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

Препринты ИПМ • Препринт № 29 за 2023 г.

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

А.А. Кислицын, М.Ю. Кислицына

Распознавание выборочных распределений среди системы эталонов: метод ближайшего соседа

Рекомендуемая форма библиографической ссылки: Кислицын А.А., Кислицына М.Ю. Распознавание выборочных распределений среди системы эталонов: метод ближайшего соседа // Препринты ИПМ им. М.В.Келдыша. 2023. № 29. 21 с. https://doi.org/10.20948/prepr-2023-29
https://library.keldysh.ru/preprint.asp?id=2023-29

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

А.А. Кислицын, М.Ю. Кислицына

Распознавание выборочных распределений среди системы эталонов: метод ближайшего соседа

Москва — 2023

Кислицын А.А., Кислицына М.Ю.

Распознавание выборочных распределений среди системы эталонов: метод ближайшего соседа

Изучается величина ошибки идентификации выборочного распределения многомерной дискретной случайной величины среди библиотеки эталонных генеральных совокупностей в зависимости от размерности случайного вектора, длины выборки и расстояния между двумя эталонными распределениями в нормах C и L1. Показано, что ошибка распознавания в норме L1 существенно ниже, чем в C. В качестве практического применения рассмотрены эталонные распределения *n*-грамм для текстов художественной литературы. Выяснилось, что точность идентификации в основном определяется индивидуальными особенностями эталонов, а не расстояниями между ними. Разработан алгоритм для тестирования системы эталонов на точность распознавания.

Ключевые слова: распознавание образов, выборочное распределение

Kislitsyn A.A., Kislitsyna M.Yu.

Recognition of sample distribution functions among a system of patterns: the nearest neighbor method

The value of the sample distribution identification error of a multidimensional discrete random variable among a library of reference patterns is studied, depending on the dimension of the random vector, the sample length and the distance between two reference distributions in the norms C and L1. It is shown that the recognition error in the L1 norm is significantly lower than in C. Reference distributions of n-grams for texts are considered as a practical application. It turned out that the accuracy of identification is mainly determined by the individual characteristics of the standards, and not by the distances between them. An algorithm has been developed to test the system of standards for recognition accuracy.

Keywords: pattern recognition, sample distribution

Содержание

1. Введение 3
2. Пример распознавания автора литературного произведения 5
3. Постановка задачи о распознавании выборочного распределения 9
4. Статистическая точность оценки эталонных частот 11
5. Результаты вычислительных экспериментов 13
6. Тестирование заданной системы эталонов на точность 18
7. Заключение 21
Литература 21

1. Введение

При практическом применении метода ближайших соседей в задачах распознавания образов точность распознавания обусловлена тем, насколько адекватно построена система классификации образов. Цель распознавания – установление соответствия между отдельным объектом и классом.

Задача распознавания состоит в том, чтобы по измеряемым значениям параметра $x \in X \subset R^n$, характеризующим объект, определить, к какому классу $s \in S \subset R^m$ этот объект принадлежит. Состояния s непосредственно не измеримы. Таким образом, распознавание – это отображение $V : X \to S$. Соответствующая функция $s = V(x)$ называется решающим правилом или управляющим функционалом.

Если решающее правило $s = V(x)$ априори не очевидно, анализируется доступный массив данных для статистического оценивания условной вероятности $P(x|s)$ того, что при наблюдении значения x система находится в состоянии s. Предполагая, что оцененные вероятности правильно отражают скрытую от наблюдателя зависимость $s(x)$, решающее правило выбирается на основе байесовского принципа максимальной вероятности [1]. Именно, считается, что наблюдаемому значению x отвечает то значение состояния s, для которого $P(x|s) = \max$. Этот подход минимизирует ошибку распознавания, понимаемую как долю неверно распознанных состояний.

Однако на практике сложно реализовать сбор данных для оценивания вероятностей $P(x|s)$ на требуемом уровне значимости. Во-первых, измеряемые показатели могут не образовывать стационарный временной ряд. Во-вторых, параметров может быть так много, что собрать нужный статистический материал невозможно из-за ограниченности самого наблюдаемого материала. Поэтому для распознавания часто используется метод ближайшего соседа: неизвестный объект относится к тому же классу, что и ближайший к нему. Функция близости выбирается в соответствии с модельной задачей. Например, если объект представляет собой функцию распределения, построенную по выборке из N данных, то расстояние можно выбрать в норме C непрерывных функций (что есть некоторое приближение, так как в численном эксперименте функции распределения имеют скачки), либо в норме L1 суммируемых плотностей (гистограммное расстояние). Могут быть и другие варианты. В данной работе объектами являются выборочные распределения – функции распределения или вероятности состояний.

Говоря об использовании метода ближайших соседей для задачи распознавания образов, следует представлять, каких практических результатов в смысле точности можно ожидать, если требуется распознать, из какой генеральной совокупности взята та или иная выборка. Формально эта задача сводится к применению статистического критерия, играющего роль расстояния.

Например, если имеется набор эталонных распределений, то вычисляются критериальные статистики для анализируемой выборки и каждого из эталонов, после чего выбирается тот, для которого значение критерия минимально. Если вариантов выбрать ближайший эталон всего два, то можно обсуждать, какова вероятность того, что распределение выборки будет ближе к одному из эталонов. Однако при наличии большего количества эталонов задача теоретического определения вероятности правильной идентификации становится трудно решаемой, хотя формальные условия на вероятность соответствующих событий могут быть выписаны.

Сложность анализа связана с тем, что выборка из некоторого многомерного эталонного распределения располагается где-то в R^d, и потому точность распознавания зависит не только и даже не столько от того, как далеко она находится от своего эталона, сколько от количества других эталонов, находящихся на примерно таком же расстоянии от данного эталона. Поэтому, если требуется определить, с какой вероятностью некоторая выборка может быть правильно распознана среди заданной системы эталонных распределений, требуется вычислительный эксперимент именно для заданной системы эталонов. Обычно на этот аспект обращается мало внимания, но следует подчеркнуть, что каждая система эталонных распределений (или библиотека эталонов) перед своим применением должна быть протестирована на априорную точность распознавания выборочного распределения в зависимости от длины выборки. Важно, что результат такого тестирования будет иметь значение только для данной системы эталонов, поскольку точность распознавания зависит также и от взаимного расположения эталонов. Система из других эталонов будет обладать, вообще говоря, другой разрешающей способностью.

В этой связи важной практической задачей является определение вероятности правильного распознавания выборки в зависимости от ее длины, числа эталонов и функции распределения расстояния между эталонами. В работе представлен численный алгоритм, позволяющий провести такое тестирование системы эталонных распределений на точность идентификации выборок, и приведен практический пример его применения для задачи идентификации автора текста по распределению буквосочетаний. Модель идентификации описана в работах [2, 3].

Вопрос, который возникает в контексте распознавания автора текста, состоит в данном случае в том, с какой вообще точностью можно ожидать распознавание своего эталона для выборки определенной длины, если в качестве альтернатив рассматриваются некоторые заданные распределения. Одной из целей настоящей работы является демонстрация того, что индивидуальные отличия между эталонами и их число являются не менее важными атрибутами библиотеки, чем функция распределения расстояний между эталонами.

2. Пример распознавания автора литературного произведения

В работах [4-6] исследовалась точность распознавания автора текста методом биграмм. Суть метода распознавания состояла в следующем. По известным текстам каждого автора на этапе обучения составлялся авторский эталон из частот двухбуквенных сочетаний. Далее на этапе тестирования брался текст одного из авторов, для которого был построен авторский эталон, но этот текст не использовался на этапе обучения. Вычислялись расстояния в норме L1 между распределением биграмм данного текста и всеми авторскими эталонами, которые были построены для отобранного корпуса текстов. Эталон, отвечающий минимальному расстоянию, назначался на роль авторского для тестируемого текста. Для корпуса из 100 авторов и примерно 2000 текстов ошибка распознавания таким методом составила 0,12.

На рис. 1 приведено совместное распределение расстояний «свой-чужой» для текстов данного корпуса. По осям отложены номера классовых интервалов, ширина которых отвечает расстоянию 0,06. Поскольку почти весь носитель этого распределения расположен справа от побочной диагонали квадрата, то это означает, что расстояние до своего эталона в основном меньше, чем до чужого.

