного отклонения дается формулой непосредственно в терминах распределений накопленных отклонений по отдельным символам:

$$V(z_{J-1}) = \int_{-\infty}^{\infty} V_{J-2}(s)\psi_{J-1}(z_{J-1}-s)ds = ... =$$

$$= \int_{-\infty}^{\infty} \psi_{J-1}(z_{J-1}-s_{J-2}) \int_{-\infty}^{\infty} \psi_{J-2}(s_{J-2}-s_{J-3})... \qquad (2.7)$$

$$\int_{-\infty}^{\infty} \psi_3(s_3-s_2) \int_{-\infty}^{\infty} \psi_2(s_2-s_1)\psi_1(s_1)ds_1 ds_2 ... ds_{J-2}.$$

Тогда плотность распределения модуля величины накопленного отклонения равна

$$Z(|z_{J-1}|) = V(z_{J-1}) + V(-z_{J-1}). \qquad (2.8)$$

Подстановка (2.6) и (2.8) в (2.3) решает задачу об определении распределения расстояний между текстом и эталоном при известных распределениях отклонений эмпирических частот от своих предположительных генеральных значений. Однако непосредственно эти распределения не позволяют оценить ошибку идентификации атрибутов текста, поскольку такая задача является операцией парного сравнения, для чего требуется знать совместное распределение расстояний до своего и чужого эталонов.

Для безошибочной идентификации автора и темы текста требуется, чтобы совместное распределение расстояний от текста до своего и до чужого эталонов имело бы в каждой из трех базисных плоскостей нижний треугольный носитель. А именно, для текста с атрибутами $(a,t)$ расстояние до своего эталона должно быть всегда меньше, чем до чужого, хотя сами по себе распределения расстояний отдельно до своего и отдельно до чужого эталонов вполне могут иметь общий носитель. Требуется же, чтобы $l_{aa,tt} < l_{ab,tt}$, $l_{aa,tt} < l_{aa,t\tau}$, $l_{aa,tt} < l_{ab,t\tau}$.

Если бы рассматриваемые случайные величины – расстояния до своего эталона и до чужого эталона – были бы независимы, то для оценки ошибки идентификации достаточно было бы ограничиться рассмотрением общей части

распределений $\rho_{aa,tt}(l)$, $\rho_{ab,tt}(l)$, $\rho_{aa,t\tau}(l)$, $\rho_{ab,t\tau}(l)$. Как показывают эксперименты, результаты которых приведены в [13], при таком подходе ошибка ожидалась бы на уровне примерно 0,5, тогда как по факту она имеет порядок 0,1. Следовательно, между указанными распределениями существует значимая корреляция.

Вычислить аналитически интегралы в (2.6), (2.7) затруднительно даже при относительно простых модельных предположениях о виде распределений $\psi(x)$. Интегралы в (2.7) могут быть вычислены для устойчивых распределений (нормальное распределение, гамма-распределение), но выражения в (2.6) аналитически вычисляются только при малых $J$ (2 или 3), тогда как, например, для биграмм $J = 33^2 = 1089$. Поэтому для качественного анализа удобно задать непосредственно распределения расстояний между текстом и эталонами в виде подходящих аппроксимаций.

Например, если аппроксимировать распределение расстояний между ЭРЧ текста и ЭРЧ своего эталона нормальным распределением с нулевым средним и дисперсией $\sigma^2_{aa,tt}$, то получим

$$\rho_{aa,tt}(l) = \frac{2}{\sqrt{2\pi}\sigma_{aa,tt}} \exp\left( -\frac{l^2}{2\sigma^2_{aa,tt}} \right), \ l \geq 0. \tag{2.9}$$

Для распределения расстояний от текста той же темы до эталона чужого автора естественно принять аппроксимацию нормальным распределением с некоторым ненулевым средним $\mu_{ab,tt}(j)$ и дисперсией $\sigma^2_{ab,tt}(j)$ (и аналогично для эталонов разных тем). В таком случае распределение расстояний до чужого эталона будет иметь вид:

$$\begin{aligned} \rho_{ab,tt}(l) &= \frac{1}{\sqrt{2\pi}\sigma_{ab,tt}} \left( \exp\left( -\frac{\left(l - \mu_{ab,tt}\right)^2}{2\sigma^2_{ab,tt}} \right) + \exp\left( -\frac{\left(-l - \mu_{ab,tt}\right)^2}{2\sigma^2_{ab,tt}} \right) \right) = \\ &= \frac{2}{\sqrt{2\pi}\sigma_{ab,tt}} \exp\left( -\frac{l^2 + \mu^2_{ab,tt}}{2\sigma^2_{ab,tt}} \right) ch\left( \frac{l\mu_{ab,tt}}{\sigma^2_{ab,tt}} \right), \ l \geq 0. \end{aligned} \tag{2.10}$$

Пусть между расстояниями до своего $x$ и до чужого $y$ эталонов имеется корреляция с коэффициентом $r$. Тогда из (2.9), (2.10) следует, что совместное распределение указанных расстояний с учетом коррелированности дается формулой (индекс $t$ темы для краткости опущен):

$$R(x,y)=\frac{1}{2\pi\sigma_{aa}\sigma_{ab}\sqrt{1-r^2}}\exp\left(-\frac{1}{2\left(1-r^2\right)}\left(\frac{x^2}{\sigma_{aa}^2}+\frac{y^2+\mu_{ab}^2}{\sigma_{ab}^2}\right)\right)\times$$

$$\left(\exp\left(-\frac{2rxy}{\left(1-r^2\right)\sigma_{aa}\sigma_{ab}}\right)ch\left(\frac{\mu_{ab}y}{\left(1-r^2\right)\sigma_{ab}^2}+\frac{2r\mu_{ab}x}{\left(1-r^2\right)\sigma_{aa}\sigma_{ab}}\right)+\right. \tag{2.11}$$

$$\left.+\exp\left(\frac{2rxy}{\left(1-r^2\right)\sigma_{aa}\sigma_{ab}}\right)ch\left(\frac{\mu_{ab}y}{\left(1-r^2\right)\sigma_{ab}^2}-\frac{2r\mu_{ab}x}{\left(1-r^2\right)\sigma_{aa}\sigma_{ab}}\right)\right),\quad x\ge 0,\quad y\ge 0.$$

Формула (2.11) получается из общего вида совместной плотности нормального распределения двух случайных величин

$$\varphi(x,y)=\frac{1}{2\pi\sqrt{\Delta}}\exp\left(-\frac{1}{2\Delta}\left(\sigma_y^2(x-\bar{x})^2+\sigma_y^2(y-\bar{y})^2-2C(x-\bar{x})(y-\bar{y})\right)\right),$$

где $\Delta=\sigma_x^2\sigma_y^2-C^2$, $C=r\sigma_x\sigma_y$ – коэффициент ковариации, $\bar{x}=0,\ \bar{y}=\mu_{ab}$. Записывая выражение $\varphi(x,y)$ для пар $(x,y),(-x,y),(x,-y),(-x,-y)$ и складывая, приходим к (2.11).

