Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании*

В. Р. Целых[1]***, К. В. Воронцов***[2]
celyh@inbox.ru, voron@forecsys.ru
1 — Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»
2 — Вычислительный центр РАН

Критерий согласия Пирсона неприменим к сильно разреженным распределениям, так как в этих случаях распределение статистики плохо описывается асимптотическим законом хи-квадрат, зависит от объёма выборки и вида исходного распределения. В данной работе предлагаются статистические критерии, основанные на вычислении эмпирических распределений статистик путём сэмплирования. Рассматривается их применение в задачах анализа текстов, в частности, для проверки гипотезы условной независимости при построении и оценивании вероятностных тематических моделей.

Ключевые слова: *критерий согласия, статистика хи-квадрат, сэмплирование, метод Монте-Карло, закон Ципфа, вероятностная тематическая модель, гипотеза условной независимости.*

Goodness-of-fit tests for sparse multinomial distributions with application to topic modeling*

V. R. Tselykh[1]***, K. V. Vorontsov***[2]
1 — Moscow Institute of Physics and Technology
2 — Computing Center of the Russian Academy of Sciences

Pearson's goodness-of-fit test is not appropriate for sparse multinomial distributions. In this case the distribution of statistic is not asymptotically chi-squared, depends on a sample size and on a form of the tested distribution. The article suggests statistical criteria based on empirical distribution of a statistic obtained from sampling. Their application to text analysis is considered, in particular, to testing the conditional independence hypothesis for probabilistic topic models evaluation.

Keywords: *goodness-of-fit test, chi-squared statistics, sampling, Zipf's law, probabilistic topic model, conditional independence.*

Введение

Стандартные критерии согласия для дискретных распределений плохо подходят, когда число возможных значений наблюдаемой переменной значительно превосходит число наблюдений, либо когда многие значения имеют близкие к нулю вероятности [12, 11]. Такие распределения называют разреженными. В этих случаях распределение статистики не описывается классической асимптотикой, может зависеть от объёма выборки и степени разреженности исходного распределения.

Разреженные распределения возникают в задачах статистического анализа текстов, когда текст рассматривается как дискретное распределение на множестве слов, и требу-

Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.

ется проверить, является ли заданный фрагмент текста случайной выборкой из известной генеральной совокупности. В данной работе рассматривается ситуация, когда генеральная совокупность фиксирована, и требуется проверять много фрагментов. Тогда становятся оправданными методы сэмплирования, выполняющие большой объём вычислений на этапе предварительной обработки генеральной совокупности.

Предлагаются непараметрические статистические тесты, основанные на сэмплировании с возвращением и без возвращения, а также параметрический метод, предназначенный для проверки согласия с законом Ципфа. Для параметрического метода строится регрессионная модель, выражающая квантиль распределения через параметры задачи. Преимущество регрессионного теста в том, что, в отличие от методов сэмплирования, его не нужно перестраивать заново для каждого распределения.

Рассматривается применение предложенных статистических тестов для проверки гипотезы условной независимости при построении и оценивании вероятностных тематических моделей коллекций текстовых документов.

Критерий согласия хи-квадрат

Пусть имеется выборка n независимых наблюдений $\{x_1, \ldots, x_n\}$ случайной величины, принимающей значения из конечного множества Ω. Её эмпирическое распределение определяется как доля наблюдений x_i, равных x:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}[x_i = x], \quad x \in \Omega.$$

Критерий хи-квадрат проверяет гипотезу о том, что случайная величина имеет заданное распределение $p(x)$, $x \in \Omega$. Для этого вычисляется статистика хи-квадрат:

$$X^2 = n\sum_{x\in\Omega}\frac{(\hat{p}(x) - p(x))^2}{p(x)}. \tag{1}$$

Распределение статистики X^2 стремится к распределению хи-квадрат с $k = |\Omega| - 1$ степенями свободы: $X^2 \sim \chi^2(k)$. Нулевая гипотеза отвергается на уровне значимости α, если значение статистики превышает $(1-\alpha)$-квантиль этого распределения: $X^2 > \chi^2_{1-\alpha}(k)$.

Считается, что асимптотика хи-квадрат применима, если объём выборки $n \geqslant 50$ и ожидаемое число наблюдений $np(x) \geqslant 5$ для каждого $x \in \Omega$. В случаях *разреженных* распределений $p(x)$, когда вероятности $p(x)$ малы для многих $x \in \Omega$ или когда $|\Omega| \gg n$, второе условие может не выполняться даже на очень больших выборках [12]. Стандартная рекомендация — объединять значения $x \in \Omega$ в группы — для сильно разреженных распределений оказывается неприемлемой, так как результат существенно зависит от способа группирования, который выбирается произвольно.

В качестве иллюстрации рассмотрим распределение, называемое *законом Ципфа*:

$$p(x) = Ax^{-s}, \quad x \in \Omega = \{1, \ldots, v\}, \tag{2}$$

где $A = \left(\sum_{x=1}^{v} x^{-s}\right)^{-1}$ — нормировочный множитель, s — параметр. Этот закон неплохо описывает частоты слов в текстах на естественных языках, если за x принимать номера слов, упорядоченных по убыванию частоты. Параметр s зависит от языка и от корпуса текстов, по которому делается оценка, но в большинстве экспериментов значение s близко к 1 и находится в пределах от 0.9 до 1.2 [1, 6].