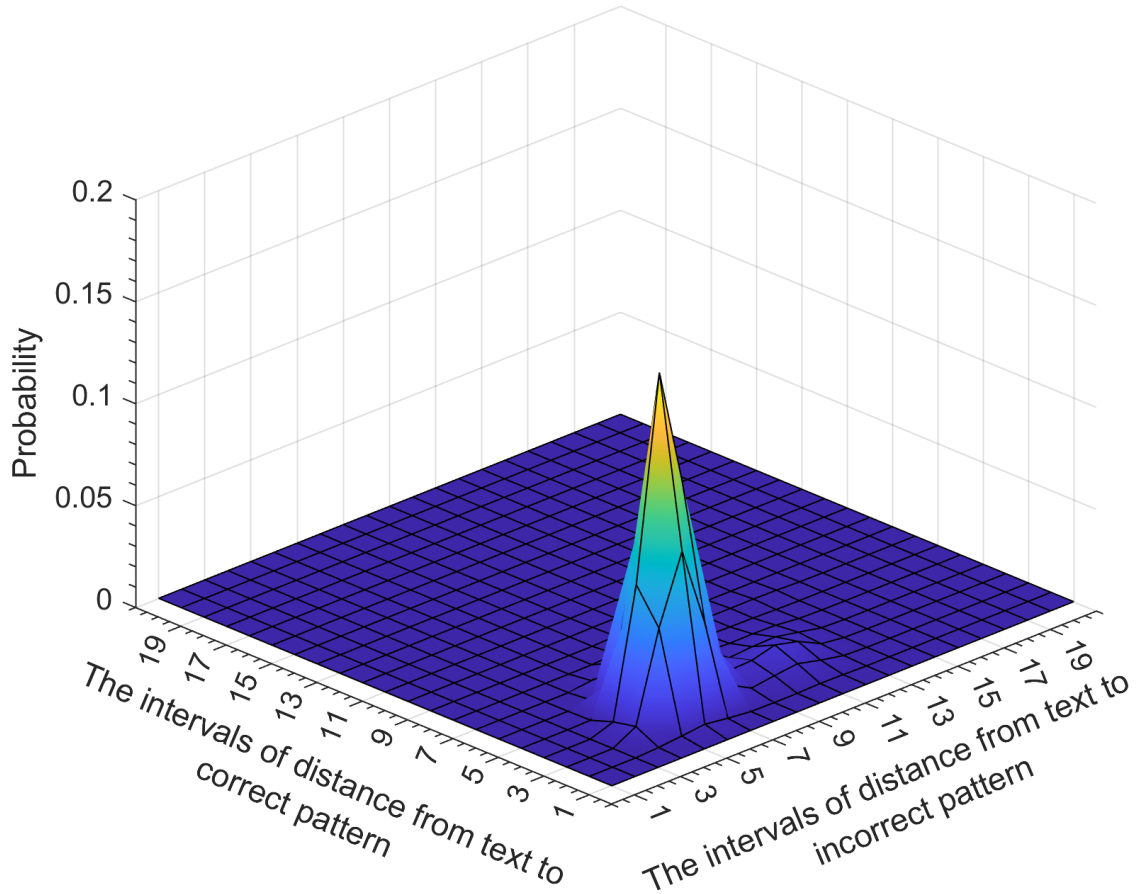

Рис. 1 – Совместная плотность распределения расстояний биграмм между текстами и эталонами «свой-чужой» в норме L1

Однако следует заметить, что, хотя точность распознавания в данном примере оказалась достаточно высокой для такого рода задач, строгого обоснования полученных результатов нет. Дело в том, что метод ближайшего соседа – эвристический, то есть не всегда ближайший означает правильный. Кроме того, нет теоремы, в рамках которой могла бы быть определена вероятность ошибки распознавания в заданной системе эталонов. Более того, величина ошибки существенно зависит от нормы, в которой вычисляется расстояние между распределениями. Например, если в той же задаче о распознавании автора текста вместо нормы L1 для эмпирических частот использовать норму C для соответствующих функций распределения, то ошибка сильно возрастет и станет равной 0,68, то есть две трети текстов будут распознаны неверно. Совместное распределение расстояний «свой-чужой» для этой нормы приведено на рис. 2. Ширина классового интервала на рис. 2 равна 0,03.

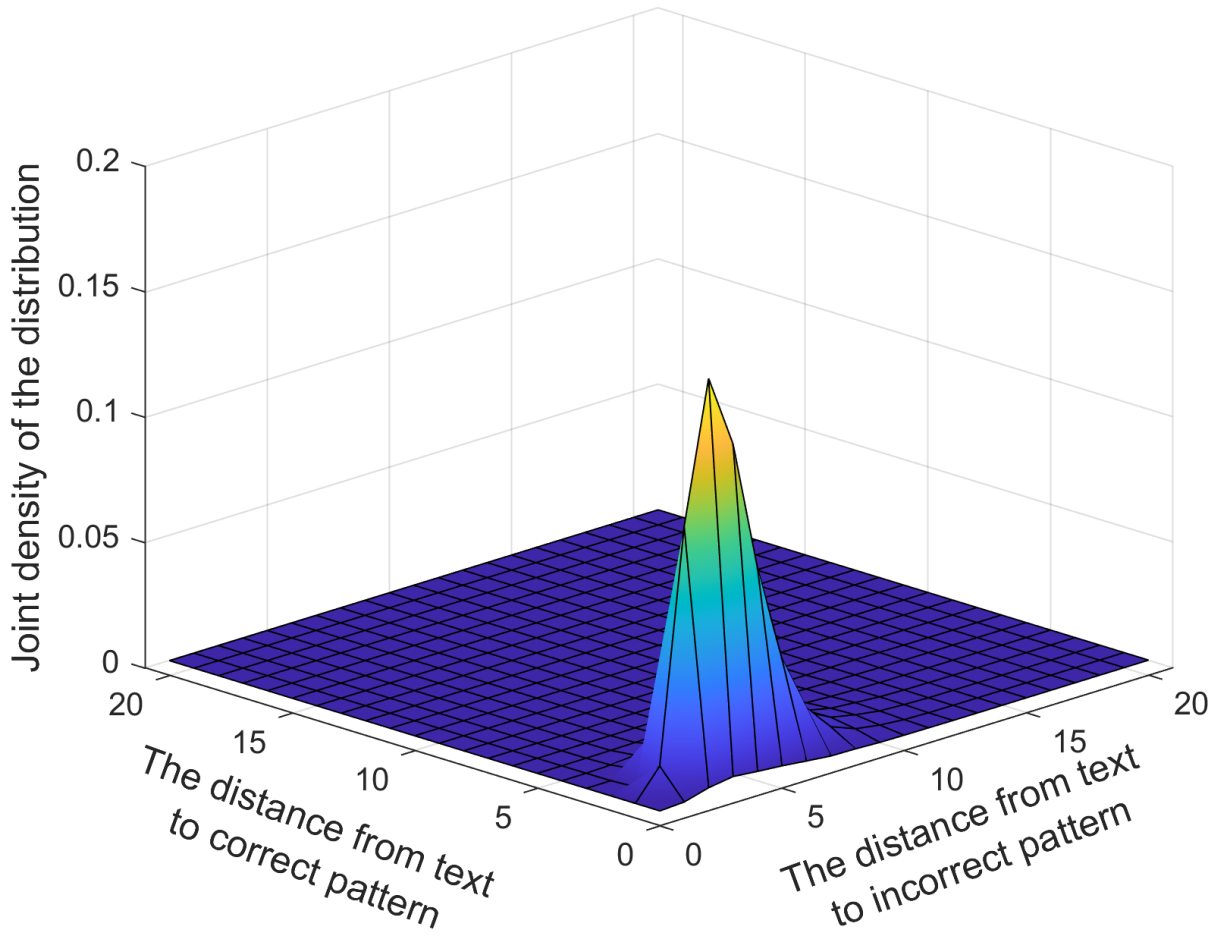

Рис. 2 – Совместная плотность распределения расстояний биграмм между текстами и эталонами «свой-чужой» в норме C

Следовательно, такое значительное расхождение в точности распознавания автора в нормах L1 и C нуждается в специальном анализе.

