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

Препринты ИПМ • Препринт № 17 за 2016 г.

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

Ивченко А.Ю., Орлов Ю.Н.

Практические аспекты задачи распознавания образов

Рекомендуемая форма библиографической ссылки: Ивченко А.Ю., Орлов Ю.Н. Практические аспекты задачи распознавания образов // Препринты ИПМ им. М.В.Келдыша. 2016. № 17. 20 с. doi:10.20948/prepr-2016-17
URL: http://library.keldysh.ru/preprint.asp?id=2016-17

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.Ю. Ивченко, Ю.Н. Орлов

Практические аспекты задачи распознавания образов

Москва — 2016

Ивченко А.Ю., Орлов Ю.Н.

Практические аспекты задачи распознавания образов

Исследуются вычислительные ограничения применения байесовского подхода к задаче распознавания образов, когда максимум вероятности соответствия текущего состояния одному из базисных эталонов определяется путем разложения изучаемого фрагмента по известному базису. Дан пример безошибочного распознавания в рамках модели близости выборочного распределения к эталонным распределениям в условиях, когда вероятностная трактовка проекций на базисные паттерны не имеет места. Построены спектральные портреты разрешающих матриц в задаче идентификации автора литературного текста.

Ключевые слова: нестационарный временной ряд, вероятности буквенных комбинаций, базисные паттерны, распознавание образов

Ivchenko A.Yu., Orlov Yu.N.

Practical aspects of pattern recognition

The numerical restrictions of Bayesian method of pattern recognition are investigated. The probability maximum of correspondence of local state to one of the basis patterns is defined through the expansion of examined vector over the patterns. The practical example of recognition without errors is presented in the frame of nearest neighbor method for the case, when the probability interpretation of the expansion coefficients is not valid. The spectral portraits of solving matrices are constructed for the literature texts author identification problem.

Key words: non-stationary time series, probabilities of letters combination, basis patterns, pattern recognition

Работа выполнена при поддержке гранта РФФИ, проект

№ 16-01-00342

Содержание

Введение

Задача распознавания состоит в том, чтобы по измеряемым значениям параметра $x \in X \subset R^n$, относящегося к исследуемой системе, определить, в каком состоянии $s \in S \subset R^m$ находится (находилась или будет находиться) система в тот или иной момент времени по отношению к моменту измерения. 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.

Измеряемым параметром может быть, например, температура воздуха в определенном географическом месте в определенное время суток, биржевой курс ценных бумаг в определенный момент времени и т.п. Это – одномерный параметр, характеризующий систему (соответственно, погоду или биржу). В качестве измеряемого параметра может выступать упорядоченная последовательность из N наблюденных значений, называемая выборкой (или выборочной траекторией) длины N из временного ряда. Такая выборка – это вектор $x \in R^N$. Объектом анализа может быть также некоторый функционал от выборочной траектории: размах, среднее значение, дисперсия, медиана или иной квантиль распределения, собственно выборочная функция распределения или ее плотность, представленная в виде гистограммы. В частности, если наблюдаемая величина принимает дискретный набор из n значений, то гистограмма представляет собой набор эмпирических частот (или выборочных вероятностей) реализации указанных значений, т.е. это вектор $\mathbf{f}_N(x) \in R^n$, где нижний индекс N показывает, что этот вектор построен по выборке длины N. Состоянием системы, о котором надо судить по выборке, является, например, тренд вверх или вниз биржевой цены, причем такой, который интересен конкретному торговцу, ожидающему определенную прибыль.

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. В предположении, 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.

Во всех прикладных задачах пространство состояний S дискретно в силу специфики вычислительных операций, хотя в теоретических моделях это пространство может быть и непрерывным. В настоящей работе мы будем предполагать, что система может пребывать в некотором конечном числе четко идентифицируемых состояний, так что s – это индекс, пробегающий

натуральные значения от 1 до некоторого m. Если других состояний нет, а обучающая выборка для нахождения вероятностей $P(x|s)$ достаточно длинная, причем сами вероятности стационарны, то задача распознавания считается решенной. На практике, однако, часто возникают проблемы вычислительного свойства, препятствующие корректному распознаванию в смысле байесовского подхода, но некоторые следствия из этого подхода могут при всем при том оставаться весьма эффективными для распознавания, не будучи строго обоснованными в отсутствие байесовской корректности. В настоящей работе мы подробно рассмотрим пример такого распознавания, которое опирается на следствие из байесовской теории в виде близости плотностей распределений в норме суммируемых функций, но при этом ключевые условия (вероятностная интерпретация относительных частот или выпуклость конуса соответствующих состояний в пространстве S) не выполнены. Конечно, если вероятностное пространство заведомо не полно, то выбранный метод приведет к неверным результатам. Но если модель представляется адекватной, не следует относиться к таким ситуациям как к неким редким парадоксальным артефактам: авторы на практике сталкиваются с ними постоянно и на собственном опыте убедились, что отказ от теоретически хорошего, но по вычислительным причинам некорректного метода в пользу менее точного, но лучше обоснованного, в ряде случаев снижает точность распознавания до неприемлемой величины. Полезно исследовать причины такого феномена.

Будем рассматривать в качестве измеряемого параметра выборочную плотность функции распределения $f_N(k)$, построенную по выборке длины N, где индекс $k = 1, 2, ..., n$ нумерует классовый интервал, в который попадает наблюдаемое значение x. Такая трактовка измерения часто бывает более содержательна, чем просто соответствие измеренной величины x номеру состояния s. Например, при анализе трендов в биржевых рядах никогда не бывает монотонного роста траектории на протяжении достаточно большого промежутка времени. На любой выборочной траектории локальные максимумы чередуются с локальными минимумами, но если приростов значений ряда одного знака оказывается заметно больше, чем другого, то такой фрагмент воспринимается как тренд. В то же время отдельный прирост $x(t+1) - x(t)$ не позволяет устойчиво идентифицировать текущую ситуацию на промежутке в N шагов как тренд. В работах [2, 3] было исследовано предположение о том, что для биржевых рядов выборочные распределения, отвечающие явно выделяемым состояниям, таким как тренд вверх (вниз), кластеризуются. Действительно, на практике часто наблюдаются состояния, которые в смысле фрагмента траектории достаточно хорошо соответствуют определенным эталонам в терминах функций распределения или их плотностей. Эталон (или базисный паттерн) распределения временного ряда представляет собой средневзвешенное состояние распределений фрагментов траектории случайного процесса, входящих в выделенный кластер. В таком подходе эталоны в виде функций распределения характеризуют типовые состояния изучаемой системы.