Чем больше значение параметра s и размер словаря v, тем более разрежено распределение $p(x)$. Проведём простой вычислительный эксперимент. Возьмём типичные значения параметра $s = 1$ и размера словаря $v \in \{50,\ 500,\ 1000,\ 5000\}$. Сгенерируем $N = 1000$ выборок (искусственных текстов) длины $n = 100$ из распределения (2), и для каждой выборки вычислим значение статистики X^2.

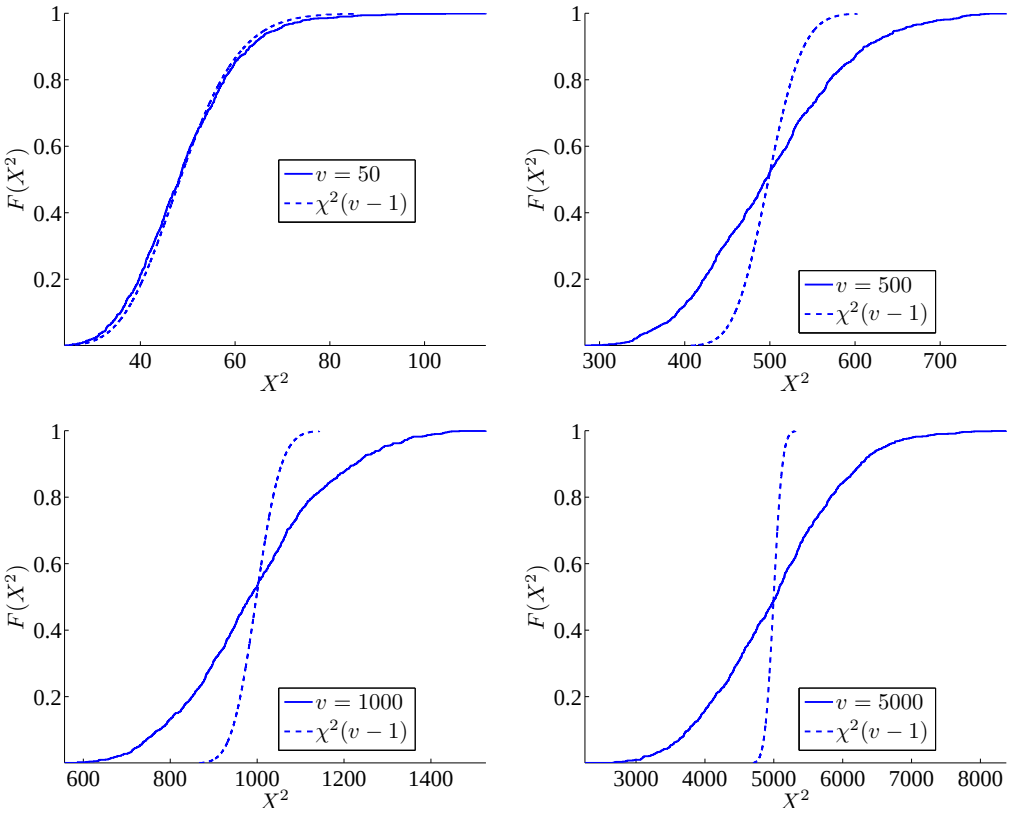

Рис. 1: Функции распределения статистики X^2 при $s = 1$, $v = 50,\ 500,\ 1000,\ 5000$, $n = 100$, $N = 1000$ и соответствующие функции распределения $\chi^2(v-1)$.

На рис. 1 сплошными линиями показаны эмпирические распределения статистики X^2, пунктирными линиями — распределения $\chi^2(v-1)$. Чем больше размер словаря, тем сильнее разрежено распределение $p(x)$, и тем сильнее отличаются $(1-\alpha)$-квантили этих распределений (при типичном значение $\alpha = 0.05$).

Таким образом, распределение хи-квадрат не может быть использовано в практических задачах анализа текстов, когда требуется проверить, является ли заданный текст $\hat{p}(x)$ случайной выборкой из корпуса текстов $p(x)$.

Тест на основе сэмплирования

Для разреженных распределений $p(x)$ предлагается вместо асимптотического распределения $\chi^2(k)$ статистики X^2 использовать эмпирическое распределение.

Построение теста. Генерируется N независимых выборок объёма n из заданного дискретного распределения $p(x)$. Для каждой выборки вычисляется эмпирическое распределение $\hat{p}_j(x)$, $j = 1, \ldots, N$ и значение статистики X_j^2 по формуле (1). По полученным значениям $X_1^2, \ldots, X_N^2$ строится эмпирическая функция распределения статистики

$$\hat{F}_n(X^2) = \frac{1}{N} \sum_{j=1}^{N} [X^2 > X_j^2]$$

и вычисляется её $(1-\alpha)$-квантиль $\hat{F}_{n,1-\alpha}$. Число N рекомендуется брать не менее 1000, если необходимо оценивать всю функцию распределения. Однако если оценивается только одна квантиль, N можно брать порядка нескольких десятков [11].