Сама по себе вероятностная идея статистического распознавания образов состоит в том, что каждый объект в идеале описывается определенной функцией распределения своих параметров, а при наблюдении объекта мы имеем дело с некоторой выборочной функцией распределения. Применительно к авторам эта идея означает, что отдельное литературное произведение представляет собой (это и есть, собственно, гипотеза) выборку из авторской генеральной совокупности, то есть из его эталона. Случайно может оказаться так, что эта выборка будет ближе не к своему эталону, а к некоторому чужому. Однако тестирование такой концепции применительно к эталонам, не являющимися специальными функциями заданного типа, ранее не проводилось. Также не проводился верификационный анализ того, насколько такая концепция распознавания авторов текстов является адекватной. А именно: не было проведено исследования, какая ошибка распознавания выборки ожидается, допустим, при наличии альтернативного эталона в зависимости от длины выборки и от расстояния между эталонами.

Пример авторских эталонных распределений биграмм приведен на рис. 3.

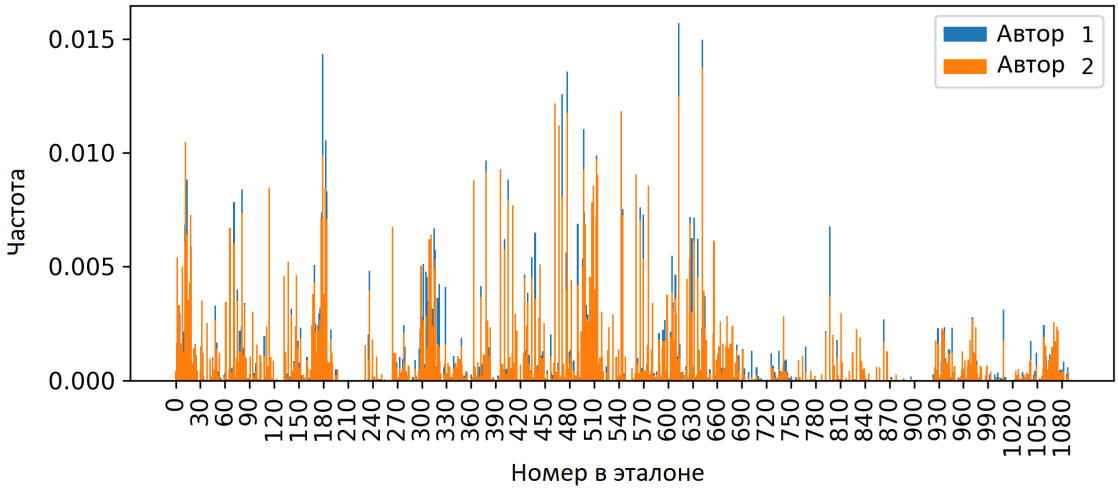

Рис. 3 – Пример эталонных распределений биграмм двух авторов

Видно, что распределения частот для разных авторов несколько различаются, но насколько существенны эти различия для распознавания выборок из данных распределений, пока не ясно. В нашем примере наборы частот биграмм представляют собой совокупность векторов, размерность которых равна $33^2 = 1089$. Будем относиться к ним как к заданным генеральным совокупностям. Возникает вопрос: если сгенерировать выборку некоторой длины N из первого эталона, то как часто она будет ошибочно принята за выборку из второго эталона?

Если за основу такого эксперимента принять имеющиеся эталоны данного корпуса текстов, то можно проанализировать зависимость ошибки от расстояния между эталонами, поскольку эталоны расположены неравномерно.

На рис. 4 приведено распределение расстояний между ними в норме L1. Как видно, минимальное расстояние между эталонами равно 0,075, а максимальное равно 0,425. Следовательно, для тестирования точности распознавания можно выбрать пары эталонов, расположенные на различных расстояниях. Кроме того, можно выбрать два эталона, находящихся примерно на одинаковом расстоянии от третьего и удаленных на большее расстояние один от другого, после чего проверить, с одинаковой ли ошибкой будут определяться выборки из третьего эталона при сравнении сначала с одним эталоном, а потом с другим. То есть в такой постановке можно будет исследовать вопрос о том, насколько расстояние между эталонами, структура которых примерно одинакова, является ключевым параметром, позволяющим провести объективную верификацию концепции распознавания автора текста, а также насколько существенны отдельные индивидуальные особенности эталонов.

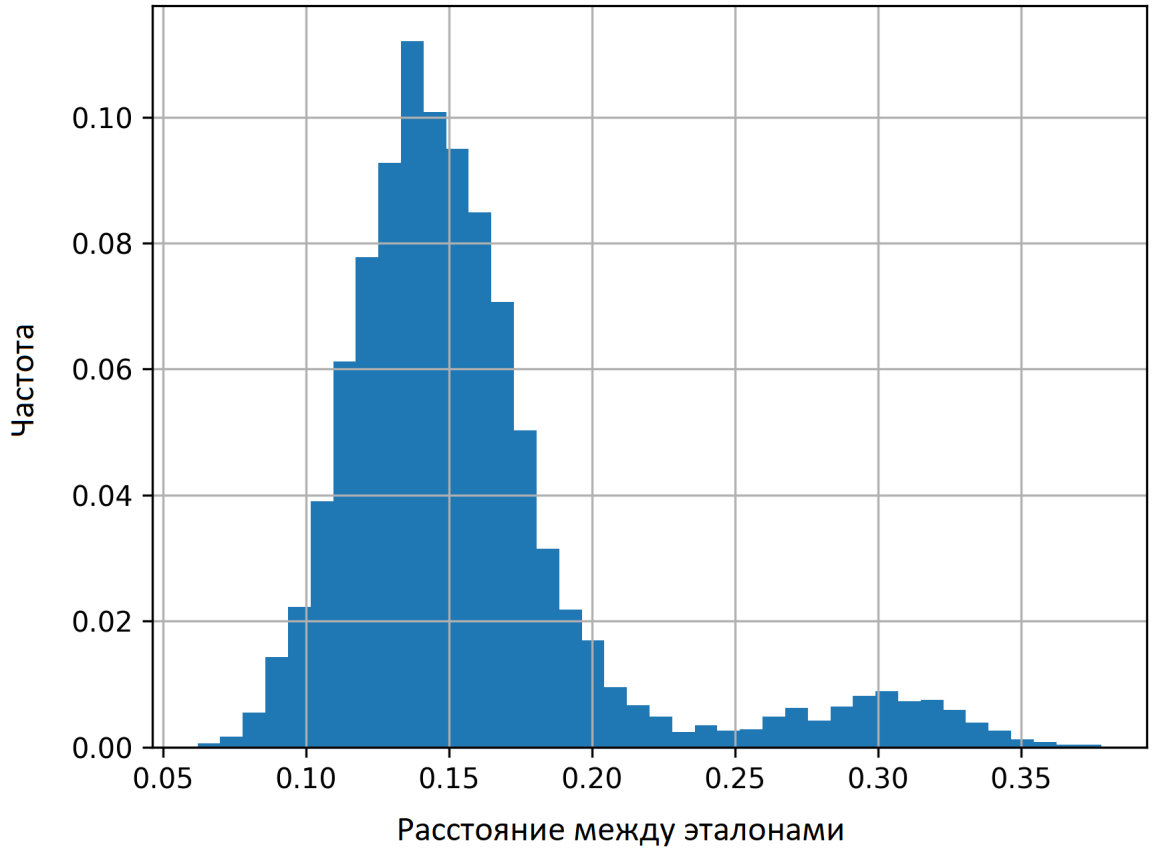

Рис. 4 – Распределение расстояний между авторскими эталонами распределений биграмм в норме L1

Далее дается математическая формулировка задачи об определении эмпирической ошибки распознавания генеральной совокупности по выборке.

3. Постановка задачи о распознавании выборочного распределения

Пусть рассматривается случайный *n*-мерный вектор, представляющий собой частоты буквосочетаний в тексте. Обозначим эти частоты через f_j, $j = 1,...,n$. При этом $f_j \geq 0$, $\sum_{j=1}^{n} f_j = 1$.