Текущее состояние, распознаваемое по близости выборочного распределения к эталонному паттерну в определенной норме, может относиться как к локально установившемуся эталонному состоянию, так и к переходному состоянию. Это последнее, в свою очередь, может быть близким к новому эталону, а может представляться в виде линейной комбинации уже имеющихся эталонов. Тем самым возникает задача оптимального, т.е. с наименьшей ошибкой, разложения текущего состояния по базисным паттернам. Представляет интерес ситуация, когда любое выборочное состояние может быть с заданной точностью представлено как линейная комбинация базисных паттернов. В результате такого разложения будет решена задача байесовского распознавания и одновременно указана вероятность, с которой состояние может относиться к какому-то другому типу.

Пусть векторы $\{\varphi_1,...,\varphi_m\}$, $\varphi_s \in R^n$ представляют собой гистограммы в количестве m штук, каждая из которых содержит n классовых интервалов, так что $\varphi_s(k)$ есть вероятность попадания наблюдаемого значения x в k-й классовый интервал при условии, что система находится в состоянии s. Тогда

$$\forall s \in \{1,...,m\} \sum_{k=1}^{n} \varphi_s(k) = 1. \qquad (1)$$

Пусть также вектор $\mathbf{f} \in R^n$, являющийся аналогичной гистограммой вероятностей, определяемой по текущей выборочной траектории длины N, принадлежит выпуклой оболочке векторов $\{\varphi_1,...,\varphi_m\}$, так что

$$\mathbf{f}(k) = \sum_{s=1}^{m} y_s \varphi_s(k), \quad 0 \le y_s \le 1. \qquad (2)$$

Тогда из (1) и (2) следует, что $\sum_{s=1}^{m} y_s = 1$.

Если разложение (2) получено, то, определив номер

$$s^* = \arg\max y_s, \qquad (3)$$

строим решающее правило байесовского распознавания: состояние s^* отвечает наблюденному распределению $\mathbf{f}$. Заметим, что в этом случае в любой норме расстояние $\left\|\mathbf{f} - \varphi_{s^*}\right\|$ минимально среди всех расстояний $\rho_s = \left\|\mathbf{f} - \varphi_s\right\|$. Обратное, очевидно, неверно, поскольку только из наличия минимального из расстояний не следует существование выпуклой комбинации вида (2).

В работе [3] на примерах биржевых рядов было показано, что хотя типовые паттерны состояний определяются с достаточной точностью, приближенное разложение вектора текущего состояния по паттернам, составляющим базис состояний временного ряда, во многих случаях не имеет вероятностной интерпретации: коэффициенты разложения либо отрицательны, либо больше единицы. Тем не менее, распознавание текущей ситуации по правилу

$$s^* = \arg\min \|\mathbf{f} - \varphi_s\| \qquad (4)$$

продолжало оставаться весьма точным. Ошибочное распознавание по формуле (4) получалось примерно в 5 % случаев, тогда как из-за потери вероятностной интерпретации метод (3) оказывался неприменим в 25 % случаев. Разумеется, для биржевых рядов разложение вида (2) является приближенным, невязка в гистограммной норме составляет примерно 0,1. Это происходит вследствие того, что вероятностное пространство состояний по факту не полно. Но также было выяснено, что увеличение количества базисных паттернов приводит к тому, что метод (3) становится неприменим в еще большем числе случаев. В [3] было высказано предположение, что наблюдаемый эффект в большей степени вычислительный, поскольку базисные паттерны близки между собой и матрица Грама базиса оказывалась плохо обусловленной. В результате проецирование из пространства большой размерности n (число классовых интервалов гистограммы) в пространство малой размерности m (число паттернов) оказывалось неустойчивым относительно малых шевелений элементов гистограмм. Отчасти это действительно так, но все же выбор паттернов в примерах биржевых рядов связан с процедурой экспертного отбора определенных ситуаций, т.е. не вполне объективен. Желательно было бы получить более веские доказательства того, что вычислительная процедура может воспрепятствовать применимости байесовского распознавания (3), но не ухудшать точности распознавания по методу (4).

Замечательно, что есть объект, к которому могут быть применены «средства объективного контроля»: это литературные произведения писателей. В монографии [4] описаны методы статистического распознавания автора текста по близости вероятностного распределения текста по буквам и буквосочетаниям к определенному авторскому эталону. Тестирование точности распознавания автора по методу (4) для трехбуквенных распределений показало его высокую эффективность: из тестируемых 300 случайных авторов с числом собственных произведений не менее 10 не было получено ни одной ошибки в распознавании автора отдельного текста. Такой статистический эксперимент позволяет детально сравнить разные методы распознавания и продемонстрировать ограничения каждого из них.

В настоящей работе исследуются вопросы точности байесовского распознавания на примерах литературных текстов достаточной длины (более 30 тыс. знаков), т.е. не просто текстов, а произведений, написанных конкретными писателями, и обсуждается парадокс сравнительно плохого распознавания по методу максимального правдоподобия и весьма точного распознавания (с нулевой эмпирической ошибкой) по методу ближайшего эталона.

1. Разложение по базисным состояниям

В общем виде задача разложения вектора $\mathbf{f} \in R^n$ по заданному набору линейно независимых векторов $\{\varphi_1,...,\varphi_m\}$, $\varphi_s \in R^n$ сводится к нахождению вектор-строки $\mathbf{y}^T = (y_1,...,y_m)$, минимизирующей в смысле 2-нормы функционал $\|\mathbf{f} - \Phi\mathbf{y}\|$, где $\Phi_{n\times m}$ есть матрица, столбцы которой составляют векторы φ_s. Минимизация этого функционала осуществляется ортогональным проектированием вектора $\mathbf{f}$ на m-мерное подпространство, натянутое на векторы $\{\varphi_1,...,\varphi_m\}$. Это проектирование представляет собой так называемое QR-разложение матрицы Φ в произведение специальной матрицы $Q_{n\times m}$, такой, что $Q^TQ = I_{m\times m}$, и верхней треугольной матрицы $R_{m\times m}$. В результате такого разложения получается следующее представление вектора $\mathbf{f}$:

$$\mathbf{f} - \Phi\mathbf{y} = I\mathbf{f} - QR\mathbf{y} = \left(I - QQ^T + QQ^T\right)\mathbf{f} - QR\mathbf{y} = Q\left(Q^T\mathbf{f} - R\mathbf{y}\right) + \left(I - QQ^T\right)\mathbf{f}. \qquad (5)$$