В этом случае вероятность ошибки, то есть того, что $x>y$, будет даваться интегралом

$$P\{x>y\}=\int_0^{+\infty}dx\int_0^{x}R(x,y)dy. \tag{2.12}$$

Задавая значение вероятности в левой части (2.12), получаем интегральное условие на параметры $r,\mu,\sigma$. Однако даже в этом модельном примере аналитически может быть вычислен только внутренний интеграл в (2.12), который представляет собой сумму выражений вида

$$\int_0^{x}e^{-A^2(x)y^2+B(x)y}dy=\frac{\sqrt{\pi}}{2A}e^{(B/2A)^2}\left(erf\left(Ax-B/(2A)\right)-erf\left(B/(2A)\right)\right), \tag{2.13}$$

где $A(x)$ и $B(x)$ – некоторые известные линейные функции от $x$. Комбинация $B/(2A)$ в терминах выражений, входящих в (2.11), имеет вид $\pm r\dfrac{\sigma_{ab}}{\sigma_{aa}}x+\mu_{ab}$. Условие сходимости внешнего интеграла по $x$ в (2.13) состоит в отрицательности коэффициента при $x^2$ в экспоненте, что дает

$$r^2\left(1-r^2\right)\sigma_{ab}^2<1. \tag{2.14}$$

При стремлении коэффициента корреляции к единице связь между расстояниями $x$ и $y$ стремится к линейной: $y\approx x+\mu_{ab}$. В результате тексты своего автора с высокой вероятностью становятся отделимы от чужого эталона, поскольку максимумы распределений «свой-чужой» сдвинуты на $\mu_{ab}$. Отметим, что для модели расстояний в виде нормальных распределений вероятность ошибки не равна нулю.

Проведенный качественный анализ позволяет сделать вывод о допустимости использования ЭРЧ *n*-грамм для создания паттернов и

идентификации авторов и тем достаточно больших текстов. В следующем разделе мы рассмотрим вопросы, связанные с кластеризацией паттернов.

### 3. Кластеризация паттернов

Рассмотрим расстояние между некоторым текстом с распределением $D_{\alpha,\tau}(j)$ и авторскими эталонами (1.3) библиотеки тестируемого корпуса. Имеем следующую цепочку равенств:

$$L(a)=\sum_j\left|f_a(j)-D_{\alpha,\tau}(j)\right|=\sum_j\left|\frac{1}{K_a}\sum_t N_{a,t}F_{a,t}(j)-D_{\alpha,\tau}(j)\right|=$$

$$=\frac{1}{K_a}\sum_j\left|\sum_t N_{a,t}F_{a,t}(j)-K_aD_{\alpha,\tau}(j)\right|=\frac{1}{K_a}\sum_j\left|\sum_t N_{a,t}F_{a,t}(j)-D_{\alpha,\tau}(j)\sum_t N_{a,t}\right|= \quad (3.1)$$

$$=\frac{1}{K_a}\sum_j\sum_t N_{a,t}\left|F_{a,t}(j)-D_{\alpha,\tau}(j)\right|=\sum_t\frac{N_{a,t}}{K_a}l(a,t)\,.$$

Поскольку мы считаем, что по каждой теме написано достаточно произведений для создания соответствующего двухфакторного эталона, то количество знаков $N_{a,t}$ достаточно для того, чтобы считать эталон стационарной генеральной совокупностью. Тогда без ухудшения точности распознавания можно ограничить обучающий корпус текстов одинаковым для всех авторов количеством знаков: это $N^*=\min\limits_{a,t}N_{a,t}$. Пусть $T_a$ – полное число тем, по которым имеются произведения данного автора. Тогда используемое в этом ограниченном корпусе количество символов в произведениях автора $a$ на разные темы равно $K_a=N^*T_a$. В результате получаем, что (3.1) запишется в виде

$$L(a)=\frac{1}{T_a}\sum_t\sum_j\left|F_{a,t}(j)-D_{\alpha,\tau}(j)\right|=\frac{1}{T_a}\sum_{t=1}^{T_a}l(a,t). \quad (3.2)$$

Таким образом, расстояние от текста до однофакторного авторского эталона равно среднему расстоянию от того же текста до двухфакторных авторских эталонов. Аналогично расстояние между текстом и тематическим эталоном (2.5) есть

$$\Lambda(t)=\sum_j\left|g_t(j)-D_{\alpha,\tau}(j)\right|=\sum_j\left|\frac{1}{M_t}\sum_a N_{a,t}F_{a,t}(j)-D_{\alpha,\tau}(j)\right|$$

$$=\frac{1}{M_t}\sum_a\sum_j N_{a,t}\left|F_{a,t}(j)-D_{\alpha,\tau}(j)\right|. \quad (3.3)$$

Далее для ограниченного вышеуказанным способом корпуса получаем $M_t=N^*A_t$, где $A_t$ – число авторов, пишущих на заданную тему. Тогда

$$\Lambda(t) = \frac{1}{A_t} \sum_{a=1}^{A_t} l(a,t) \ . \qquad (3.4)$$

В результате получаем, что для укрупненных классов расстояние от исследуемого текста до однофакторного эталона равно среднему по второму фактору расстоянию от этого текста до двухфакторных эталонов. Следовательно, если средние расстояния (3.2) и (3.4) минимальны при тех же значениях одного из параметров $a$ и $t$, что и идентификаторы двухфакторного эталона автора и темы данного текста, то идентифицирующее свойство укрупненных эталонов сохраняется.

Без упрощающего предположения об ограничении корпуса текстов формулы (3.1) и (3.3) означают построение средневзвешенных расстояний – соответственно для автора по его разным темам и для темы по разным авторам.

## 4. Практические примеры

Рассмотрим некоторые варианты удачной и неудачной кластеризации тематик и/или авторов в более крупные классы с целью снижения ошибки идентификации принадлежности отдельного текста данному классу. Для целей настоящей работы были отобраны авторы, которые имеют различные более или менее узкие тематические серии. Определенная сложность отбора таких примеров состоит в том, что автор каждого текста известен точно, тогда как тематика текста определяется неоднозначно и представляет точку зрения эксперта. Тем самым оценка ошибки тематической идентификации, вообще говоря, субъективна. В этом смысле интересна сама возможность тематической идентификации с малой ошибкой. Представляется естественным, что если тема достаточно узкая, то она довлеет над автором, и в рамках такой темы авторские эталоны имеют низкую точность. Таковым является, например, определенное научное направление: допустим, историческое развитие теории вероятностей. И наоборот, если тематика более широкая – история науки или, более общо, история как таковая, то низкую точность имеют тематические эталоны, построенные по совокупности авторов, пишущих в этих жанрах.