Применение теста. Пусть задана выборка объёма n, по которой построено эмпирическое распределение $\hat{p}(x)$ и вычислено значение статистики X^2 согласно (1). Если

Алгоритм 1 Построение теста путём рекуррентного вычисления значений статистики X^2 по N одновременно растущим выборкам объёма n.

Вход: $p(x)$, N, $n_{\max}$, α;
Выход: $\hat{F}_{n,1-\alpha}$ для всех $n = 1, \ldots, n_{\max}$;

для всех $j := 1, \ldots, N$
  сэмплировать первый элемент j-й выборки $\xi \sim p(x)$;
  инициализировать эмпирическую гистограмму для j-й выборки: $H_j(x) := [x = \xi]$ для всех $x \in \Omega$;
  инициализировать значение статистики X^2 для j-й выборки: $X^2_{j1} := 1/p(\xi) - 1$;
для всех $n := 1, \ldots, n_{\max} - 1$
  для всех $j := 1, \ldots, N$
    сэмплировать $(n+1)$-й элемент j-й выборки $\xi \sim p(x)$;
    обновить эмпирическую гистограмму для j-й выборки: $H_j(\xi) := H_j(\xi) + 1$;
    обновить значение статистики X^2 для j-й выборки:
$$X^2_{j,n+1} := \frac{nX^2_{j,n} + 1}{n+1} + \frac{2H_j(\xi) - 1}{(n+1)p(\xi)} - 2;$$
для всех $n := 1, \ldots, n_{\max}$
  упорядочить $X^2_{1,n}, \ldots, X^2_{N,n}$ по возрастанию;
  $\hat{F}_{n,1-\alpha} := X^2_{N(1-\alpha),n}$;

$X^2 > \hat{F}_{n,1-\alpha}$, то нулевая гипотеза о том, что данная выборка порождена распределением $p(x)$, отклоняется.

Рекуррентное построение теста. Как будет показано ниже, в случае разреженных распределений значение квантили $\hat{F}_{n,1-\alpha}$ может зависеть от объёма выборки n. Строить тест заново для каждой выборки довольно накладно. Поэтому предлагается рекуррентный метод, позволяющий при заданном распределении $p(x)$ вычислить квантили для всех значений n один раз, и затем быстро осуществлять проверку нулевой гипотезы для выборок различного объёма n.

В рекуррентном методе N выборок $\{x_{j1}, \ldots, x_{jn}\}$ наращиваются одновременно, где $j = 1, \ldots, N$ — номер выборки, $n = 1, \ldots, n_{\max}$ — объём выборки. Для каждого j строится эмпирическая гистограмма $H_j(x) = n\hat{p}_j(x)$. При добавлении каждого нового наблюдения $\xi = x_{j,n+1}$, сэмплированного из распределения $p(x)$, обновляется гистограмма и пересчитывается значение статистики $X^2_{j,n+1}$ по значению $X^2_{j,n}$. Сэмплированные выборки не сохраняются. В процессе работы алгоритм формирует двумерный массив значений статистики $X^2_{j,n}$ и одномерный массив эмпирических гистограмм $H_j(x)$. В случае $|\Omega| \gg n_{\max}$ для хранения эмпирических гистограмм лучше использовать специальные структуры данных — разреженные векторы, не выделяющие память под нулевые значения $H_j(x)$. В таком случае расход памяти для данного алгоритма составляет $O(n_{\max}N)$; вычислительная сложность $O(n_{\max}N \log N)$. Детали реализации показаны в Алгоритме 1.

Регрессионный тест

Рассмотрим частную постановку задачи: проверяется нулевая гипотеза о том, что выборка с эмпирическим распределением $\hat{p}(x)$ порождена распределением Ципфа (2) с параметром s. Будем строить распределение статистики X^2 с помощью сэмплирования и исследовать зависимость квантиля $\hat{F}_{n,1-\alpha}$ от параметров n, s и v.

На рис. 2 показана зависимость 0.95-квантиля от объёма выборки n и её интерполяция функцией $\tilde{F}_{1-\alpha}(n) = A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4}$ с параметрами A, B, C, D, E.

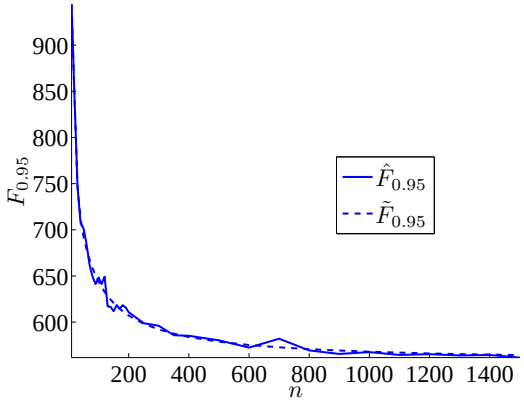

Рис. 2: Зависимость 0.95-квантиля X^2 от объёма выборки n при $s = 1$, $v = 500$, $N = 1000$ и ее интерполяция.