Для генерации выборки из данного дискретного распределения используем стандартный метод обращения выборки из равномерного распределения на $[0;1]$. Если $\{x_1,...,x_N\}$ есть выборка из равномерного распределения, то выборку $\{y_1,...,y_N\}$ из непрерывного распределения с функцией F можно получить как решение уравнения $x_k = F(y_k)$, то есть $y_k = F^{-1}(x_k)$. При численном решении такого уравнения надо найти номер классового интервала, которому соответствует данное число x_k. Исходный набор эталонных частот f_j представляет собой ступенчатую гистограмму, определяемую формулой:

$$\begin{aligned} &f(y) = f_j,\ y \in [j-1;\ j),\ j = 1,...,n-1; \\ &f(y) = f_n,\ y \in [n-1;\ n]. \end{aligned} \tag{3.1}$$

Обозначим указанные классовые промежутки через Δ_j. Непрерывная аппроксимация соответствующей функции распределения имеет вид:

$$F(y)\Big|_{y\in\Delta_j} = F(j-1) + f_j \cdot (y - j + 1). \tag{3.2}$$

Как видно, функция (3.2) является строго монотонной, только если все частоты положительны: $f_j > 0$. Заметим, что применительно к буквосочетаниям это, вообще говоря, не так, поскольку существуют невозможные сочетания вида «гласная – мягкий знак» и ряд других. Важно понимать, что в разных текстах одного и того же писателя носители эмпирических распределений частот не совпадают, как не совпадают они и для эталонных распределений разных авторов. Поэтому наши эталоны все равно имеют некоторые «выборочные» черты в том плане, что нулевое значение эталонной частоты не означает невозможного события, ибо на стадии тестирования может появиться текст с таким буквосочетанием. Этот факт несколько затрудняет корректное обоснование верификационного эксперимента. Указанную трудность предлагается обойти следующим образом. Упорядочим частоты эталона, из которого предполагается генерировать выборки, по убыванию. В результате получим $n' < n$ ненулевых частот. Перенумеруем затем частоты g_j второго эталона (альтернативы) в соответствии с нумерацией частот первого эталона. Естественно, эти вторые

частоты уже не будут в общем случае упорядочены по убыванию. Соответствующую гистограмму обозначим $g(y)$, а функцию распределения – $G(y)$. Формулы для них аналогичны (3.1), (3.2). Далее для каждого элемента x_k равномерной выборки определяем номер классового интервала j_k так, что $F(j_k-1)\le x_k < F(j_k)$. В результате находим число $m_N(j)$ попаданий элементов выборки $\{x_1,...,x_N\}$ в промежутки $R_j=[F(j-1); F(j))$, $j=1,...,n'$. Конкретные значения элементов y_k находятся из (3.2) обращением линейной функции:

$$y_k = j_k - 1 + \frac{x_k - F(j_k-1)}{f_{j_k}}. \qquad (3.3)$$

Если мы хотим протестировать полученную выборку на предмет ее близости к другому распределению с функцией $G(y)$,

$$G(y)\Big|_{y\in\Delta_j} = G(j-1) + g_j\cdot(y-j+1),$$

то следует проверить, насколько элементы $z_k = G(y_k)$ близки к равномерному распределению. Если получаемая таким образом выборка элементов

$$z_k = G(j_k-1) + \frac{g_{j_k}}{f_{j_k}}(x_k - F(j_k-1)) \qquad (3.4)$$

оказывается ближе к равномерному распределению, чем выборка x_k, то будет фиксироваться ошибка идентификации.

Выборка $\{x_1,...,x_N\}$, как выборка из равномерного распределения, имеет некоторые выборочные частоты, которые обозначим через $p_N(j)$, $j=1,...,n$. Получаемая в результате преобразований (3.4) выборка $\{z_1,...,z_N\}$ имеет частоты, обозначаемые через $q_N(j)$, $j=1,...,n$.

В норме L1 рассматриваем суммы

$$S_1 = \sum_{j=1}^{n}\left|p_N(j) - 1/n\right|,\ S_2 = \sum_{j=1}^{n}\left|q_N(j) - 1/n\right|. \qquad (3.5)$$

Если $S_1 < S_2$, то распознавание имеет правильный результат, а если наоборот, то ошибочный.

В норме C рассматриваются соответствующие функции распределения $P_N(j)$ и $Q_N(j)$. Сравниваются величины

$$M_1 = \max_j\left|P_N(j) - j/n\right|,\ M_2 = \max_j\left|Q_N(j) - j/n\right|. \qquad (3.6)$$

Если $M_1 < M_2$, то распознавание имеет правильный результат, а если наоборот, то ошибочный.

4. Статистическая точность оценки эталонных частот

Перед проведением вычислительного эксперимента следует дать априорную оценку точности в определении эмпирических частот буквосочетаний в предположении случайного появления букв в тексте. Метод такой оценки был предложен в [2]. Здесь мы конкретизируем его применительно к рассматриваемой задаче.

Пусть $f^{(n)}(j)$ есть генеральная совокупность частот для *n*-грамм, где j есть номер *n*-граммы, а $p^{(n)}(j;N)$ есть выборочная совокупность частот по выборке длины *N*. Тогда (см., напр., [7]) статистика

$$t = \sqrt{N-1}\,\frac{\left|p^{(n)}(j;N) - f^{(n)}(j)\right|}{s(j;N)} \tag{4.1}$$

имеет распределение Стьюдента с $N-1$ степенью свободы. Здесь $s^2(j;N)$ есть выборочная дисперсия частоты, равная

$$s^2(j;N) = p^{(n)}(j;N)\cdot\left(1 - p^{(n)}(j;N)\right). \tag{4.2}$$

Поскольку $N >> 1$, то вместо квантиля распределения Стьюдента в оценке доверительного интервала для вероятности $p^{(n)}(j;N)$ можно взять квантиль нормального распределения $u_{1-\varepsilon/2}$, где ε – уровень значимости, на котором принимается решение о близости распределений. Тогда в приближении $N-1 \approx N$ на уровне значимости ε выражение $\left|p^{(n)}(j;N) - f^{(n)}(j)\right|$ не превосходит величины $s(j;N)u_{1-\varepsilon/2}/\sqrt{N}$.

Заметим теперь (см. рис. 3), что эталонные частоты имеют весьма разный порядок – от 10^{-1} до 10^{-6}. Поэтому требовать одинаковой точности позиционирования для малых и для больших частот статистически неоправданно. Для наших целей интерес представляет собственно распределение, для чего введем средневзвешенную точность ε_j для оценки $p^{(n)}(j;N)$. Нас будет интересовать выполнение условия

$$\sum_{j=1}^{n}\left|p^{(n)}(j;N) - f^{(n)}(j)\right| \le \varepsilon \tag{4.3}$$

в предположении, что для каждой *n*-граммы будет соблюдена своя точность ε_j, так что

$$u_{1-\varepsilon_j/2}\,\frac{s(j;N)}{\sqrt{N}} \le \varepsilon_j f^{(n)}(j). \tag{4.4}$$

Сумма всех частот нормирована на единицу, поэтому после подстановки (4.4) в (4.3) получаем

$$\sum_{j=1}^{n}\left|p^{(n)}(j;N) - f^{(n)}(j)\right| \le \sum_{j=1}^{n} u_{1-\varepsilon_j/2}\,\frac{s(j;N)}{\sqrt{N}} = \frac{u_{1-\varepsilon/2}\Sigma_N(n)}{\sqrt{N}}, \tag{4.5}$$

где введен средневзвешенный по эмпирическим дисперсиям квантиль

$$u_{1-\varepsilon/2} = \frac{1}{\Sigma_N(n)} \sum_{j=1}^{n} s(j;N) u_{1-\varepsilon_j/2} \,. \tag{4.6}$$

Величина $\Sigma_N(n)$ в (4.5) есть нормировка весов, т.е. сумма

$$\Sigma_N(n) = \sum_{j=1}^{n} s(j;N) = \sum_{i=1}^{n} \sqrt{p^{(n)}(j;N) \cdot \left(1 - p^{(n)}(j;N)\right)}. \tag{4.7}$$

Потребуем теперь, чтобы уровень значимости ε в (4.5) был бы равен уровню неопределенности в оценке совокупности частот в норме L1 в (4.3). Тогда получаем условие

$$\frac{u_{1-\varepsilon/2}}{\varepsilon} \le \frac{\sqrt{N}}{\Sigma_N(n)}. \tag{4.8}$$

При заданной точности ε состояний формула (4.8) в случае знака равенства дает оценку на минимальную длину выборки, при которой эта точность достигается в среднем. Обозначим $\varepsilon/2$ через α. Тогда условие (4.8) перепишется в виде $u_{1-\alpha}/\alpha \le 2\sqrt{N}/\Sigma_N(n)$. Функция $u_{1-\alpha}/\alpha$ монотонно убывает с ростом α, поэтому к ней существует обратная, значение которой и дает верхнюю оценку точности определения эмпирических вероятностей.