Применительно к литературным текстам жанр также имеет довольно субъективное толкование. Мы здесь трактуем жанр не в смысле литературной формы произведения (роман, эссе, очерк и т.п.), для которой принят определенный стиль изложения, а как тематику произведения. В соответствии с такими «сюжетами» книги расставляются, например, на полках книжных магазинов или в соответствующих рубриках электронных библиотек. В этом случае весьма сложно построить объективный эталон жанровой подборки. Например, можно ввести категорию «детектив», понимая под этим описание работы полиции или частного сыска, или в более широком смысле – некий сюжет, затрагивающий криминальную сторону жизни. Тогда в такой категории существуют довольно большие подкатегории: политический детектив, иронический детектив, социально-психологический детектив, фантастический детектив, мемуары бухгалтеров коммерческих компаний, собственно

криминалистика (список легко продолжить). Не имея целью проводить литературоведческое исследование, мы приведем здесь ряд результатов pro et contra методов кластеризации.

Пример 1. Рассмотрим формальное объединение близких эталонов писателей в один совместный эталон. Этот прием весьма важен в контексте использования эвристического метода ближайшего соседа для кластеризации объектов. В работе [12] был рассмотрен корпус из 1783 текстов 100 авторов на русском языке. Авторы по жанрам не распределялись. Были построены авторские эталоны биграмм вида (1.3) для цели идентификации автора текста методом кросс-валидации. Ошибка оказалась на уровне 12 %. Объединим теперь близких авторов, взяв в качестве инструмента поиска такой близости граф ближайших соседей среди авторских эталонов. Тогда можно попробовать объединить в один класс авторов, чьи эталоны оказались взаимно близки. Процесс объединения итерационный. Изначально такой граф для рассматриваемого корпуса текстов состоял из 16 несвязных фрагментов, из которых 4 фрагмента содержали 2 вершины, 3 фрагмента – 3 вершины, 3 фрагмента – 4 вершины, 1 фрагмент из 9 вершин, 2 фрагмента по 10 вершин, 1 фрагмент из 12 вершин и 2 фрагмента по 14 вершин. Поскольку каждый фрагмент графа первых ближайших соседей заканчивается циклом длины 2, позволяющим выделить пару взаимно ближайших авторских эталонов, то на первом этапе отбора такие пары можно объединить каждую в свой кластер. Оставшиеся 68 эталонов снова перегруппировываются, после чего опять находятся взаимно ближайшие пары. Таким способом можно провести попарное объединение авторов. Если предположить теперь, что близкие эталоны достаточно далеко отстоят от других, то совокупный эталон такой пары будет, видимо, обладать более высоким идентифицирующим свойством. Во-первых, в нем будет примерно в два раза больше текстов, чем у каждого из авторов в отдельности, то есть он будет более достоверен как генеральная совокупность. Во-вторых, если все тексты будут верно приписаны таким парным эталонам, то ошибка последующего распознавания автора внутри пары сведется всего лишь к ошибке, когда второй эталон ошибочно признается за первый. Таких примеров в указанном корпусе текстов было 7 % от числа ошибок, т.е. реально менее 1 % книг. Тем самым могла бы открыться возможность почти безошибочного распознавания автора. Конечно, с другой стороны, если при объединении эталонов их носители увеличиваются и начинают более сильно перекрываться с носителями других пар, то точность идентификации может и понизиться. Именно это и произошло в данном эксперименте: ошибка распределения текстов по взаимно близким парным эталонам составила 19 %, то есть увеличилась почти в два раза по сравнению с исходным тестированием. Это означает, что формальное, а не экспертное объединение по некоторому принципу близости в данной задаче привело к значительному ухудшению точности распознавания. Этот результат можно трактовать и иначе: если изначально были некоторые крупные эталоны, не позволявшие провести идентификацию достаточно точно, то путем выделения из них подклассов можно получить существенное улучшение качества

распознавания. Итак, это пример того, что «укрупнять плохо, мельчить хорошо». Также это важное свидетельство того, что существуют задачи, в которых метод ближайшего соседа не вполне адекватен для кластеризации объектов.

Пример 2. Рассмотрим теперь противоположную ситуацию, когда «укрупнять хорошо, мельчить плохо». В качестве примера возьмем детективные произведения Дарьи Донцовой. Будучи интегрированной в корпус [12], о котором шла речь в примере 1 выше, она идентифицировалась весьма успешно: только один ее текст из примерно ста отобранных был распознан неверно. С другой стороны, в ее творчестве имеются разные сюжетные линии в соответствии с главными действующими лицами: «Евлампия Романова», «Виола Тараканова» и ряд других. Если выделить их в самостоятельные классы, то выясняется, что примерно 30 % текстов каждого класса идентифицируются как тексты Донцовой, но из другой подборки. При этом три текста были вообще неверно идентифицированы не только по подклассу, но и по автору. Следовательно, выделение подкласса не всегда эффективно. Можно построить распределение расстояний от текстов до полного эталона Донцовой и убедиться, что это распределение имеет весьма узкий носитель. Распределения же расстояний от текстов Донцовой до эталонов подклассов более широкие, что и приводит к появлению ошибок.

Отметим, однако, что есть примеры текстов, когда вполне хорошо идентифицируются и авторы с широким носителем. Проблема в том, что метод идентификации не абсолютный в лингвистическом смысле, а зависящий от корпуса текстов. В одном случае соседи автора таковы, что надо укрупнить тематику, а в другом – что надо, наоборот, выделить подклассы.

Пример 3. Бывают также ситуации, когда укрупнение или выделение классов не принципиально. Рассмотрим творчество Агаты Мэри Клариссы леди Маллоуэн, которая под псевдонимом Агаты Кристи опубликовала порядка 60 детективных романов, а также шесть «обычных» романов под псевдонимом Мэри Уэстмакотт. Статистика показала, что среди авторов выбранного корпуса [12] на роль автора этих шести романов претендует только Агата Кристи. Далее, собственно детективы Агаты Кристи можно разбить на три серии по «главным следователям»: «Пуаро», «Марпл», «Прочие: Рейс, Томми и Таппенс, и др.». Если каждую из этих серий трактовать как написанную отдельным автором, то выясняется, что все три группы произведений могут быть распознаны без ошибок, т.е. произведения с Пуаро написаны тем же автором, кто написал и другие тексты с этим действующим лицом, а произведения с мисс Марпл также написаны своим автором (отличным от автора серии с Пуаро). В то же время объединение всех трех серий в один эталон также дает безошибочное распознавание их общего автора. Но означает ли это, что Кристи действительно несколько изменяла манеру письма в зависимости от серии, оставаясь тем не менее собой в глобальном смысле? Оказалось, что если заменить в надлежащих романах словосочетание «мисс Марпл» на «Эркюль Пуаро», то тексты с «бывшей Марпл» на 50 % будут относиться к эталону «Марпл», а на 50% - к эталону «Пуаро». Следовательно, эти серии различаются

не тематически или стилистически, а только наименованием главных лиц, что и приводит к определенным изменениям в распределении буквосочетаний. В этом смысле Кристи, по сути, Донцова (или, хронологически, наоборот).

Пример 4.