Векторы, в виде суммы которых в последнем равенстве (5) представлено данное разложение, ортогональны:

$$\left(R\mathbf{y} - Q^T\mathbf{f}\right)^T Q^T\left(I - QQ^T\right)\mathbf{f} = \left(R\mathbf{y} - Q^T\mathbf{f}\right)^T\left(Q^T{}_{p\times n}I_{n\times n} - I_{p\times p}Q^T{}_{p\times n}\right)\mathbf{f} =$$
$$= \left(R\mathbf{y} - Q^T\mathbf{f}\right)^T\left(Q^T{}_{p\times n} - Q^T{}_{p\times n}\right)\mathbf{f} = \left(R\mathbf{y} - Q^T\mathbf{f}\right)^T O_{p\times n}\mathbf{f} = \mathbf{o}.$$

Второе слагаемое в (5) не зависит от коэффициентов разложения $\mathbf{y}$. Следовательно, с учетом ортогональности указанных слагаемых, минимальное по $\mathbf{y}$ значение нормы $\|\mathbf{f} - \Phi\mathbf{y}\|$ равно норме этого второго слагаемого и достигается тогда, когда первое слагаемое равно нулю: $R\mathbf{y} - Q^T\mathbf{f} = \mathbf{o}$.

Итак, оптимальное разложение определяется вектором

$$\mathbf{y}_{opt} = R^{-1}Q^T\mathbf{f}. \qquad (6)$$

Величина

$$\mathbf{r} = \mathbf{f} - \Phi\mathbf{y}_{opt} = \left(I - QQ^T\right)\mathbf{f} \qquad (7)$$

есть невязка разложения (5). Ошибкой разложения считается 2-норма невязки, т.е. величина $\delta = \|\mathbf{r}\| = \left\|\left(I - QQ^T\right)\mathbf{f}\right\|$. Относительная ошибка определяется как

$$\varepsilon = \frac{\delta}{\|\mathbf{f}\|} = \frac{\left\|\left(I - QQ^T\right)\mathbf{f}\right\|}{\|\mathbf{f}\|}. \qquad (8)$$

Пусть теперь, как это и бывает на практике, вектор текущего состояния $\mathbf{f}$ и матрица Φ базисных паттернов известны неточно. Неточность здесь имеет не измерительную, а статистическую природу, поскольку вместо генеральных совокупностей приходится иметь дело с выборочными распределениями. Возникает вопрос: как эта неточность повлияет на вычисление оптимального

разложения, насколько эта процедура устойчива к малым возмущениям, какова в этом случае невязка? Положим

$$\xi = \max\left(\frac{\|\Delta\Phi\|}{\|\Phi\|}, \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}\right) \tag{9}$$

и введем число обусловленности $\kappa(\Phi)$ матрицы Φ в смысле 2-нормы как отношение наибольшего и наименьшего ее сингулярных чисел. Поскольку матрица Φ по построению имеет полный столбцевой ранг, ее наименьшее сингулярное число строго больше нуля. Однако если базисные векторы оказываются близкими, то число обусловленности может быть очень большим. Согласно [5], 2-норма относительной вариации оптимального разложения оценивается сверху следующим образом:

$$\frac{\|\Delta\mathbf{y}\|}{\|\mathbf{y}\|} \le \xi \cdot \left(\frac{2\kappa(\Phi)}{\cos\theta} + \kappa^2(\Phi) tg\theta\right) + O\left(\xi^2\right), \tag{10}$$

где $\sin\theta = \varepsilon$ есть синус угла между раскладываемым вектором $\mathbf{f}$ и вектором $\Phi\mathbf{y}_{opt}$ в соответствии с (8). В результате может оказаться так, что разложение, например, в двумерное подпространство является более точным, чем в трехмерное. Это связано с тем, что если раскладывается n-мерный вектор, лежащий в плоскости двух базисных паттернов, но определенный с ошибкой порядка ξ (9), то возможен вычислительный артефакт нахождения большой проекции на третий паттерн согласно (10). При этом существенной оказывается и последовательность, в которой исследуемый вектор проецируется на эталонные распределения.

Рассмотрим в этом контексте разложение вероятностных распределений буквосочетаний литературного текста по известным авторским эталонам.

2. Статистические свойства авторских эталонов

Приведем здесь некоторые результаты статистического анализа по построению авторских эталонов, т.е. базисных паттернов, с которыми надо будет сравнивать отдельные тексты. Под текстом в нашем анализе понимается совокупность вероятностей буквосочетаний определенной длины (1, 2, 3 и т.д.), если текст представить как одну строку без пробелов и знаков препинания. Для краткости совокупность вероятностей сочетаний из n букв, идущих в тексте подряд, будем обозначать n-ПФР (плотность функции распределения). Так, например, 1-ПФР есть просто распределение текста по буквам (массив из 33 символов), 2-ПФР представляет распределение текста по парам букв («аа», «аб», «ав»,…, «ба», …, всего $33^2 = 1089$ символов) и т.д.

Задача идентификации автора неизвестного текста состоит в следующем [4, 6]. Имеется библиотека, содержащая тексты, представленные в виде ПФР для A известных авторов. Пусть K_a – имеющееся количество текстов a-го

автора, и $N_{i,a}$ – количество букв в i-м тексте этого автора, $i = 1,2,...,K_a$. Обозначим $f_{i,a}^{(n)}(j)$ n-ПФР соответствующего текста, где аргумент j меняется от 1 до $J(n) = 33^n$.

Для каждого автора определим его средневзвешенную n-ПФР, которую трактуем как авторский эталон:

$$F_a^{(n)}(j) = \frac{1}{N_a}\sum_{i=1}^{K_a} f_{i,a}^{(n)}(j) N_{i,a}, \quad N_a = \sum_{i=1}^{K_a} N_{i,a}, \quad j = 1,2,...,J(n). \tag{11}$$

В (11) для краткости записи формул мы пренебрегли единицей по сравнению с N_a при подсчете пар букв в тексте для 2-ПФР (или двойкой при подсчете триграмм, если речь идет о 3-ПФР и т.д.), т.к. $N_a >> n$. Также для упрощения записи мы будем опускать верхний индекс n в выражении n-ПФР, если указание на длину массива несущественно.