При $N >> n$ сумма (4.7) весьма слабо зависит от N и варьируется в пределах нескольких процентов. Округленные средние значения $\Sigma_N(n)$ по анализируемому корпусу текстов приведены в таблице 1.

Таблица 1. Значения $\Sigma_N(n)$ и точность оценки $f^{(n)}$

n	1	2	3	4
$\Sigma_N(n)$	5	20	50	100
Значения ε по формуле (4.8)				
N = 10 тыс.	0,08	0,22	0,40	0,60
N = 30 тыс.	0,05	0,15	0,30	0,42
N = 50 тыс.	0,04	0,12	0,25	0,39
N = 100 тыс.	0,03	0,10	0,20	0,31
N = 500 тыс.	0,02	0,05	0,15	0,20
N = 1 млн	0,01	0,04	0,10	0,15

Авторский эталон составляется по достаточно большому числу произведений, так что число знаков в нем порядка миллиона или больше. Обращаясь к рис. 4, видим, что минимальное расстояние между эталонами биграмм данного корпуса равно примерно 0,08, то есть в два раза больше неопределенности в оценке самих эталонов, что является критическим условием для возможности их разделения статистическими методами. Следовательно, взятые нами для эксперимента эталоны корректно считать различными генеральными совокупностями, а не случайными выборками из

одного и того же распределения, поэтому эксперимент по идентификации выборок может быть проведен адекватно.

5. Результаты вычислительных экспериментов

Прежде чем проводить основной эксперимент по анализу зависимости точности идентификации от длины выборки и расстояния между эталонами, надо оценить субъективную составляющую нашей модели. Следует понять, существенно ли для результатов анализа то, какой именно паттерн взят в качестве первой генеральной совокупности, а какие – в качестве альтернативы.

Эксперимент 1. Проверка независимости распознавания от выбора конкретного эталона.

Поскольку носители эталонных распределений не совпадают, то следует проверить, насколько отличается ошибка распознавания выборки при изменении альтернативы на противоположную. В разделе 3 была описана задача об идентификации выборки при наличии двух эталонов, которые заданы своими функциями распределения F и G. Эксперимент состоит в том, что генерируется выборка $\{x_1,...,x_N\}$ из равномерного распределения на [0;1], а затем с ее элементами делаются преобразования, приводящие к двум выборкам: $a_k = G\left(F^{-1}(x_k)\right)$ и $b_k = F\left(G^{-1}(x_k)\right)$. Таких выборок генерируется, скажем, 1000.

Ошибка идентификации первого варианта – это доля выборок $\{a_k\}$, которые оказались ближе к равномерному распределению, чем исходная выборка $\{x_k\}$. Ошибка идентификации второго варианта – это доля выборок $\{b_k\}$, которые оказались ближе к равномерному распределению, чем исходная выборка $\{x_k\}$. В идеале эти доли должны совпадать. Качественно влияние асимметрии, если оно проявится, связано с разными долями носителя одного из эталонов в носителе другого.

Рассматриваются все пары эталонных биграмм, расстояние между которыми в L1 незначительно превышает уровень разделения из таблицы 1 и равно 0,1. Генерируется 1000 выборок $\{x_k\}$ разных длин. Определяется ошибка идентификации по выборкам $\{a_k\}$, и $\{b_k\}$, после чего находится средняя по парам ошибка и дисперсия ошибки. Если корень из дисперсии много меньше средней ошибки, то индивидуальные различия в эталонах несущественны. В противном случае идентификация носит субъективный характер.

Такие же статистические вычисления проводятся для других пар эталонов, находящихся между собой на расстояниях 0,11, 0,12 и т.д.

Интересно провести сравнение с точностью распознавания выборок из эталонов тех же авторов, которые были отобраны в пары по расстояниям между биграммами, если взять эталоны в виде униграмм и триграмм. Это позволит оценить, насколько размерность пространства представляющих векторов влияет на точность распознавания. Для униграмм длина вектора частот равна $n = 33$, биграммам соответствует $n = 1089$, а для триграмм $n = 35937$.

Описанные вычисления повторяются в норме C.

Результаты этих экспериментов будем обозначать в легенде графиков как «case 1 a,b».

Эксперимент 2. Проверка зависимости ошибки от числа эталонов.

Добавим в систему из двух эталонов (см. описание эксперимента 1) третий эталон так, что попарные расстояния между ними примерно одинаковы и равны 0,1. Один из них обозначим как F, а два других – как G_1 и G_2. Рассмотрим две серии экспериментов по получению выборок с элементами $a_{k,1} = G_1\left(F^{-1}(x_k)\right)$ и $a_{k,2} = G_2\left(F^{-1}(x_k)\right)$. Генерируется 1000 выборок разных длин. Определяется ошибка идентификации этих двух множеств выборок. Далее повторяем те же процедуры, что и в эксперименте 1. Результаты обозначаем в легенде графиков как «case 2».

Эксперимент 3. Исследование зависимости ошибки бинарного распознавания от размерности эталонов, длины выборки и расстояния между эталонами в нормах L1 и C.

Выбирается некоторый условно первый эталон F и несколько эталонов $G_1 \div G_k$ с последовательно увеличивающимися расстояниями до первого эталона в норме L1. Авторы выбранных эталонов при изменении размерности и нормы остаются прежними. Первоначальный отбор эталонов проводился на основе анализа расстояний между биграммами.

Результаты этого эксперимента обозначим как «case 3.1», «case 3.2» и, если вариантов достаточно, ««case 3.3».

Эксперимент 4. Исследование зависимости ошибки распознавания выборки среди нескольких эталонов. Разовьем эксперимент 2 и расположим теперь шесть эталонов «вокруг» данного первого так, что они находятся примерно на одинаковом расстоянии от него и на не меньшем расстоянии один от другого. Вычисляем ошибку распознавания выборки. Результат обозначим как «case 4». Через T обозначаем расстояние между эталонами в L1. Длина выборки обозначается как «N gramm».