Другой тип укрупнения класса распознавания относится к авторам, которые пишут тексты на существенно разные темы. Тогда детализация признаков «автор-жанр» приводит к улучшению классификации по сравнению с объединенным эталоном автора. Мы приведем результаты анализа для 15 авторов, 220 текстов которых распределены по следующим темам. Айзек Азимов (позитронные роботы, прочая фантастика, научно-популярные книги); Борис Акунин (Фандорин, остальные тексты); Кир Булычев (Алиса, Гусляр, исторические романы); Николай Гоголь (деревня, город); Роджер Желязны (Амбер, остальная фантастика); Александр Казанцев (космос, Арктика); Артур Конан Дойл (Холмс, приключения); Станислав Лем (Ийон Тихий, остальная фантастика); Сергей Лукьяненко (дозоры, остальная фантастика); Александр Мазин (инквизитор, варяг, варвары, дракон); Жорж Сименон (Мегрэ, прочие романы); Рекс Стаут (Вульф, прочие романы); Лев Толстой (романы, повести, религия и нравственность); Барбара Хэмбли (Дарвет, Эшер, драконы); Алексей Яшин (публицистика, философия). Все двумерные эталоны «автор-тема», кроме Яшина (о чем ниже), дали точное распознавание произведений в рамках этого корпуса, тогда как при объединении всех произведений автора в один эталон примерно 5 % текстов данного корпуса были идентифицированы неверно. То есть литературные тематические направления достаточно хорошо отделяются, тогда как с текстами иной специфики возникают проблемы.

На примере Алексея Яшина интересно сравнить тематические эталоны литературных авторов, являющихся в то же время учеными определенных специальностей. Отметим, что тексты по техническим наукам очень сильно отличаются от литературного творчества, поэтому не удается построить хорошо работающий эталон «писателя-математика» или «писателя-физика (химика и т.п.)». Но есть достаточно много научных направлений, труды по которым носят в той или иной степени литературный характер: это история, философия, политология, культурология, искусствоведение. В определенной степени многим художественным произведениям присущи аспекты указанных наук. Насколько тогда отличаются, допустим, научные труды историка от его же художественных исторических романов, и можно ли объединить те и другие в один авторский эталон? Детальный анализ в этой области требует специального исследования, мы же дадим одну иллюстрацию, обладающую, на наш взгляд, определенной общностью.

Творчество российского писателя-прозаика и ученого в области так называемой «биоинформатики» Алексея Афанасьевича Яшина насчитывает несколько десятков публицистических повестей и рассказов, а также около десятка крупных научных работ, тематика которых отражена в их названиях: «Глобализация как ноосферный процесс», «Феноменология ноосферы» и т.п. Не обсуждая здесь, насколько понятия «биоинформатика, глобализация и ноосфера» корректно определены с формально научной точки зрения,

рассмотрим возможность разделения творчества этого автора. Оказалось, что из построенных двух авторских эталонов (научного и публицистического) только публицистический эталон обладает полным идентифицирующим свойством. Научные же труды на 30% были отнесены к эталонам «политология» и «культурология», построенным в [10] по совокупности работ соответствующих ученых без учета трудов Яшина, а не к «биоинформатике». Следовательно, принадлежность научных трудов определенному классу в гуманитарных областях только отчасти обусловлена тематической спецификой, влияние самого автора может быть весьма заметным. Впрочем, возможно, что определенные научные труды Яшина на самом деле относятся к политологии, а их априорное отнесение к иным областям (даже, возможно, сделанное самим автором), является ошибочным. Заметим, что в технических науках все работы хорошо идентифицируются по теме, а не по автору. Если же теперь объединить в один авторский эталон все произведения Яшина, то каждое из них идентифицируется точно методом кросс-валидации. Таким образом, разделение творчества на литературу и «гуманитарную науку» может привести к ухудшению точности распознавания автора по сравнению с объединенным корпусом его текстов.

Пример 5. Рассмотрим теперь более сложный пример двухфакторной идентификации. Имеются две группы авторов (точнее, авторш), пишущих в жанрах «иронический детектив» и «сентиментальный роман». К первому корпусу отнесены четыре группы текстов Натальи Александровой в виде тематических серий «Наследники Остапа Бендера» (28 текстов), «Надежда Лебедева» (27 текстов), «Три подруги» (9 текстов) и прочие произведения числом 20; также в этот корпус включены Светлана Алешина (19 текстов), Валентина Андреева (10 текстов), Наталья Андреева (11 текстов), Наталья Борохова (13), Сандра Браун (8), Татьяна Полякова (45), Иоанна Хмелевская (53), Елена Яковлева (7), и три тематических группы Дарьи Донцовой «Даша Васильева» (36), «Евлампия Романова» (28), «Виола Тараканова» (25). Во второй корпус вошли: Александра Авророва (12 текстов), Барбара Босуэл (9), Вирджиния Браун (12), опять же Сандра Браун (45 текстов), Хелен Брукс (13), Барбара Брэдфорд (12), Ронда Бэйс (5), Татьяна Веденская (9), Джулия Грайс (6), Кристин Григ (10), Лаура Гурк (9), Барбара Делински (8), Джулия Джеймс (6), Луис Ламур (95), Миранда Ли (18). Таким образом, мы рассматриваем 16 детализированных подгрупп в каждом жанре.

Задача 1. Идентификация автора и подтемы внутри каждого жанра. Подтема относится в данном случае только к первому жанру. Из 339 текстов первого корпуса правильно идентифицированы 280. Из 59 неверно определенных классов принадлежности текстов почти все ошибки относятся к неверной идентификации тематической рубрики внутри одного и того же автора: эти ошибки, которые фактически не являются ошибками идентификации автора, составили 49 текстов: 15 Донцовой и 34 Александровой. Общая ошибка по первому корпусу составила примерно 0,17. Среди 269 текстов второго жанра неверно опознаны 40, что составляет около 0,15 от текстов этого корпуса. Следовательно, здесь мы имеем ошибки двух

видов. Во-первых, неправильно указан автор текста. Во-вторых, неверно указана тематическая подгруппа, хотя автор указан правильно. Отметим, что вторая ошибка фактически не приводит к неверному указанию автора.

В результате получаем, что средневзвешенная ошибка определения автора по двум этим корпусам равна (59+40)/(269+339)=0,16.

Проведем далее объединение обоих корпусов в один, сохранив авторские подтемы, т.е. фактически увеличим число авторов. Задача та же – определить тексты, принадлежащие выделенным классам. Ошибка оказалась равной 0,18, то есть несколько увеличилась по сравнению с 0,16. Неверно идентифицированных текстов стало 108 вместо 99. Из девяти добавленных ошибок пять обусловлены новыми авторскими эталонами, а четыре связаны с тем, что в таком объединенном корпусе оказалось две Сандры Браун, которые, как оказалось, плохо различаются между собой (в переводах).

Таким образом, если подтемы автора близки между собой и находятся реально в одном жанре, то их разделение в виде эталонов ухудшает качество распознавания.