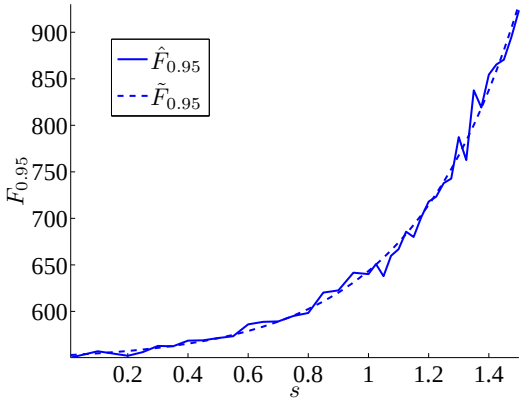

Рис. 3: Зависимость 0.95-квантиля X^2 от параметра s при $n = 100$, $v = 500$, $N = 1000$ и ее интерполяция.

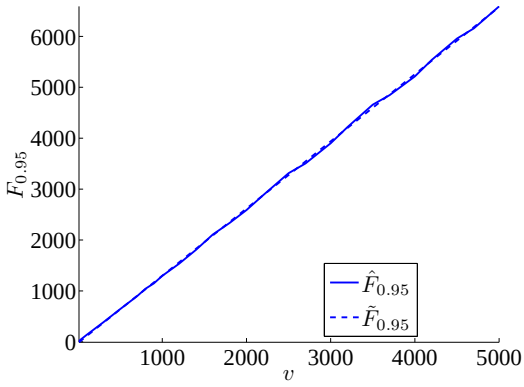

Рис. 4: Зависимость 0.95-квантиля X^2 от $v = |\Omega|$ при $s = 1$, $n = 100$, $N = 1000$ и ее интерполяция.

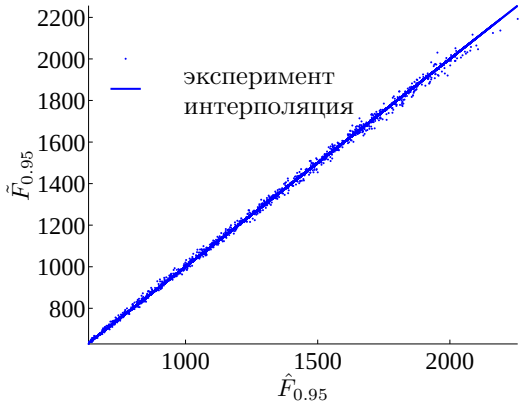

Рис. 5: Зависимость 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{1-\alpha}$, от их эмпирических значений при различных s, n, v.

На рис. 3 показана зависимость 0.95-квантиля от показателя s в законе Ципфа и её интерполяция функцией $\tilde{F}_{1-\alpha}(s) = F + GH^s$ с параметрами F, G, H.

На рис. 4 показана зависимости 0.95-квантиля от параметра $v = |\Omega|$ и её линейная интерполяция $\tilde{F}_{1-\alpha}(v) = I + Jv$ с параметрами I, J.

Построение регрессионного теста. Чтобы найти общий вид зависимости $\tilde{F}_{1-\alpha}(s, v, n)$, применим эмпирический подход. Сформируем обучающую выборку из 1000 троек (s, v, n), равномерно выбранных из параллелепипеда $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки вычислим значение $\hat{F}_{n,0.95}$.

Для поиска нелинейной регрессионной зависимости используем алгоритм символьной регрессии MVR-composer [9, 10]. Преимущество этого алгоритма в том, что он автоматически подбирает формулу регрессии среди всевозможных суперпозиций заданного множества элементарных функций. В нашем случае MVR-composer находит следующую модель регрессии: $\tilde{F}^1_{1-\alpha}(s, v, n) = (A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(F + GH^s)(I + Jv)$ и определяет оптимальные значения 10 параметров $A, B, C, D, E, F, G, H, I, J$. Рассмотрим также

некоторые упрощения этой модели:

$$\tilde{F}^2_{1-\alpha}(s,v,n) = A(1 + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^3_{1-\alpha}(s,v,n) = A(1 + Bn^{-c})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^4_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c})(1 + GH^s);$$
$$\tilde{F}^5_{1-\alpha}(s,v,n) = Av(1 + GH^s);$$
$$\tilde{F}^6_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c}).$$

Параметры этих моделей настроим с помощью функции nlinfit программы Matlab. Начальные приближения всех параметров положим равными 1, кроме параметра A, который инициализируем средним значением $\hat{F}_{n,1-\alpha}/v$ по всей выборке. Получим следующие значения среднеквадратичной ошибки (СКО) на обучающей и контрольной выборках из 1000 случайных троек (s,v,n) каждая:

модель	$\tilde{F}^1_{1-\alpha}$	$\tilde{F}^2_{1-\alpha}$	$\tilde{F}^3_{1-\alpha}$	$\tilde{F}^4_{1-\alpha}$	$\tilde{F}^5_{1-\alpha}$	$\tilde{F}^6_{1-\alpha}$
число параметров	10	8	6	5	3	3
СКО (обучение)	16.3	16.8	16.8	16.7	52.2	43.7
СКО (контроль)	15.8	16.1	16.0	16.0	50.9	43.8

Сравнение СКО на обучающей и контрольной выборках показывает, что переобучения нет ни в одной из моделей. Модель $\tilde{F}^4_{1-\alpha}$ представляется оптимальной по точности и числу параметров. Дальнейшее упрощение модели приводит к резкому увеличению СКО. Оптимальные значения параметров для неё: $A = 0.913$, $B = 3.98$, $c = 0.636$, $G = 0.00458$, $H = 36.8$.