Введем «библиотечную норму» ρ_{ik} как расстояние между ПФР текстов i и k в норме суммируемых функций:

$$\rho_{ik} = \| f_i - f_k \| = \sum_{j=1}^{J(n)} | f_i(j) - f_k(j) |. \tag{12}$$

Для каждого автора a построим плотность функции распределения $g_a^+(\rho)$ отклонений $\rho_{i_a,a}$ «его» текстов, а также распределение $g_a^-(\rho)$ отклонений $\rho_{k_b,a}$ «чужих» произведений от его средней ПФР F_a, определенной в (11). Построение этих плотностей реализовывалось в статистическом эксперименте следующим образом. Были взяты 30 авторов, классиков и современных, имеющих по 10 крупных произведений (более 60 тыс. знаков), и для них строились эталонные 3-ПФР по формуле (11). Затем тестировалось отклонение текстов от «своих» и «чужих» эталонов. Естественно, на момент сравнения «свое» произведение исключалось из эталона (11), так что ПФР i-го произведения a-го автора сравнивалась с его квази-эталоном:

$$F'_{i,a}(j) = \frac{1}{1 - N_{i,a}/N_a}\left(F_a(j) - f_{i,a}(j)\frac{N_{i,a}}{N_a} \right), \tag{13}$$

$$\| f_{i,a} - F'_{i,a} \| = \frac{1}{1 - N_{i,a}/N_a}\sum_{j=1}^{J(n)} | f_{i,a}(j) - F_a(j) | = \frac{\| f_{i,a} - F_a \|}{1 - N_{i,a}/N_a}. \tag{14}$$

Те авторы, которые пишут примерно в одном стиле («сериальные» писатели), имеют узкое распределение $g_a^+(\rho)$, а у классиков более широкое распределение расстояний от своих текстов до эталона. Распознать последних авторов можно только в том случае, если расстояния от их текстов до чужих

эталонов заметно превосходят характерное расстояние до своего эталона. Нам важно сравнить эти распределения в целом по группе разных авторов.

Средние плотности $g^{\pm}(\rho)$ для 3-ПФР совокупности авторов приведены на рис. 1. Из графиков видно, что если расстояния между текстами и эталонами полностью случайны, то точность распознавания автора ожидается не очень высокой в силу довольно значительного перекрытия плотностей, отвечающих своим и чужим текстам.

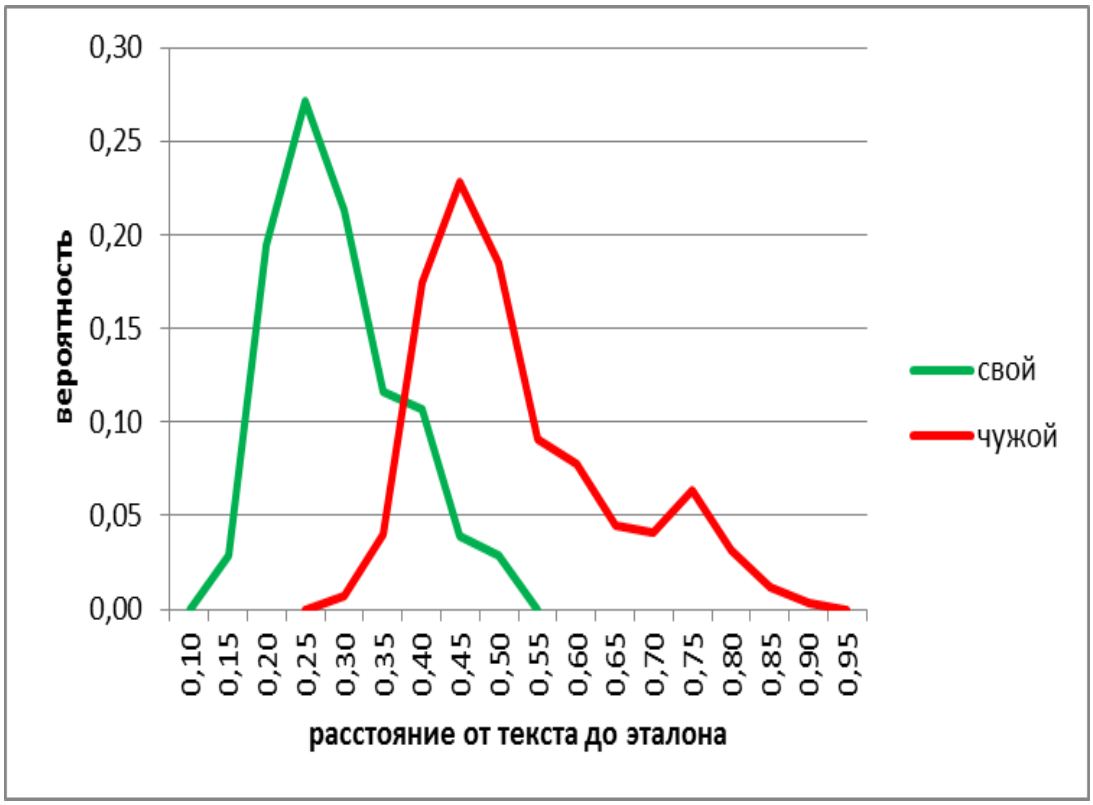

Рис. 1. Среднее распределение расстояний между 3-ПФР текста и эталона

Обозначим через $G^{\pm}(\rho)=\int_0^{\rho} g^{\pm}(r)dr$ соответствующие интегральные функции распределения расстояний между текстами и эталонами. Минимальное значение ρ, при котором $G^{+}(\rho)=1$, обозначим ρ^{+}, а максимальное значение ρ, при котором $G^{-}(\rho)=0$, обозначим ρ^{-}. Смысл введенных величин в том, что все ПФР текстов находятся на расстоянии не более ρ^{+} от соответствующих авторских эталонов и на расстоянии не менее ρ^{-} от чужих эталонов. Величина $1-G^{+}\left(\rho^{-}\right)$ есть вероятность ошибочно признать за произведение автора «*a*» чужой текст (ошибка второго рода, пропуск цели), а величина $G^{-}\left(\rho^{+}\right)$ есть вероятность ошибочно отвергнуть произведение автора «*a*», посчитав его за чужое (ошибка первого рода, ложная тревога). Назовем

расстоянием разделения авторов такое значение ρ^*, для которого ошибка идентификации автора текста минимальна:

$$\rho^* = \arg\min\left(1 - G^+(\rho) + G^-(\rho)\right) = \arg\max\left(G^+(\rho) - G^-(\rho)\right). \qquad (15)$$

Построенная величина может служить верхним уровнем для кластеризации текстов по авторам или для кластеризации текстов одного автора по жанрам.

Некоторые характеристики распределений рис. 1 приведены в табл. 1.