Задача 2. Идентификация автора безотносительно подтемы. В первом корпусе можно провести объединение всех текстов Александровой и, соответственно, Донцовой, и рассмотреть ту же задачу, что и выше. Это – стандартная задача определения автора по его эталону, которая решалась в [12, 13]. В результате число ошибок в первом (детективном) корпусе снизилось до 0,03. Следовательно, объединенные авторские эталоны в данном примере приводят к более высокой точности, а разделение творчества на близкие серии неэффективно.

Однако отметим, что улучшение качества распознавания именно автора при объединении подтем не позволяет увидеть более тонкую структуру, состоящую в данном случае в вариации серий внутри жанра. Серии при таком подходе статистически не выделяются.

Если теперь, как и в задаче 1, объединить два корпуса текстов, где авторы присутствуют своими авторскими эталонами без разбивки на подтемы, то общая ошибка идентификации автора текста становится равной 0,09.

Задача 3. Рассмотрим определение жанра произведения безотносительно автора. Для этого объединим все тексты первого корпуса в эталон «иронический детектив», а все тексты второго – в эталон «сентиментальный роман». После этого объединим все тексты обоих корпусов в один корпус и попытаемся его разделить на жанровые компоненты в соответствии с близостью к эталону жанра, согласно формуле (3.4). Жанровая ошибка идентификации текстов оказалась равной 0,05. Следовательно, можно считать, что для данных двух тематических рубрик понятие эталона жанра существует.

При этом ошибочно определяются преимущественно тексты жанра «сентиментальный роман»: ошибка по ним составила 0,04, то есть их доля в ошибке составляет 80 %.

После разделения текстов по жанрам попробуем сделать вторичную классификацию – по авторам. При этом неверно идентифицированные тексты по жанрам считаем исходной ошибкой. Для первого корпуса таких ошибочных

текстов 6, а для второго – 26. Оставшиеся жанрово верно определенные тексты были распределены по авторским эталонам с ошибкой соответственно 10 текстов (первый корпус) и 38 текстов (второй корпус). Суммарная ошибка такой двухэтапной идентификации составила: по первому корпусу 0,05, а по второму корпусу 0,24. В итоге средневзвешенная ошибка составила по обоим корпусам величину 0,13. Это больше, чем величина 0,09 ошибки одноэтапной идентификации автора. Следовательно, увеличение этапов распознавания приводит к увеличению ошибки. С одной стороны, это представляется естественным, поскольку каждый этап порождает свою ошибку. С другой стороны, можно было бы ожидать, что предварительная классификация окажется точной, поскольку она отвечает укрупненным параметрам. Однако этого не произошло. Как уже говорилось выше, это связано с тем, что при усреднении нескольких авторов даже в пределах одной, но литературной, а не научной тематики, происходит расширение носителя распределения расстояний от текста до своего эталона, причем возрастает область перекрытия распределения расстояний до эталонов «свой-чужой», что и порождает дополнительную ошибку.

## 5. Анализ спектральных портретов матриц условных биграмм

С ЭРЧ в виде биграмм связан еще один метод классификации, который использует информацию о спектре матрицы условных вероятностей $P_{ij}$ того, что в тексте следом за символом $i$ идет символ $j$. Переформатируем биграмму «$ij$» в матрицу $F_{ij}$, отвечающую эмпирической вероятности данного символа. Введем также однобуквенное ЭРЧ

$$f_j = \sum_i F_{ij} \tag{5.1}$$

и определим условную вероятность

$$P_{ij} = F_{ij} / f_j . \tag{5.2}$$

Из (5.1), (5.2) следует, что $f_j$ является собственным вектором для матрицы $P_{ij}$, принадлежащим собственному значению $\lambda = 1$. Другие собственные значения лежат в некоторых областях в комплексной плоскости, причем точность их расположения обусловлена устойчивостью спектра матрицы $P_{ij}$ к малому возмущению ее элементов.

В терминах резольвенты $R(\lambda) = (\lambda I - P)^{-1}$ матрицы $P$ $\varepsilon$-спектр определяется следующим образом: число $\lambda$ принадлежит $\varepsilon$-спектру $\Lambda_\varepsilon(P)$ матрицы $P$, если

$$\|R(\lambda)\| \geq \frac{1}{\varepsilon \|P\|} . \tag{5.3}$$

При исследовании расположения точек спектра удобно рассматривать замкнутые гладкие кривые $\gamma_\varepsilon$, представляющие изолинии $\varepsilon$-спектра. Контур

$\gamma_\varepsilon$ разбивает весь $\varepsilon$-спектр $\Lambda_\varepsilon(P)$ на две части – лежащие внутри и вне его. Тем самым $\gamma_\varepsilon$ осуществляет дихотомию $\varepsilon$-спектра матрицы. Качество дихотомии $\kappa_\gamma(P)$ оценивается нормой квадрата резольвенты (9) на данной кривой:

$$\kappa_\gamma(P) = \frac{\|P\|^2}{l_\gamma} \oint_\gamma \|R(\lambda)\|^2 d\lambda . \qquad (5.4)$$

Таким образом, спектральный портрет – это представление собственных значений матрицы в виде зон, где могут находиться эти собственные значения при заданной точности элементов матрицы. В работах [10, 13] метод спектральных портретов был применен для построения индикатора принадлежности научного текста определенному тематическому кластеру. В настоящей работе мы интересуемся возможностью создания эталонного спектрального портрета жанра для целей идентификации текста. Естественно, такой эталон эффективен только для случая, когда жанры могут быть разделены безотносительно авторов отдельных текстов. Поэтому ниже рассмотрен последний пример 5 из предыдущего раздела, когда оказалось возможным разделить тексты иронического детектива и сентиментального романа только по жанрам с весьма малой ошибкой 0,05, что почти в два раза меньше, чем ошибка распознавания авторов текстов.

Ниже на рис. 1-6 представлены спектральные портреты матриц в следующей очередности. Сначала на рис. 1-2 приведены портреты матриц условных биграмм, характеризующих жанровый эталон в целом, т.е. эти матрицы построены по всем текстам, входящим в данный жанр. Отмечаются различия в формах центральных областей, что и позволяет различать жанры один от другого.

Затем на рис. 3-4 показан спектральный портрет одной из авторш, которые пишут в соответствующих направлениях. Цель сравнения – проверить, что эти авторские портреты близки каждый к своему жанровому эталону.

Наконец, на рис. 5-6 даны спектральные портреты отдельного текста каждого из жанров тех же авторш. Хотя они и не наследуют специфику жанра спектрального портрета, тем не менее выяснилось, что спектры отдельных текстов отличаются от своего жанра менее, чем от чужого.