На рис. 4 показан график зависимости 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{0.95}$, от их эмпирических значений при различных s, n, v. Сплошной линией изображена «идеальная» прямая $\tilde{F} = \hat{F}$.

Таким образом, в отличие от классического критерия хи-квадрат, квантиль распределения статистики X^2 существенно зависит от объёма выборки n и от вида распределения $p(x)$, в частности, от показателя степени s в законе Ципфа, отвечающего за разреженность распределения. Построенная регрессионная модель довольно точно описывает зависимость 0.95-квантили от параметров s, n, v. Эту зависимость можно построить один раз вместо того, чтобы строить тест для каждого распределения $p(x)$. Предварительно необходимо убедиться, что распределение $p(x)$ описывается законом Ципфа и найти значение параметра s. Данное обстоятельство сужает область применимости регрессионного теста.

Анализ качества регрессионного теста. Оценим вероятности ошибок первого и второго рода предложенного регрессионного теста в эксперименте.

Ошибкой первого рода называется отклонение нулевой гипотезы при условии её истинности. Вероятность ошибки первого рода равна уровню значимости $\alpha = 0.05$. Для эксперимента сгенерируем контрольную выборку из 500 различных троек (s,v,n), равномерно распределённых на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки сгенерируем 1000 выборок объёма n из распределения Ципфа $p(x)$ с параметрами v и s и вычислим значение статистики X^2. Оценим вероятность ошибки первого рода как долю выборок, для которых нулевая гипотеза отклонялась: $X^2 > \tilde{F}^4_{0.95}(s,v,n)$. Оценка вероятности ошибки первого рода составляет 0.0496 ± 0.0141 с доверительной вероятностью 0.95.

Ошибкой второго рода называется принятие гипотезы $H_0\colon p(x)$ при условии истинности ее альтернативы $H_1\colon p'(x)$. Вероятность ошибки второго рода существенно зависит от альтернативы — чем более похожи распределения $p(x)$ и $p'(x)$, тем больше вероятность ошибки. Исследуем способность теста различать распределения, отличающиеся на

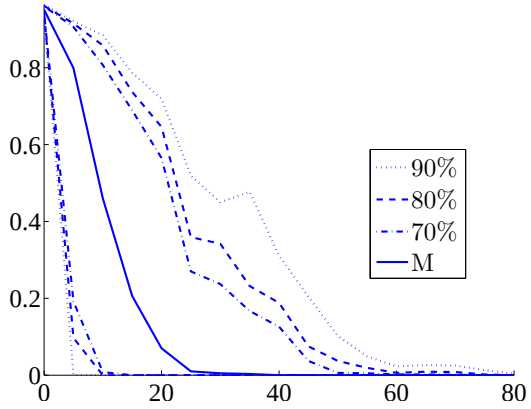

Рис. 6: Зависимость вероятности ошибки второго рода от K при $\mu = 0.01$.

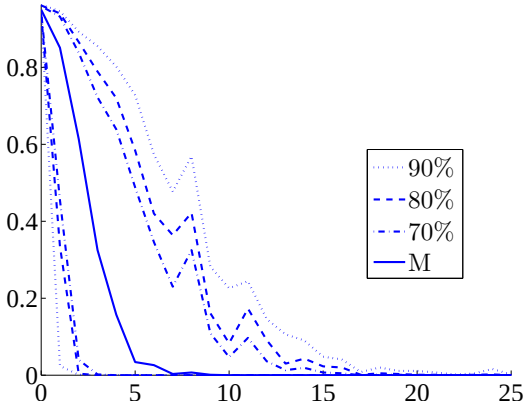

Рис. 7: Зависимость вероятности ошибки второго рода от K при $\mu = 0.05$.

небольшом числе элементов x из Ω. Выделим из множества $\Omega = \{1, \ldots, v\}$ подмножество элементов с наибольшими вероятностями: $\Omega_0 = \{x\colon p(x) > \mu p(1)\}$ при заданном $\mu \in (0, 1)$. Построим распределение $p'(x)$ из $p(x)$ следующим образом: выберем K различных случайных элементов множества Ω_0 и их вероятности поменяем местами с вероятностями K различных случайных элементов множества $\Omega \backslash \Omega_0$.

Из полученного распределения $p'(x)$ сгенерируем выборки, для каждой построим эмпирическое распределение $\hat{p}(x)$ и вычислим статистику X^2. Если $X^2 \leqslant \tilde{F}^4_{0.95}(s, v, n)$, то для данной выборки гипотеза H_0 ошибочно принимается. Долю выборок, при которых это происходит, примем в качестве оценки вероятности ошибки второго рода.