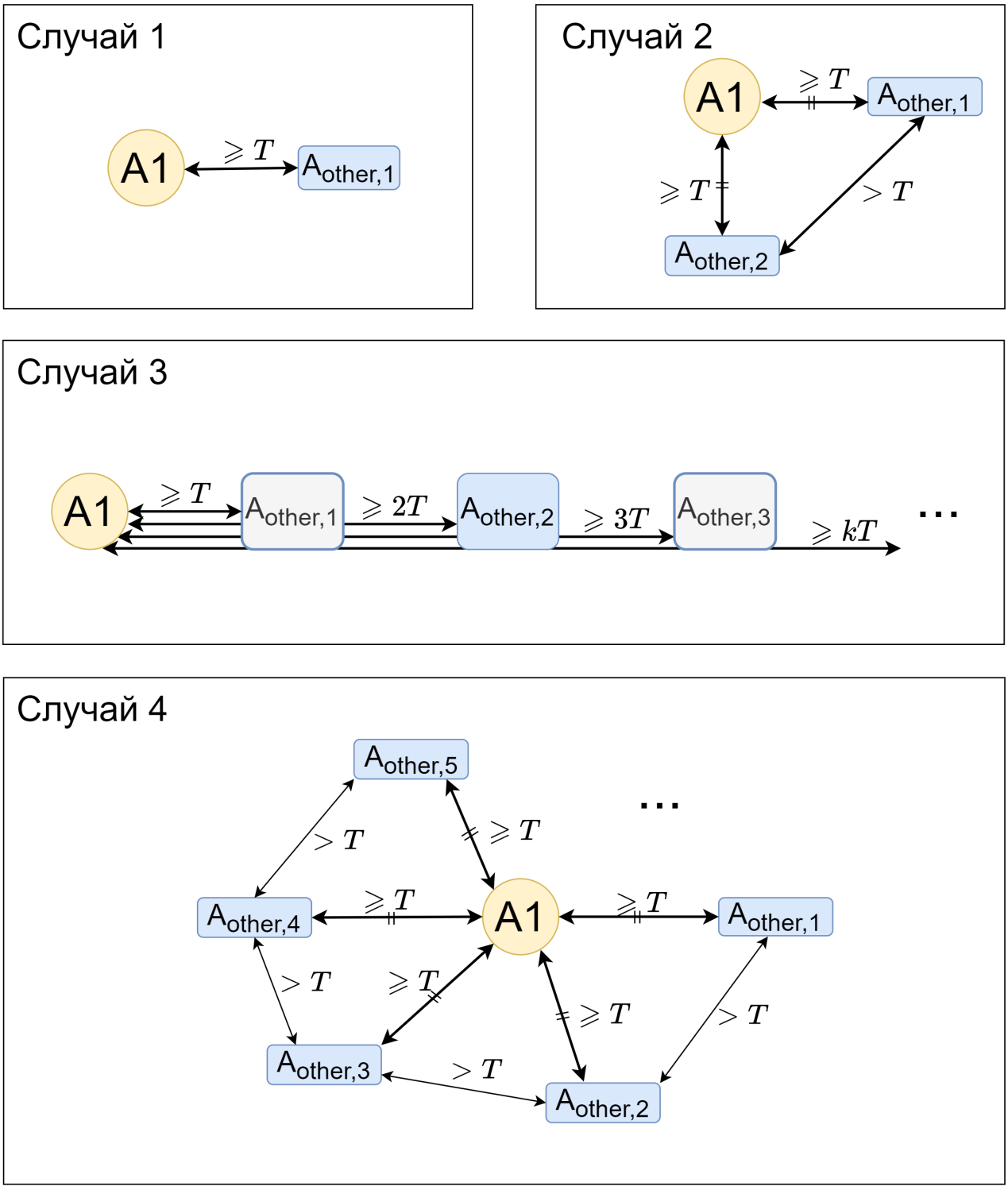

Рис. 5 – Схемы проводимых экспериментов

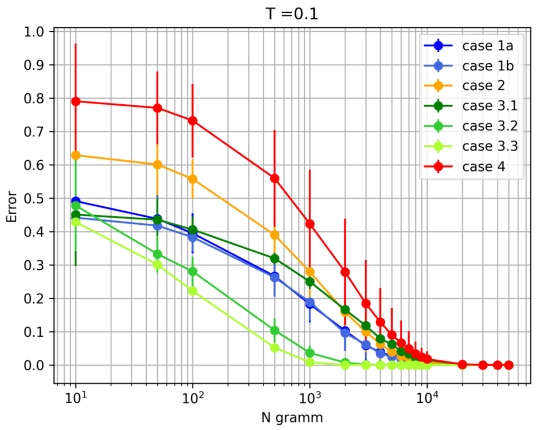

Рис. 6 – Ошибка униграмм, L1

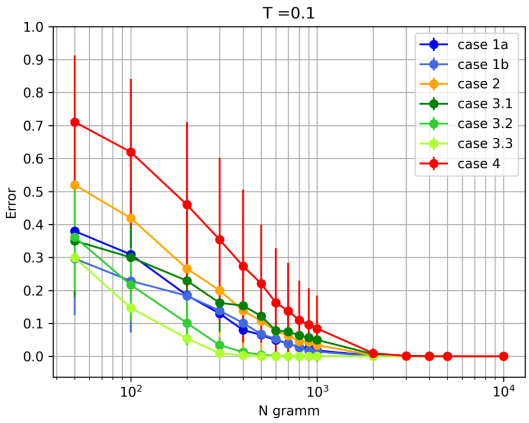

Рис. 7 – Ошибка биграмм, L1

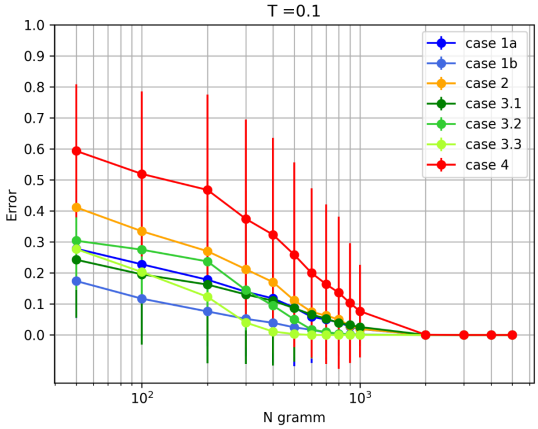

Рис. 8 – Ошибка триграмм, L1

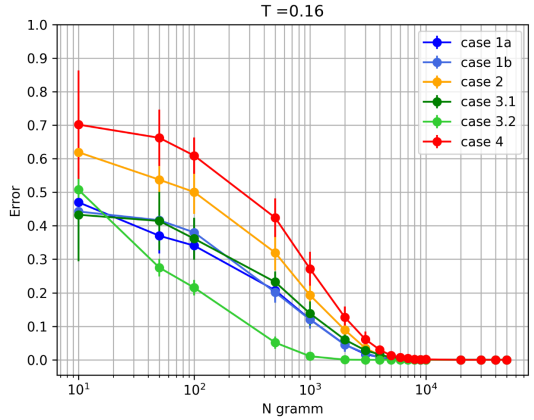

Рис. 9 – Ошибка униграмм, L1

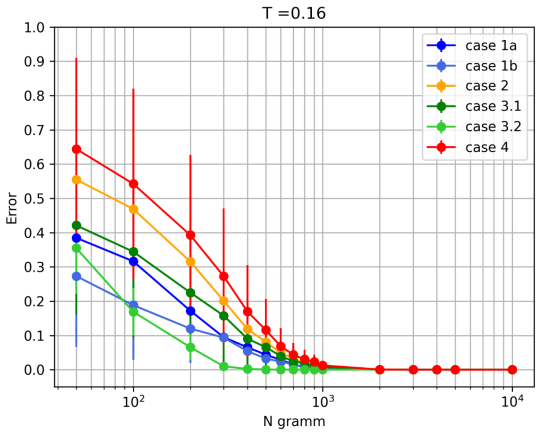

Рис. 10 – Ошибка биграмм, L1

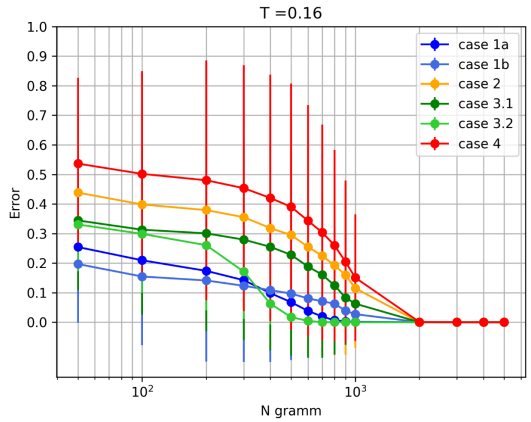

Рис. 11 – Ошибка триграмм, L1

Из рис. 6-11 следует, что увеличение размерности вектора состояний приводит к снижению ошибки распознавания. Увеличение расстояния между эталонами также снижает ошибку, но несущественно. Увеличение числа эталонов (переход от «case 1» к «case 4») увеличивает ошибку. Вертикальные черточки показывают среднеквадратичное отклонение. Оно достаточно велико, поэтому распознавание существенно зависит от выбранной системы эталонов, то есть в большой степени субъективно.