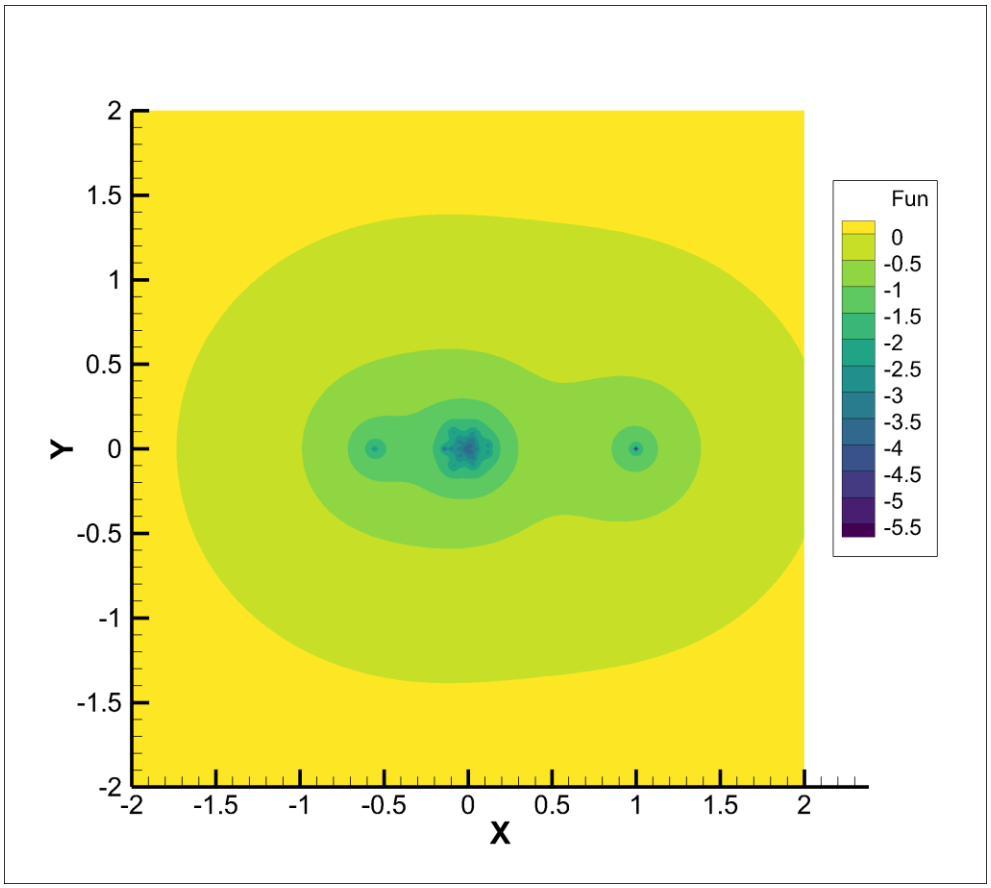


Рис. 1 – Спектральный портрет темы «сентиментальный роман»

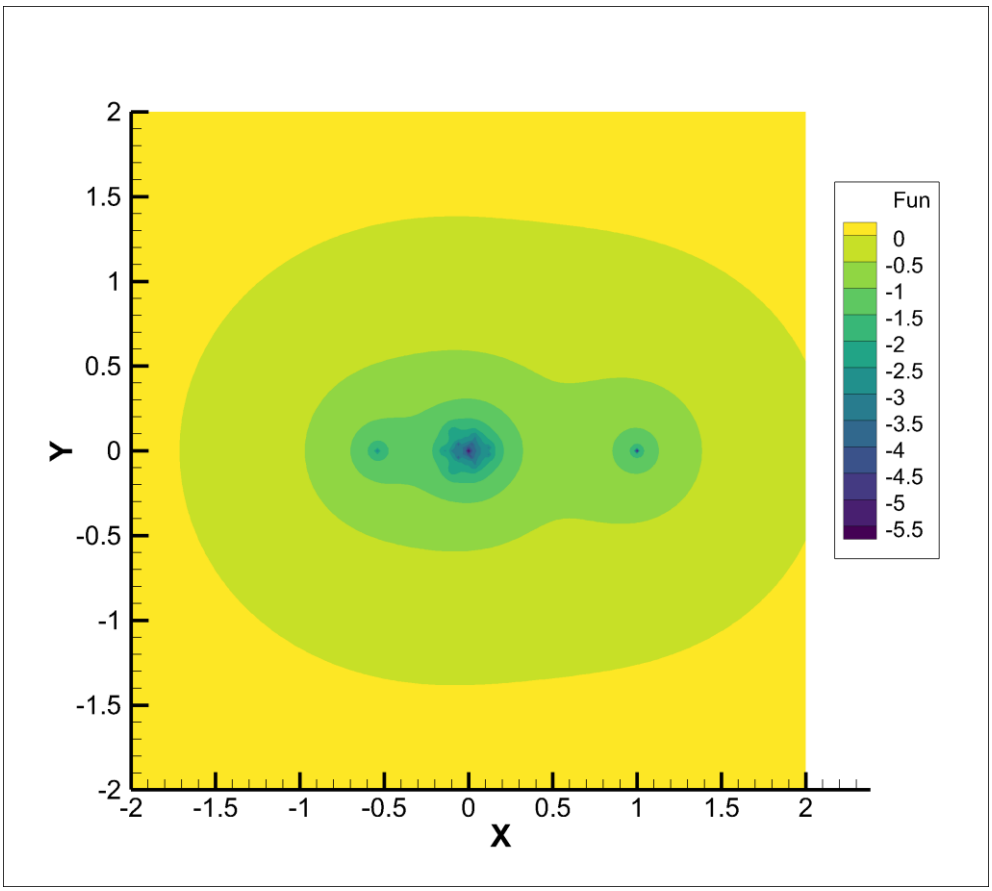


Рис. 2 – Спектральный портрет темы «дамский детектив»