Для каждого K сгенерируем 200 различных троек (s, v, n) из равномерного распределения на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$ и вычислим 200 оценок вероятности ошибки второго рода. На рис. 6 и рис. 7 показаны зависимости медианы M и доверительных границ 90%, 80%, 70% вероятности ошибки второго рода от числа перестановок K при $\mu = 0.01$ и $\mu = 0.05$. По мере увеличения K распределения $p(x)$ и $p'(x)$ все сильнее отличаются, и вероятность ошибки второго рода уменьшается. По мере увеличения μ различия становятся менее контрастными, и вероятность ошибки второго рода убывает медленнее. При $\mu = 0.01$ она становится меньше 0.1 при $K = 20$, при $\mu = 0.05$ она достигает этого значения при $K = 5$.

Отсюда, в частности, можно сделать вывод, что различные тексты, отличающиеся лишь 5 высокочастотными терминами, в среднем довольно надёжно различаются по их случайным фрагментам.

Вероятностные тематические модели

Тематическое моделирование (topic modeling) — одно из активно развивающихся приложений машинного обучения к анализу текстов [5]. *Тематическая модель* коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие термины образуют каждую тему. *Вероятностная тематическая модель* описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Это позволяет решать задачи классификации, кластеризации и категоризации текстов, а также создавать тематические поисковые системы, позволяющие по тексту произвольной длины находить документы схожей тематики.

Исходными данными для тематической модели является множество (коллекция) текстовых документов D и множество (словарь) терминов W. Каждый документ $d \in D$ представляется последовательностью терминов $(w_1, \ldots, w_{n_d})$ из W, где n_d — длина документа. Через n_{dw} обозначается число вхождений термина w в документ d.

Вероятностные модели основаны на следующих предположениях [7, 4].

Во-первых, предполагается, что для выявления тематики достаточно знать, какие термины встречаются в каких документах, но не важен ни порядок терминов в документах (*гипотеза «мешка слов»*), ни порядок документов в коллекции (*гипотеза «мешка документов»*). Другими словами, предполагается, что тематику документа можно узнать даже после случайной перестановки терминов, хотя для человека такой текст теряет смысл.

Во-вторых, предполагается, что существует конечное множество тем T и дискретное распределение $p(d, w, t)$ на $D \times W \times T$, порождающее последовательность независимых наблюдений — троек (d_i, w_i, t_i), $i = 1, \ldots, n$. Переменная t является латентной (скрытой), и наблюдаемая коллекция документов представляет собой последовательность пар (d_i, w_i), $i = 1, \ldots, n$, оставшихся после отбрасывания всех тем.

В-третьих, предполагается, что условное распределение вероятностей терминов $p(w \mid d, t)$ в любом документе d зависит только от темы t, но не от самого документа. Это предположение называется *гипотезой условной независимости*:

$$p(w \mid d, t) = p(w \mid t). \tag{3}$$

Согласно формуле полной вероятности и гипотезе условной независимости,

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d). \tag{4}$$

Построить тематическую модель коллекции — означает по известной левой части $p(w \mid d) = n_{dw}/n_d$ найти неизвестные условные распределения в правой части: $p(w \mid t)$ для каждой темы $t \in T$ и $p(t \mid d)$ для каждого документа $d \in D$, а также определить оптимальное число тем $|T|$.

Большинство тематических моделей [7, 4, 8, 3] оценивают вероятности тем $p(t \mid d, w)$ для каждого слова w в каждом документе d. Зная эти вероятности, возможно оценить число троек:

$n_{dwt} = n_{dw} p(t \mid d, w)$ — в которых термин w документа d связан с темой t,
$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — в которых термин документа d связан с темой t,
$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — в которых термин w связан с темой t,
$n_t = \sum\limits_{d \in D} \sum\limits_{w \in W} n_{dwt}$ — связанных с темой t,

и затем по ним найти частотные оценки искомых условных вероятностей:

$$\hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(w \mid d, t) = \frac{n_{dwt}}{n_{dt}}. \tag{5}$$

Чтобы оценить качество тематической модели, необходимо проверить, выполняется ли гипотеза условной независимости (3) — важнейшее базовое предположение модели (4) — для каждой пары документ–тема (d, t). Тема t описывается распределением $\hat{p}(w \mid t)$. Выборка слов документа d, относящихся к теме t, согласно модели, образует эмпирическое распределение $\hat{p}(w \mid d, t)$. Оба распределения оцениваются согласно (5) в процессе построения тематической модели. Чтобы проверить, действительно ли данная выборка могла быть получена из распределения $\hat{p}(w \mid t)$, воспользуемся критерием согласия, основанным на статистике хи-квадрат (1):

$$X^2_{dt} = n_{dt} \sum_{w:\, n_{wt} > 0} \frac{\big(\hat{p}(w \mid d, t) - \hat{p}(w \mid t)\big)^2}{\hat{p}(w \mid t)}. \tag{6}$$

Число различных слов в теме может быть намного больше, чем число слов в документе. Следовательно, мы имеем дело с разреженными распределениями, к которым неприменим асимптотический критерий хи-квадрат. Поэтому будем строить статистические тесты методом сэмплирования, для каждой темы $t \in T$ отдельно.