Табл. 1. Характеристики распределений расстояний между текстом и эталоном

Показатель	1-ПФР	2-ПФР	3-ПФР
Среднее значение l_s (свой текст)	0,04	0,13	0,27
Среднее значение l_d (чужой текст)	0,08	0,23	0,50
Стандартное отклонение σ_s (свой текст)	0,02	0,04	0,08
Стандартное отклонение σ_d (чужой текст)	0,03	0,07	0,12
Расстояние разделения ρ^*	0,05	0,19	0,35
$\rho+$	0,11	0,30	0,55
$\rho-$	0,03	0,10	0,25
Вероятность ошибки I: $G^-(\rho+)$	0,47	0,65	0,72
Вероятность ошибки II: $1-G^+(\rho-)$	0,72	0,57	0,50

Из табл. 1 следует, что ключевая для распознавания ошибка II рода (пропуск цели) уменьшается с увеличением размерности ПФР. Подчеркнем, что оцененные величины ошибок идентификации будут наблюдаться на практике, если расстояния до своих и чужих эталонов случайны. Будем строить решающую функцию распознавания по методу (4). Именно, пусть имеется текст «0» неизвестного автора, который надо идентифицировать внутри данной библиотеки. Автором текста «0» считается тот из авторов «*a*», для которого норма $\rho_a^0 = \| f_0 - F_a \|$ разности между ПФР $f_0(j)$ текста «0» и авторского эталона $F_a(j)$ минимальна:

$$\rho_a^0 = \| f_0 - F_a \|, \quad a^0 = \arg\min_a \rho_a^0. \qquad (16)$$

Оказалось, что ошибка идентификации авторов текстов по выбранной группе равна нулю, а вовсе не 0,5, как это следует из последней строки табл. 1. Причина такой высокой точности выясняется из рис. 2.

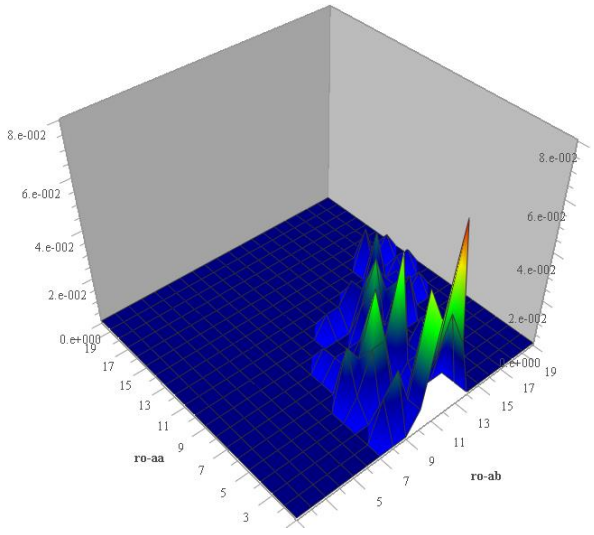

Рис. 2. Совместное распределение расстояний до своих и чужих эталонов

На рис. 2 показано двумерное распределение вероятностей расстояний от текста до своего эталона (левая ось «aa») и одновременно до чужого эталона (правая ось «ab»). Поскольку весь носитель этого распределения лежит справа от главной диагонали, то нет примеров, когда расстояние до своего эталона оказалось больше, чем до чужого. Следовательно, можно построить решающее правило с нулевой (для данного примера) ошибкой.

Такой высокоточный результат естественным образом наводит на мысль, что в его основе лежит байесовское решающее правило (3), а авторские эталоны являются базисными паттернами, по которым можно разложить любой текст. Именно, если $\mathbf{f}$ есть ПФР неизвестного текста и матрица Ф авторских эталонов задана, то рассматривается аппроксимация $\tilde{\mathbf{f}}$ вектора $\mathbf{f}$ по этой системе: $\tilde{\mathbf{f}} = \Phi \mathbf{y}_{opt}$. Коэффициенты разложения $\mathbf{y}_{opt}$, определяемые по формуле (6), дадут вероятности того, что данный текст написан тем или иным автором, максимальный коэффициент укажет на наиболее вероятного автора. Заметим, что если расстояние до этого автора окажется больше, чем расстояние разделения ρ^* из табл. 1, то следует сделать вывод, что соответствующего эталона в библиотеке, скорее всего, нет.

В результате тестирования оказалось, что разложение по базисным паттернам является в этом примере крайне неточной операцией. Некоторые парадоксальные свойства метода (3) в сравнении с методом (4) обсуждаются ниже. Здесь же отметим следующее. Распознавание по методу (4) можно напрямую свести к байесовскому, если использовать критерий Колмогорова-Смирнова о принадлежности выборки (т.е. данного текста) определенной генеральной совокупности (одному из авторских эталонов). По сути, это использование метода (3) в норме С, для которой есть априорное распределение вероятностей отклонения выборочной функции распределения от эталона. Однако точность этого критерия заметно хуже, чем метода сравнения ПФР текстов в норме L1. Это связано не только с тем, что сравнение в норме С

несколько менее точное, чем в L1, но и с тем, что появление букв в тексте не полностью случайно, и потому сходимость выборочных функций распределения к эталонному распределению автора не имеет асимптотики в виде функции Колмогорова.

3. Точность разложения ПФР текста по базисным паттернам

Рассмотрим теперь следующий статистический эксперимент. Спроецируем по методу, описанному в п.1, векторы *n*-ПФР отдельных текстов, авторы которых и только они находятся среди эталонов, составляющих матрицу Φ. Судя по результативности метода (4) распознавания авторов по близости к эталону, следовало бы ожидать, что в результате проецирования на эталоны по формулам (5-6) будет получено разложение, коэффициенты которого имеют вероятностную интерпретацию. Возможно, сумма коэффициентов при этом будет несколько отличаться от единицы, но естественно ожидать, что отличие это будет иметь порядок невязки (7-8).

На практике же выяснилось, что проецирование вектора *n*-ПФР размерности 33^n в пространство малой размерности (согласно числу авторских эталонов) является весьма неточной операцией и, более того, коэффициенты разложения зачастую не имеют предполагаемого вероятностного смысла.

В табл. 2 приведены величины среднеквадратичного отклонения коэффициентов $\mathbf{y}_{opt}$ (в среднем по текстам авторов) от промежутка $[0;1]$ в зависимости от числа авторских паттернов (2, 3 или 4) и в зависимости от длины вектора ПФР (33, 33^2 и 33^3). Этот показатель увеличивается при увеличении числа авторов и для четырех эталонов становится практически неприемлемым. Увеличение среднеквадратичного отклонения связано с тем, что эталоны всех авторов весьма близки между собой, т.е. как векторы они почти коллинеарны. Так, например, для двух эталонов 1-ПФР определитель матрицы Грама имеет порядок 10^{-5}, а для трех эталонов он уже равен 10^{-9}. Косинус угла между любой парой эталонов 1-ПФР равен примерно 0,995, поэтому численное решение плохо обусловленной системы даже из двух линейных уравнений приводит к заметным ошибкам.