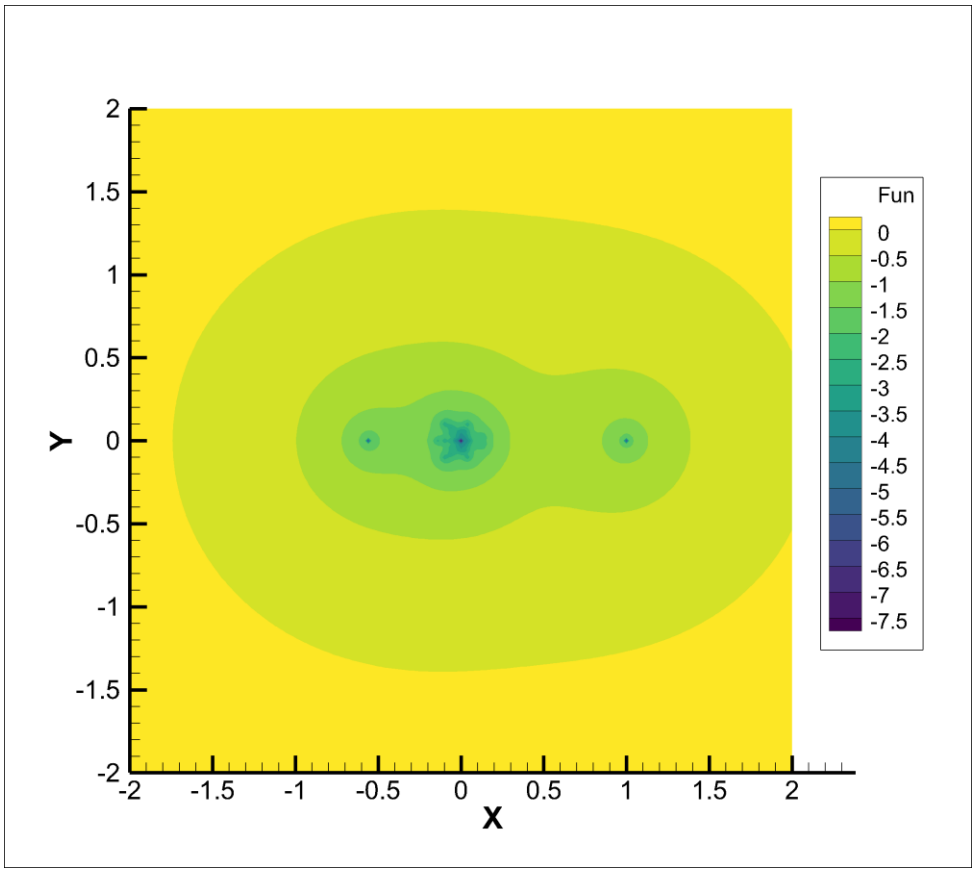


Рис. 3 – Спектральный портрет автора из темы «сентиментальный роман»

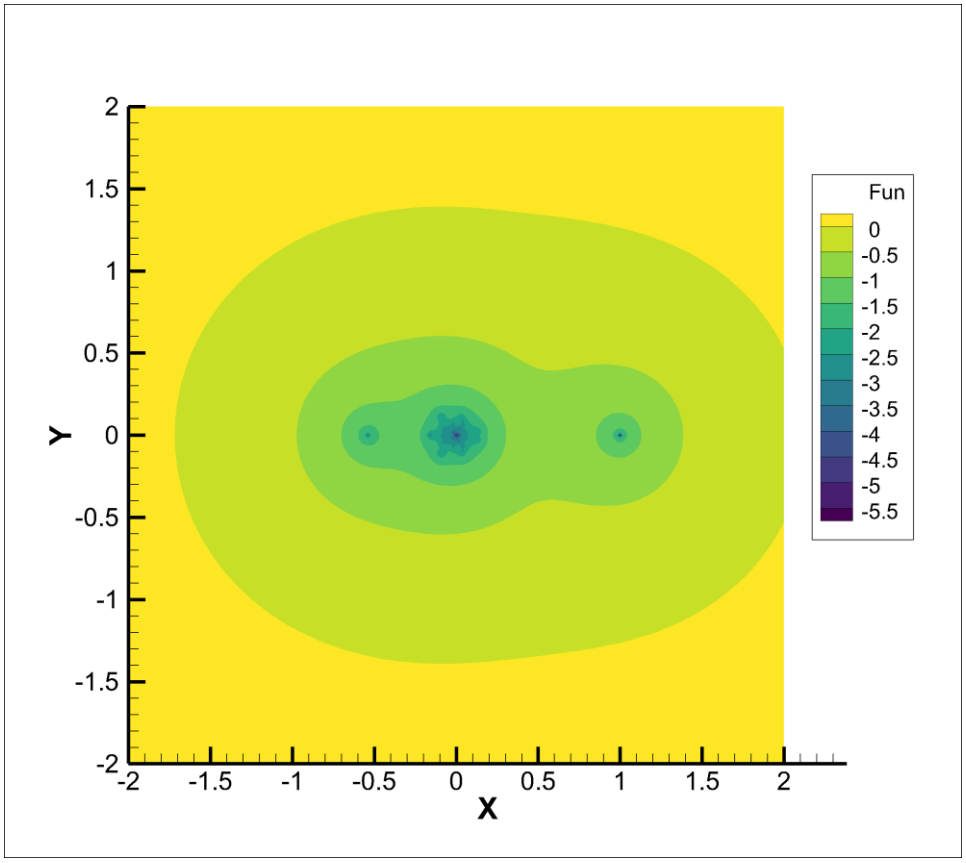


Рис. 4 – Спектральный портрет автора из темы «дамский детектив»

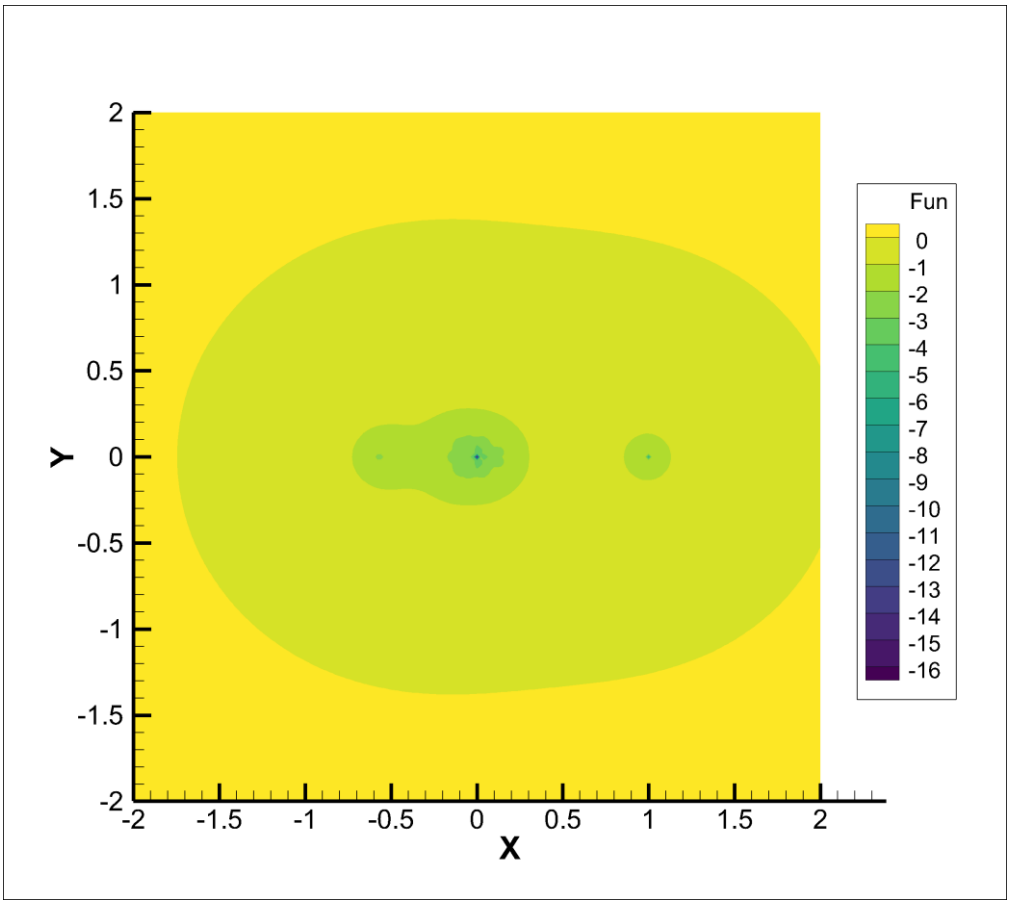


Рис. 5 – Спектральный портрет текста из темы «сентиментальный роман»