Экспериментально установлено, что для больших корпусов текстов на естественных языках закон Ципфа или более сложные параметрические законы (например Ципфа–Мандельброта) выполняются с неплохой точностью [1, 6]. Для ускорения проверки гипотезы условной независимости предлагается двухэтапный тест. Сначала проверяется согласие каждой темы t с выбранным параметрическим законом. Если согласие есть, то строится один регрессионный тест для всех таких тем. Для каждой из остальных тем строится отдельный тест на основе сэмплирования.

Сэмплирование без возвращений

Проверки согласия документных эмпирических распределений $\hat{p}(w \mid d, t)$, $d \in D$ с распределением $\hat{p}(w \mid t)$, вообще говоря, не являются независимыми, поскольку имеется тождество, связывающие эти распределения друг с другом:

$$\hat{p}(w \mid t) = \sum_{d \in D} \hat{p}(w \mid d, t)\hat{p}(d \mid t). \tag{7}$$

Документы являются выборками без возвращений из распределения $\hat{p}(w \mid t)$, тогда как обычно критерии согласия предполагают выборку с возвращениями. Наличие дополнительного ограничения (7) может и не влиять на результаты тестов или влиять несущественно, особенно на коллекциях большого размера. Однако это лишь предположение, которое необходимо проверить. Для этого построим более точный тест на основе сэмплирования *без возвращений*, учитывающий, что последовательность слов, образующих тему t, разрезается на документы в пропорциях $\hat{p}(d \mid t)$.

Построение теста сэмплированием без возвращений. Возьмём последовательность терминов длины n_t, образующую распределение $\hat{p}(w \mid t)$. Сгенерируем N случайных перестановок этой последовательности. Разрежем каждую из полученных последовательностей W_j, $j = 1, \ldots, N$ на «документы» — подпоследовательности терминов W_{jd} длины n_{dt} каждая, $d \in D$. По каждому «документу» W_{jd} построим эмпирическое распределение $\hat{p}_j(w \mid d, t)$ и вычислим значение статистики хи-квадрат X^2_{jd}. Для каждого $d \in D$ по множеству значений статистики $X^2_{1d}, \ldots, X^2_{Nd}$ построим эмпирическую функцию распределения $\hat{F}_d(X^2)$ и вычислим её $(1-\alpha)$-квантиль $\hat{F}_{d,1-\alpha}$. Число N рекомендуется брать не менее 1000 при типичном значении $\alpha = 0.05$.

Отметим, что в тесте без возвращений квантиль строится для каждого документа d, тогда как тест с возвращениями строился для каждого значения длины документа n. Построение теста без возвращений более ресурсоёмко и требует $O(n_t N \log N)$ операций вместо $O(n_{\max} N \log N)$, где $n_{\max} = \max\limits_{d \in D} n_{td}$.

Применение теста сэмплированием без возвращений. Проверка гипотезы условной независимости для пары документ–тема (d, t) заключается в вычислении статистики X^2_{dt} по формуле (6) и проверке неравенства $X^2_{dt} > \hat{F}_{d,1-\alpha}$. Если оно выполнено, то гипотеза условной независимости отвергается для данной пары (d, t).

Вычислительные эксперименты

Эксперименты проводились на коллекции из $|D| = 2000$ авторефератов диссертаций на русском языке. Мощность словаря после предварительной обработки данных (лемматизации и удаления стоп-слов) составляет $|W| = 20211$ слов, длина документов от 1000 до 4000 слов. Строились две тематические модели — PLSA [7] и LDA-GS [4, 8] с помощью алгоритма, описанного в [2]. Число тем $|T| = 100$.

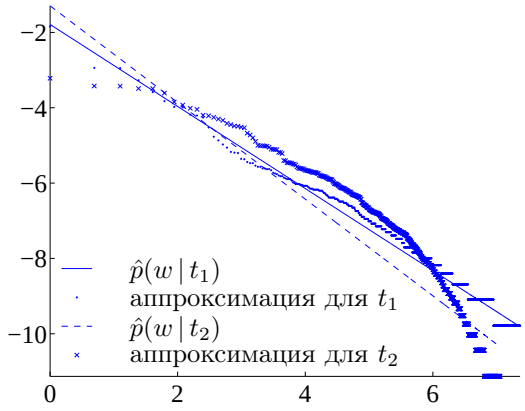

Рис. 8: Аппроксимация эмпирических распределений слов законом Ципфа (для двух тем, в логарифмических осях).

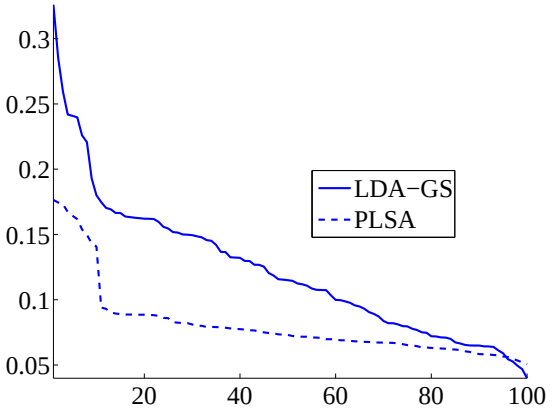

Рис. 9: Доля документов, для которых гипотеза условной независимости отклоняется (в порядке убывания).