В то же время рассматриваемая величина среднеквадратичного отклонения уменьшается при увеличении размерности вектора ПФР. Это связано с тем, что с увеличением размерности ПФР эффект коллинеарности эталонов снижается, поскольку частоты употребления отдельных букв более устойчивы по разным авторам, чем пар букв, а те, в свою очередь, более устойчивы, чем тройки букв и т.д. Но тем не менее эталоны все же остаются близкими между собой. Например, для 2-ПФР косинус угла между парой эталонов имеет величину 0,977, а для 3-ПФР он равен 0,910.

Табл. 2. Среднеквадратичное нарушение вероятностной интерпретации

Число авторов	**2**	**3**	**4**
1-ПФР	0,15	0,26	0,54
2-ПФР	0,12	0,20	0,35
3-ПФР	0,09	0,18	0,29

Определенная неожиданность состоит в том, что алгебраическая сумма коэффициентов разложения $\mathbf{y}_{opt}$ (т.е. несохранение нормировки функции распределения при аппроксимации) ведет себя противоположно отмеченной выше тенденции, характерной для отдельных коэффициентов того же разложения. Именно, с увеличением числа эталонов несохранение нормировки в среднем уменьшается, а с увеличением размерности вектора вероятности – увеличивается (табл. 3). Такая же тенденция имеет место и для среднеквадратичной невязки (7), что видно из табл. 4: при увеличении количества авторов невязка уменьшается, а при увеличении размерности вектора ПФР невязка увеличивается.

Табл. 3. Среднеквадратичное несохранение вероятности

Число авторов	**2**	**3**	**4**
1-ПФР	0,006	0,005	0,005
2-ПФР	0,011	0,009	0,008
3-ПФР	0,028	0,027	0,025

Табл. 4. Среднеквадратичная невязка аппроксимации

Число авторов	**2**	**3**	**4**
1-ПФР	0,060	0,056	0,052
2-ПФР	0,171	0,163	0,161
3-ПФР	0,352	0,335	0,322

Однако из того, что наименьшая и вполне приемлемая невязка (кстати, примерно равная расстоянию разделения текстов ρ^* из табл. 1) достигается для 1-ПФР, не следует, что этот инструмент дает наиболее точное распознавание. Наиболее точное распознавание автора получается при использовании 3-ПФР, а для 1-ПФР оно как раз самое плохое. Вычислительно это связано с тем, что число обусловленности увеличивается с ростом числа эталонов и уменьшается с увеличением размерности ПФР (табл. 5), т.е. ведет себя примерно так же, как и среднеквадратичное нарушение вероятностной интерпретации (табл. 2). Однако то, что среднеквадратичная невязка ведет себя противоположно изменению числа обусловленности, вызывает большое удивление. Обычно в вычислительном эксперименте неточность аппроксимации с использованием плохо обусловленной матрицы заметно выше, чем в случае матрицы с близкими сингулярными числами.

Табл. 5. Округленное число обусловленности
матрицы Грама системы эталонов

Число авторов	**2**	**3**	**4**
1-ПФР	600	1800	3500
2-ПФР	100	300	560
3-ПФР	20	70	110

Таким образом, налицо парадоксальность поведения вычислительной точности байесовского метода распознавания образов на примере близких эталонов.

Важно понимать, что ошибка разложения вектора по системе паттернов связана не только с приближенностью вычислений, но и с тем, что сами паттерны известны не точно в статистическом смысле, поскольку они всего лишь приближают некую гипотетическую генеральную совокупность, которой предположительно обладает автор текста.

В [4, 6] была определена достаточная длина текста для того, чтобы говорить о статистической точности в оценке соответствующего авторского эталона в смысле *n*-ПФР. Пусть эталон автора $F_{N,a}^{(n)}(j)$, определенный в (11), построен по произведениям этого автора совокупной длины N знаков. Тогда условие его отклонения от генерального распределения $U_a^{(n)}(j)$ в норме L1 на величину ε, имеющее вид

$$\sum_{j=1}^{J(n)} \left| F_{N,a}^{(n)}(j) - U_a^{(n)}(j) \right| \le \varepsilon, \qquad (17)$$

достигается при совокупной длине текста, не меньшей, чем $N^*(\varepsilon)$, а эта длина определяется как решение относительно N уравнения

$$\frac{t_{1-\varepsilon/2}}{\varepsilon} = \frac{\sqrt{N}}{\Sigma_N(n)}, \tag{18}$$

где $t_{1-\varepsilon/2}$ есть квантиль распределения Стьюдента порядка $N-1$, а $\Sigma_N(n)$ определяется эмпирической частотой $F_{N,a}^{(n)}(j)$:

$$\Sigma_N(n) = \sum_{i=1}^{J(n)} \sqrt{F_{N,a}^{(n)}(j) \cdot \left(1 - F_{N,a}^{(n)}(j)\right)}. \tag{19}$$

Величина $\Sigma_N(n)$ при фиксированном N приблизительно одинакова для всех авторов. Результаты расчетов статистической погрешности в оценке авторских эталонов, т.е. в оценке элементов матрицы Φ, приведены в табл. 6.

Табл. 6. Значения $\Sigma_N(n)$ и точность оценки *n*-ПФР

	1-ПФР	2-ПФР	3-ПФР	4-ПФР
$\Sigma_N(n)$	5	18	51	105
Значения ε по формуле (18)				
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

Возникает вопрос: с какой точностью необходимо знать элементы матрицы Φ, чтобы различать ее собственные значения, т.е. чтобы корректно получать численные решения относительно коэффициентов разложения $\mathbf{y}_{opt}$? Для ответа надо знать ε-спектр матрицы Грама базиса из этих эталонов, т.е. области расположения собственных значений в зависимости от возможной нормы возмущения элементов матрицы Грама. Алгоритм построения ε-спектра матрицы описан в монографии С.К. Годунова [7]. Ниже на рис. 3-5 приведены спектральные портреты матриц Грама для четырех эталонов, представленных в виде соответственно 1,2,3-ПФР. Здесь мы считаем своим приятным долгом поблагодарить к.ф.-м.н. О.Б. Феодоритову за консультации по вопросам построения ε-спектра матрицы и проведению соответствующих вычислений.