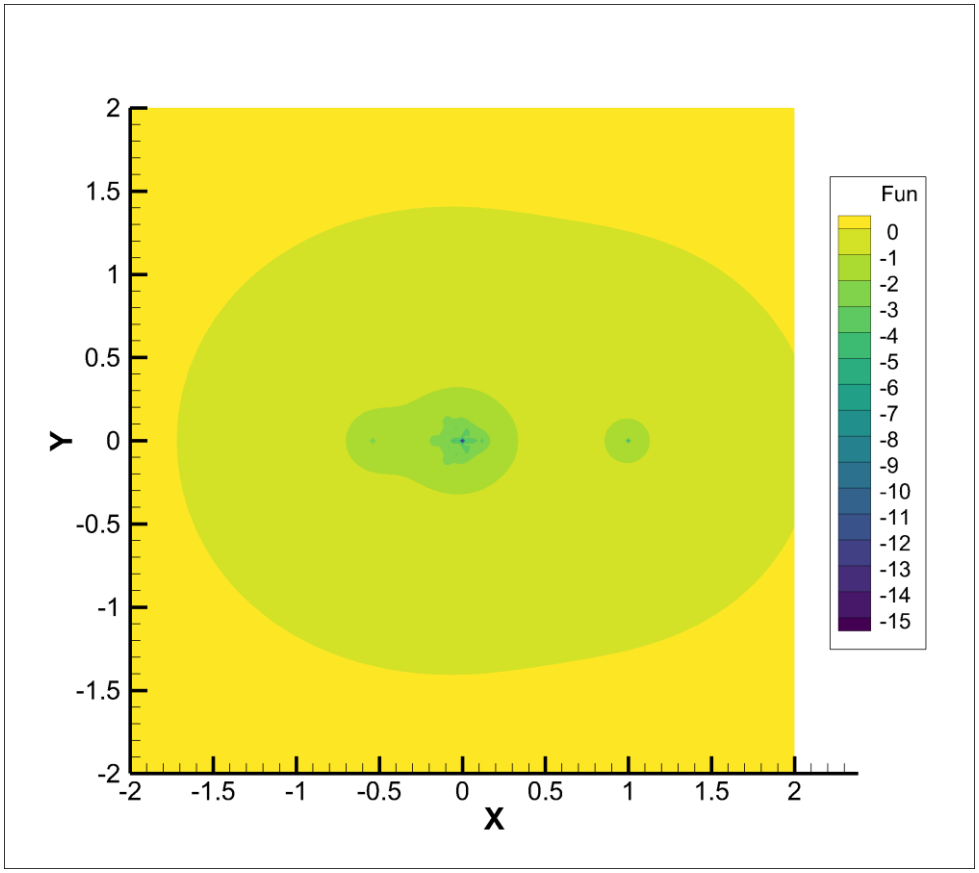


Рис. 6 – Спектральный портрет текста из темы «дамский детектив»

Основные различия в спектрах относятся к форме области во внутренней части круга радиусом примерно 0,3. На рис. 1 имеются три луча, исходящие из центральной точки, тогда как на рис. 2 имеется в основном центральное пятно. На авторских портретах на рис. 3 и 4 эти особенности сохранены, и на рис. 4 наблюдаются не лучи, а фактически небольшими выступы. Характерно и то, что точность, требуемая для позиционирования собственных значений, повышается с уменьшением длины текста, что статистически обоснованно.

**Заключение**

В данной работе исследована возможность двухфакторной идентификации автора и темы литературного произведения методом построения соответствующего паттерна как эмпирического распределения частот символов. Выяснилось, что априорный синтез двухфакторного паттерна в однофакторный не приводит к улучшению идентификации, равно как и выделение в однофакторном паттерне нескольких составляющих. Это означает, что методы кластеризации близких объектов не универсальны, а зависят от неизвестных заранее свойств этих объектов, то есть являются эвристическими.

В то же время существуют ситуации, когда такие методы, обоснование которых дается постфактум, весьма точны. Важным результатом работы является демонстрация изначально не очевидной возможности классификации текстов по жанрам на основе близости эталонных спектральных портретов матриц условных биграмм, отвечающих выбранным жанрам.

Введение написано Ю.Н. Орловым, раздел 1 написан Ю.Н. Орловым и А.А. Кислицыным, разделы 2, 3 и 4 написаны М.Ю. Ворониной, раздел 5 написан А.А. Кислицыным.

**Литература**

1. Фролов В.В., Слипченко С.Е., Приходько О.Ю. Метод расчета числа кластеров для алгоритма k-means // Экономика. Информатика. 2020. Т.47. № 1. С. 213-225.

2. Rezaei M., Fränti P. Set-matching measures for external cluster validity // IEEE Transactions on Knowledge and Data Engineering. 2016. V. 28. No. 8. P. 2173-2186.

3. Кольцов С.Н. Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования // Письма в ЖТФ. 2017. Т. 43. № 12. С. 90-95.

4. Шалымов Д.С. Рандомизированный метод определения количества кластеров на множестве данных // Научно-технический вестник Санкт-Петербургского государственного университета информационных технологий, механики и оптики. 2009. № 5. С. 111-116.

5. Ложкинс А., Буре В.М. Вероятностный подход к определению локально-оптимального числа кластеров // Вестник СПбГУ. Серия 10. Прикладная математика. Информатика. Процессы управления. 2016. № 1. С. 28-37.

6. Елизаров С.И., Куприянов М.С. Проблема определения количества кластеров при использовании методов разбиения // Изв. вузов. Приборостроение. 2009. Т. 52. № 12. С. 3-8.

7. Орлов Ю.Н. Оптимальное разбиение гистограммы для оценивания выборочной плотности распределения нестационарного временного ряда // Препринты ИПМ им. М.В. Келдыша. 2013. № 14. 26 с.

8. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы, 2017. Т. 30. № 1. С. 85-99.

9. Орлов Ю.Н., Осминин К.П. Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика, 2010. Т. 26. № 2. С. 95-108.

10. Митин Н.А., Орлов Ю.Н. Статистический анализ биграмм специализированных текстов // Компьютерные исследования и моделирование, 2020. Т. 12. № 1. С. 243-254.

11. Данилов Г.В., Жуков В.В., Куликов А.С., Макашова Е.С., Митин Н.А., Орлов Ю.Н. Сравнительный анализ статистических методов классификации научных публикаций в области медицины // Компьютерные исследования и моделирование, 2020. Т. 12. № 4.

12. Кислицын А.А., Орлов Ю.Н. Исследование статистик графов ближайших соседей // Препринты ИПМ им. М.В. Келдыша. 2021. № 85. 23 с.

13. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 312 с.