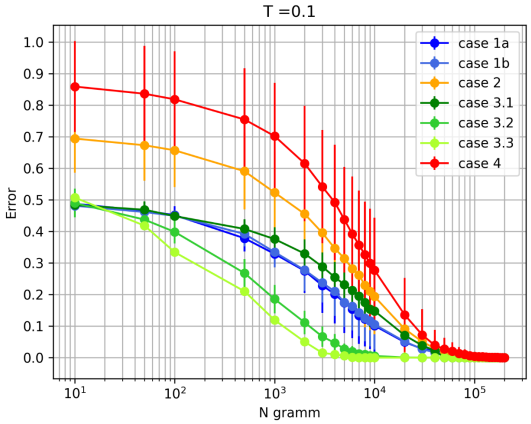

Рис. 12 – Ошибка униграмм, C

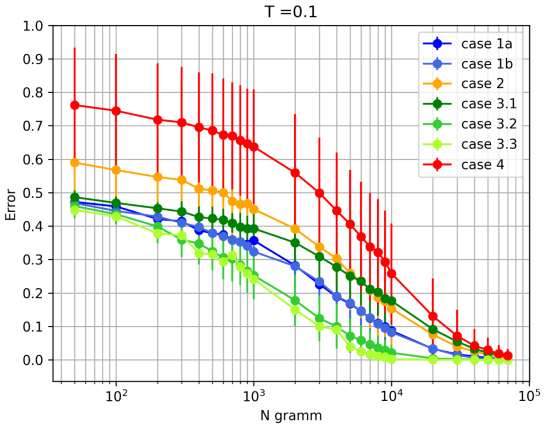

Рис. 13 – Ошибка биграмм, C

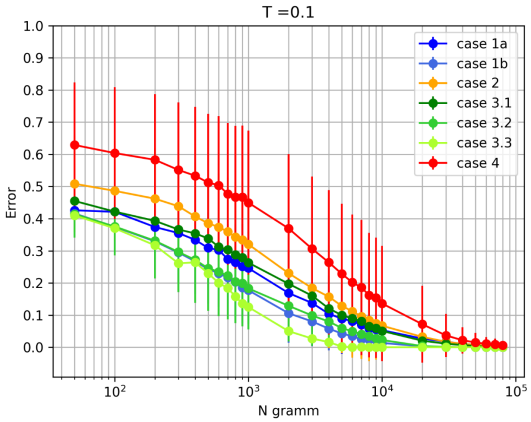

Рис. 14 – Ошибка триграмм, C

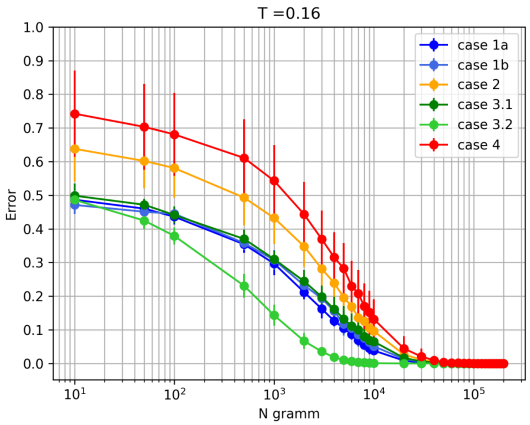

Рис. 15 – Ошибка униграмм, C

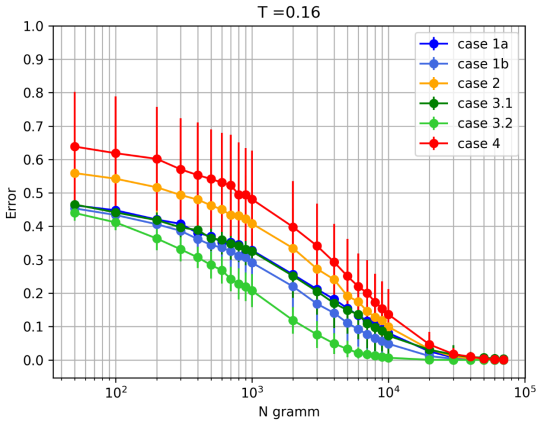

Рис. 16 – Ошибка биграмм, C

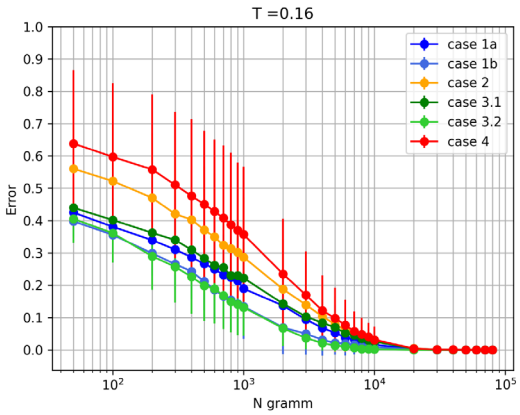

Рис. 17 – Ошибка триграмм, C

Как в L1, так и в C нормах наблюдается снижение ошибки с увеличением размерности вектора распределения вероятностей и увеличение ошибки с ростом числа эталонов. Критически важным моментом является то, что ошибка в L1 равна нулю для выборок длиной более чем 10^3, тогда как в норме C для этого требуется длина на уровне 10^5. Также следует отметить, что дисперсия в C несколько больше, чем в L1. Эти результаты объясняют аномальную ошибку распознавания в C для системы эталонов из 100 векторов.

Однако проведенные расчеты показали, что гипотеза о «выборочном» характере отдельных произведений писателя из его «авторского эталона», строго говоря, не соответствует реальности. Поскольку выборочное распределение распознается безошибочно на выборках длин порядка 10^4, а отдельное произведение автора насчитывает в среднем 400 тыс. знаков, то распознавание произведений должно было пройти без ошибок. Фактическая ошибка распознавания по данному корпусу текстов составила, как уже говорилось выше, величину 0,12. Эта ошибка сравнительно невелика, поэтому отдельный текст все же несет существенные черты авторского эталона. В то же время вариабельность отдельных произведений свидетельствует о нестационарности этого творческого процесса.

Отметим также, что оценка точности эмпирических частот (таблица 1) не связана с точностью распознавания выборки при наличии альтернативы. Например, для выборок длин 10^4 точность биграмм составляет 0,2, тогда как ошибка распознавания при расстояниях между эталонами 0,1 близка к нулю (вместо ожидаемой ситуации 50/50). Следовательно, проведение численного анализа точности распознавания в рамках данной системы эталонов является необходимой процедурой, которую нельзя заменить априорными оценками.

6. Тестирование заданной системы эталонов на точность

Алгоритм тестирования состоит в следующем.

Шаг 1. На вход подается (то есть читается) система из p векторов $f_i(j), i = 1,2,..., p$ в пространстве размерности n, координаты которых положительны и в сумме равны единице по каждому вектору.

Шаг 2. Выбираются параметры тестирования: минимальная длина выборки $N_{\min}$, шаг перебора длин h.

Шаг 3. Генерируются 10^4 выборок длины $N_k = N_{\min} + (k-1)h$ из равномерного распределения на $[0;1]$. Это количество выбрано исходя из тех соображений, чтобы статистическая точность оценивания ошибки была порядка 0,01 (то есть порядка корня из числа генераций).

Шаг 4. Каждая из выборок используется для генерации выборочного распределения из каждого эталона. В результате получается набор $\varphi_i^m(j;N)$ выборочных распределений, где i индексирует номер эталона, m соответствует номеру генерации, j отвечает компоненте вектора.

Шаг 5. Определяются расстояния $r_{il}^m(N) = \left\| \varphi_i^m(j;N) - f_l(j) \right\|$ между данным выборочным распределением и одним из эталонов.

Шаг 6. Вычисляется минимальное расстояние $\rho_i^m(N) = \min_l r_{il}^m(N)$ и подсчитывается доля случаев $\nu_i(N)$, в которых $\rho_i^m(N) \neq r_{ii}^m(N)$.

Шаг 7. На выходе алгоритма даются распределение ошибки $\nu(N)$ в заданной системе эталонов, а также средняя ошибка и ее среднеквадратичное отклонение. Распределение представляется в виде графика и таблицы квантилей.

Построенный численный алгоритм позволяет провести статистическую «паспортизацию» эталонной системы в задаче распознавания образов методом ближайшего соседа. На рис. 18-20 приведены такие распределения ошибок в зависимости от длины выборки для униграмм, биграмм и триграмм в норме L1 для рассматриваемой системы из 100 авторских эталонов.