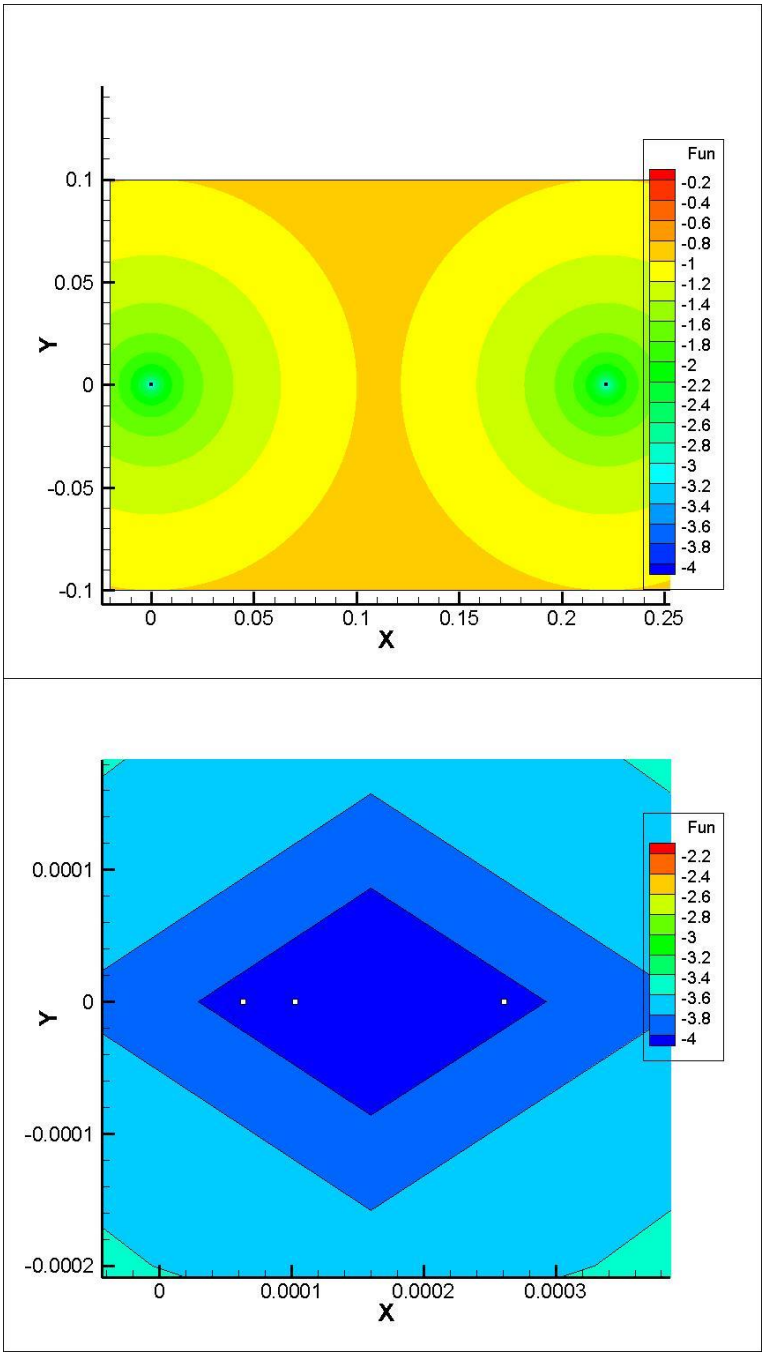

Рис. 3. Спектральный портрет 1-ПФР, 4 эталона

Вычисленные собственные значения: 0.22172E+00, 0.26068E-03, 0.10266E-03 и 0.62983E-04. Из первого рис. 3 видно, что при точности 10^{-4} (правая шкала легенды) спектр еще полностью не разделяется. Три значения, близких к нулю, вынесены на второй рисунок. Области, в которых находится спектр при определенной точности относительно элементов матрицы, окрашены в один цвет.

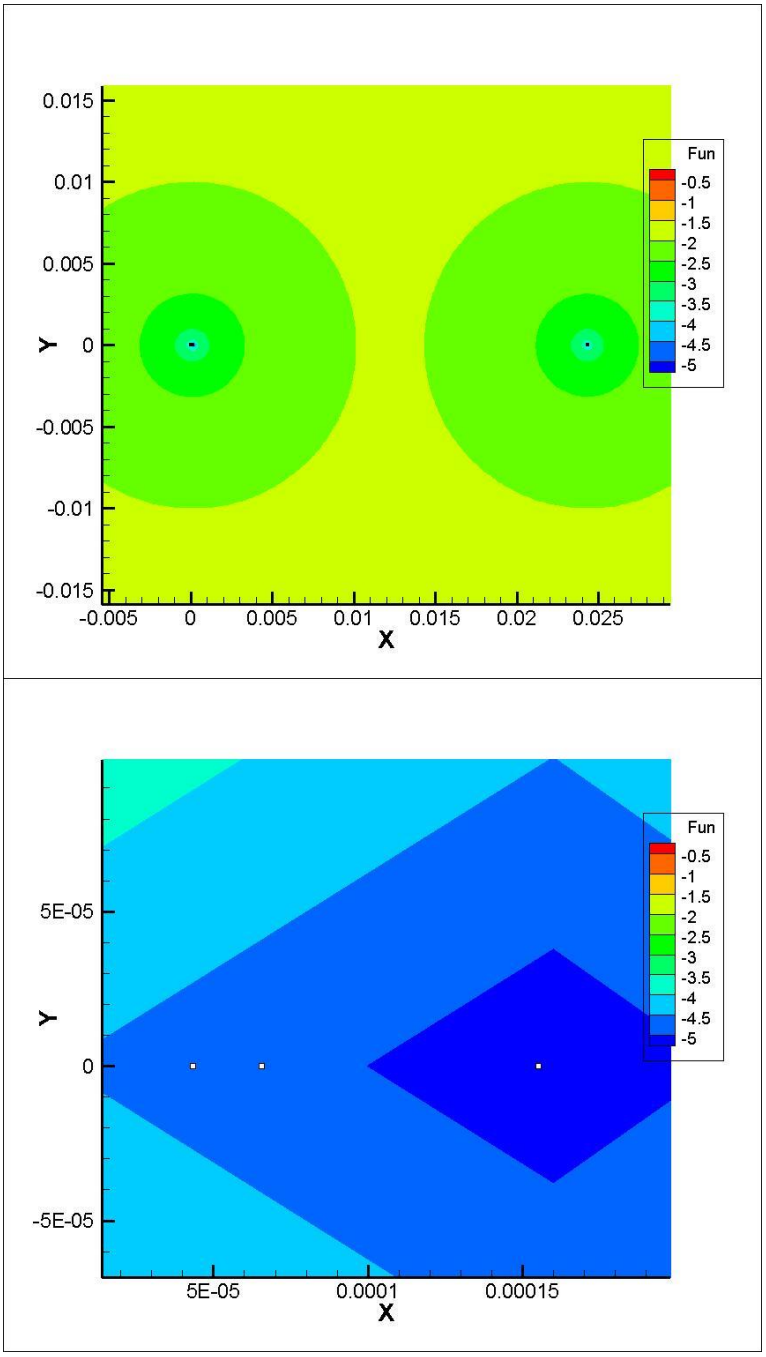

Рис. 4. Спектральный портрет 2-ПФР, 4 эталона

Собственные значения матрицы Грама для 2-ПФР: 0.24333E-01, 0.15504E-03, 0.65528E-04 и 0.43401E-04. Отметим, что хотя число обусловленности для 2-ПФР меньше, чем для 1-ПФР (табл. 5), ее спектр разделяется на два значения при точности 10^{-2}, тогда как для 1-ПФР это достигалось уже при точности 10^{-1}. Для 3-ПФР аналогичное разделение наступает только при точности 10^{-3}.

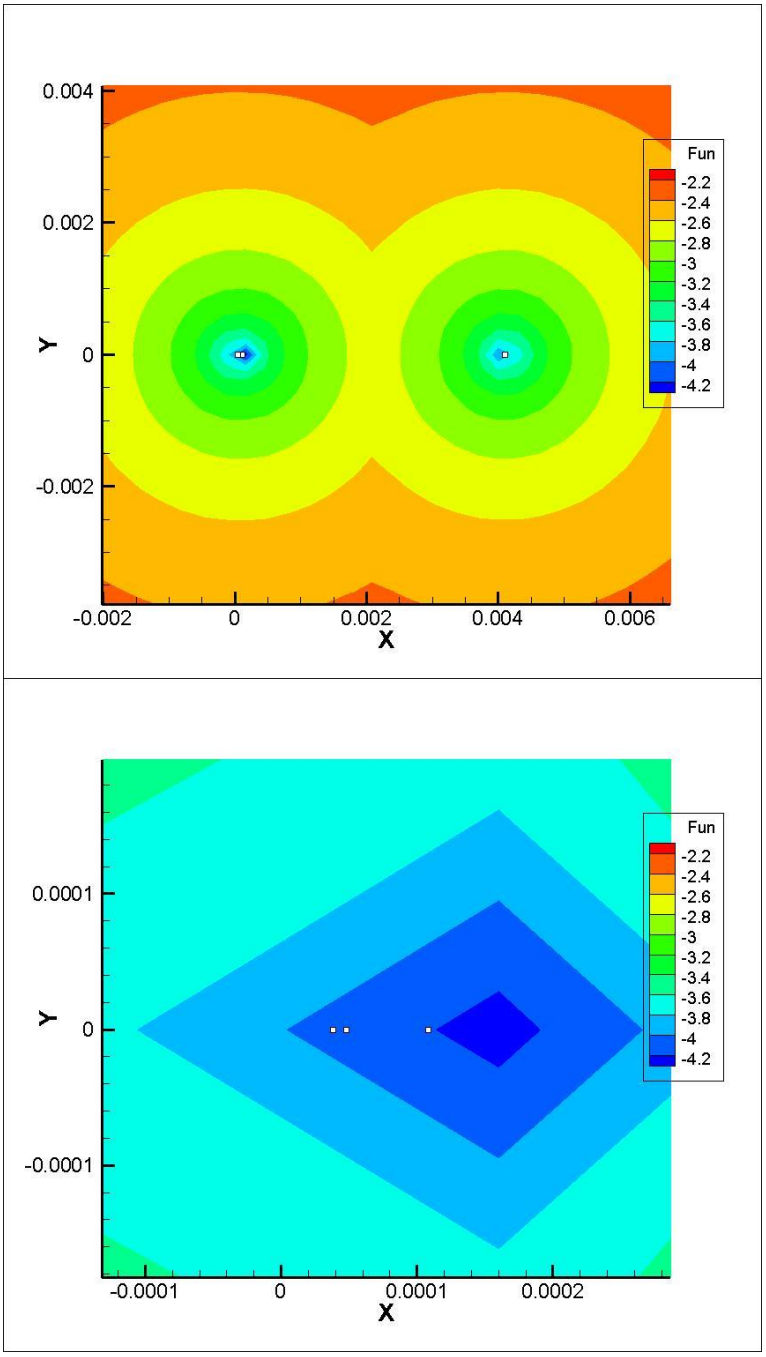

Рис. 5. Спектральный портрет 3-ПФР, 4 эталона

Собственные значения матрицы Грама для 3-ПФР: 0.41078E-02, 0.10802E-03, 0.47800E-04 и 0.38349E-04.

Сравнивая спектральные требования к точности элементов матрицы Ф с оценками, приведенными в табл. 6, видим, что даже при длине эталона в 1 млн знаков спектр не может быть достоверно различим из статистических соображений. Для различения собственных значений для 3-ПФР на уровне 10^{-5} длина эталона должна быть более 100 млрд знаков, тогда как характерная длина эталона из 10 романов примерно 5 млн знаков. Следовательно, проекции на эталонные базисные паттерны не могут быть вычислительно устойчивыми. Поэтому метод (3) определения максимальной вероятности образа не дает в

рассматриваемом примере практически полезных результатов. В то же время метод (4) распознавания по близости к эталону дает поразительно точную решающую функцию.

Проведенное исследование было направлено на то, чтобы подчеркнуть различие между теоретическим методом распознавания и его практической реализацией, которая в силу применяемых вычислительных процедур может оказаться бесполезной. Анализ погрешности вычислений является ключевым, поскольку даже в ситуации с абсолютным распознаванием можно получить вычислительные артефакты, препятствующие применению метода.

Литература

1. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. – М.: Наука, 1974. – 416 с.

2. Босов А.Д., Орлов Ю.Н., Федоров С.Л. О распределении рядов абсолютных приростов цен на финансовых рынках // Препринты ИПМ им. М.В. Келдыша. 2014. № 96. 15 с.
URL: http://library.keldysh.ru/preprint.asp?id=2014-96

3. Кирина-Лилинская Е.П., Орлов Ю.Н., Федоров С.Л. Метод базисных паттернов в анализе нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша. 2016. № 7. 20 с.
URL: http://library.keldysh.ru/preprint.asp?id=2016-7

4. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.

5. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения (пер. с англ.). – М.: Мир, 2001. – 436 с.

6. 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.
URL: http://library.keldysh.ru/preprint.asp?id=2013-27

7. Годунов С.К. Современные аспекты линейной алгебры. – Новосибирск: Научная книга, 1997. – 388 с.