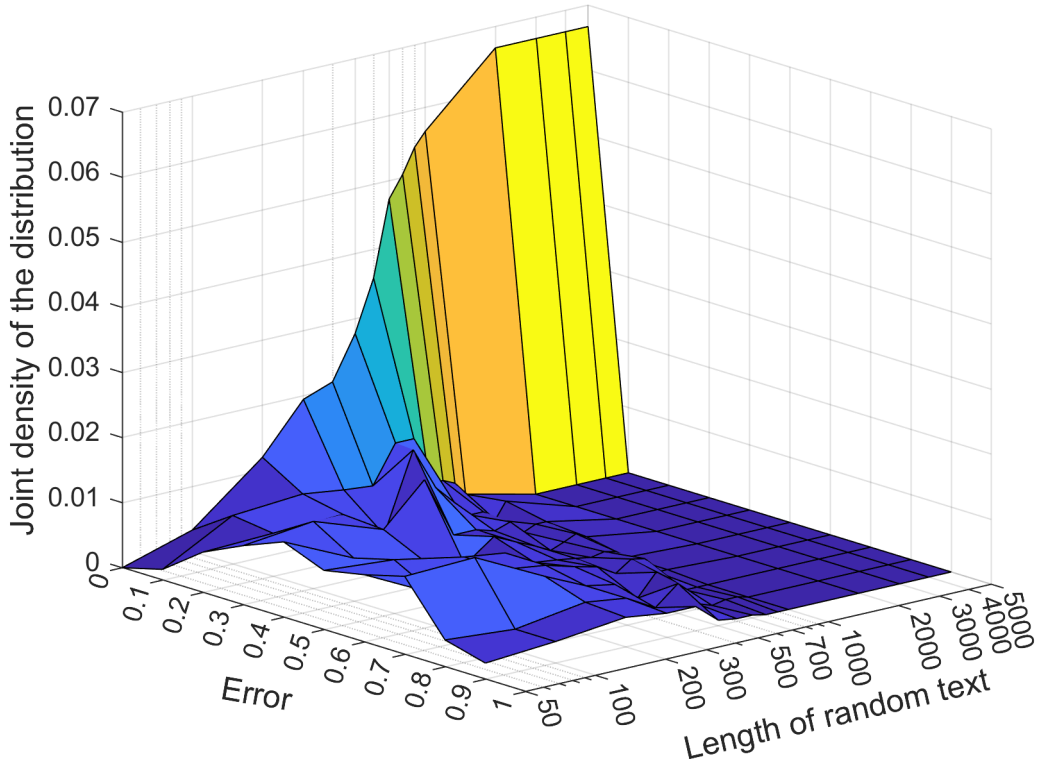

Рис. 18 – Вероятность ошибки идентификации выборок из униграмм

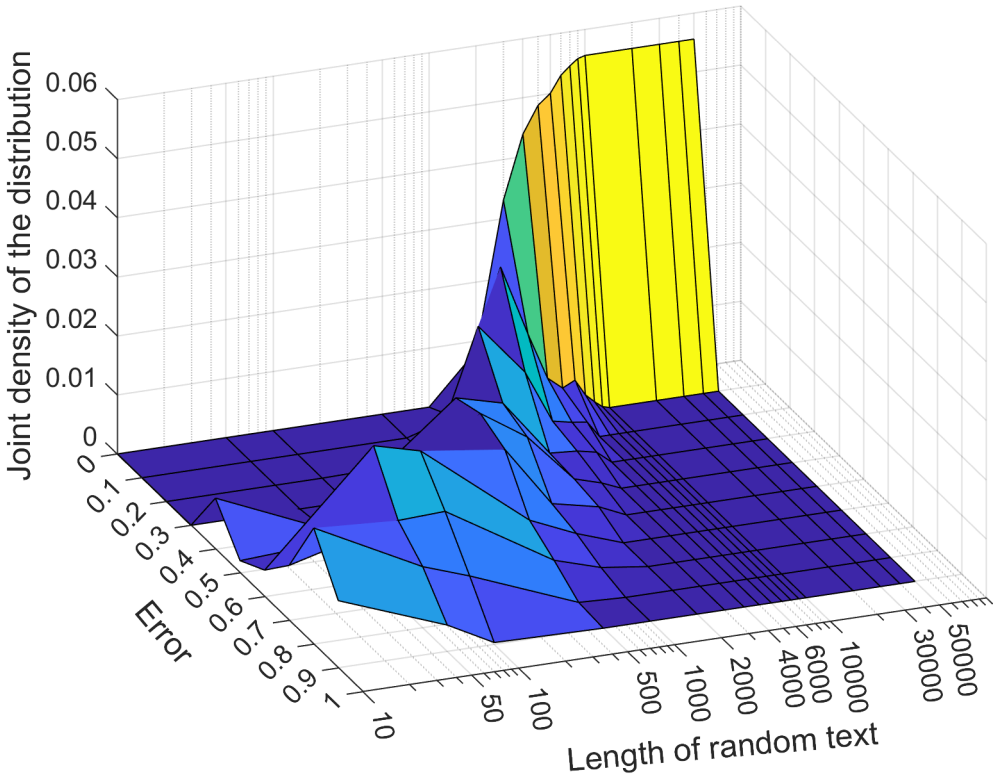

Рис. 19 – Вероятность ошибки идентификации выборок из биграмм

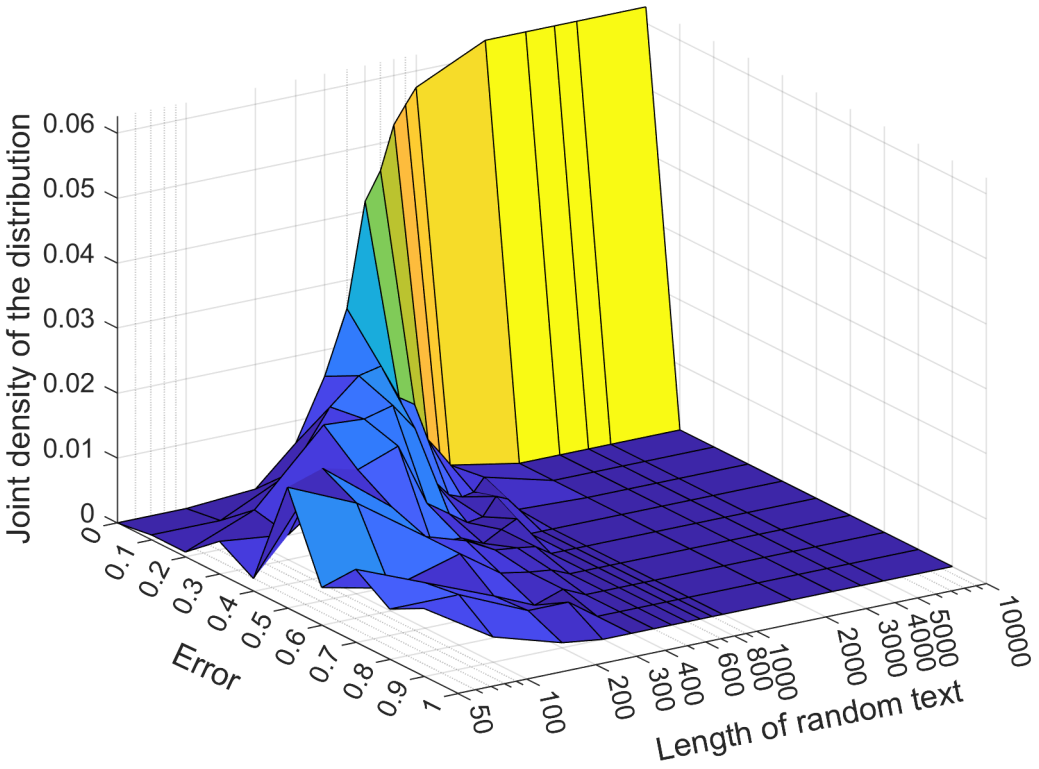

Рис. 20 – Вероятность ошибки идентификации выборок из триграмм

7. Заключение

В работе проведен вычислительный эксперимент, показывающий, что точность распознавания методом ближайшего соседа зависит от конкретных эталонных представителей классов. Важным практическим результатом явилась демонстрация того, что ошибка распознавания в большей степени определяется самими эталонными распределениями, а не расстояниями между ними. Следовательно, для применения данного метода распознавания следует предварительно провести тестирование заданной системы эталонов на точность, которую можно ожидать, в зависимости от длины выборки.

Разделы 1 и 3 написаны А.А. Кислицыным, раздел 2 написан М.Ю. Кислицыной, остальные разделы написаны совместно.

Литература

1. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. – М.: Наука, 1974. – 416 с.
2. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.
3. Борисов Л.А., Орлов Ю.Н., Осминин К.П. Идентификация автора текста по распределению частот буквосочетаний // Препринты ИПМ им. М.В. Келдыша. 2013. № 27. 27 с.
4. 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.
5. Воронина М.Ю., Кислицын А.А., Орлов Ю.Н. Алгоритм коррекции метода биграмм в задаче идентификации автора текста // Математическое моделирование. 2022. Т. 34. № 9. С. 3-20.
6. Воронина М.Ю., Орлов Ю.Н. Определение автора текста методом сегментации // Компьютерные исследования и моделирование. 2022. Т. 14. № 5. С. 1199-1210.
7. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.