Выполняется ли закон Ципфа для тем? На рис. 8 показаны графики эмпирических распределений и закона Ципфа для двух из 100 тем t_1 и t_2 в модели LDA, в логарифмических осях. По горизонтальной оси откладывается логарифм номера слова, слова упорядочены по частоте. По вертикальной оси откладывается логарифм вероятности слова. Оптимальные значения параметра закона Ципфа: $s = 1.04$ для t_1, $s = 1.28$ для t_2. Хотя «на глаз» соответствие неплохое, особенно для t_1, нулевая гипотеза отклоняется для обоих тем. Более того, большинство тем согласуются с законом Ципфа лишь при крайне низких уровнях значимости, меньших 0.05. Это объясняется тем, что при выборках длины n_t порядка 10^3–10^5 критерии согласия чувствительны даже к незначительным различиям распределений, и одного параметра в законе Ципфа не достаточно для описания эмпирических распределений.

Сравнение тестов без возвращений и с возвращениями.

Для модели PLSA рассматривается одна тема из $|D_t| = 1992$ документов суммарной длины $n_t = 87026$ слов. В тестах без возвращений и с возвращениями нулевая гипотеза принимается для 1674 и 1688 документов соответственно. Решения отличаются на 22 документах из 1992. Оба теста дают примерно одинаковый результат: гипотеза условной независимости отклоняется для 15% документов.

Для модели LDA-GS рассматривается тема из $|D_t| = 1114$ документов суммарной длины $n_t = 63805$ слов. Нулевая гипотеза принимается для 1032 и 1035 документов соответственно. Решения отличаются на 7 документах из 1114. Оба теста снова дают примерно одинаковый результат: нулевая гипотеза отклоняется для 7% документов.

Таким образом, результаты тестов без возвращений и с возвращениями почти одинаковы, однако тест с возвращениями менее ресурсоёмкий.

На рис. 9 показан результат сравнения моделей PLSA и LDA по всем темам. По вертикальной оси откладывается доля документов, для которых отклоняется гипотеза условной независимости. По горизонтальной оси откладываются темы в порядке убывания долей (порядки тем для двух моделей, естественно, не совпадают). Модель LDA строит темы менее аккуратно, что, возможно, объясняется применением смещённых (сглаженных) частотных оценок условных вероятностей в LDA, в то время как PLSA основан на несмещённых оценках максимального правдоподобия. Однако в обеих моделях доля документов, не прошедших тест, превышает уровень значимости 0.05 для всех тем. Это может быть объяснено выбором неоптимального (заниженного) числа тем $|T| = 100$. Более полный статистический анализ качества тематических моделей PLSA и LDA выходит за рамки данной работы.

Выводы

Предложены критерии согласия на основе сэмплирования для разреженных дискретных распределений, выходящих за границы применимости классических асимптотических критериев. Предложен рекуррентный алгоритм построения теста на основе сэмплирования. Для параметрического случая, когда проверяется согласие эмпирических данных с распределением Ципфа, построен регрессионный тест, подобрана модель регрессии и проведён анализ ошибок первого и второго рода. Рассмотрено применение предложенных тестов для проверки гипотезы условной независимости — ключевого предположения вероятностных тематических моделей коллекций текстовых документов. Экспериментально показано, что в случае большого числа документов нет необходимости строить точный тест без возвращений, и можно пользоваться вычислительно более эффективным тестом с возвращениями.

Работа выполнена при поддержке Министерства образования и науки Российской Федерации (Государственный контракт 07.524.11.4002).

Литература

[1] *Бриллюэн Л.* Наука и теория информации. — М.: «Государственное издательство физико-математической литературы», 1960. — 391 с.

[2] *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование.* — 2012. — Т. 4, № 4. — С. 693–706.

[3] *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — 2009.

[4] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 993–1022.

[5] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // *Frontiers of Computer Science in China.* — 2010. — Vol. 4, no. 2. — Pp. 280–301.

[6] *Gelbukh A., Sidorov G.* Zipf and heaps laws' coefficients depend on language // Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science. — Springer-Verlag, 2001. — P. 332–335.

[7] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.

[8] *Steyvers M., Griffiths T.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.

[9] *Strijov V.* Search for a parametric regression model in an inductive-generated set // *Computational technologies.* — 2007. — Vol. 12, no. 1. — Pp. 93–102.

[10] *Strijov V.* MVR Composer. — 2012. http://strijov.com/?p=84.

[11] *von Davier M.* Bootstrapping goodness-of-fit statistics for sparse categorical data — results of a monte carlo study // *Methods of Psychological Research Online.* — 1997. — Vol. 2, no. 2.

[12] *Zelterman D.* Goodness-of-fit tests for large sparse multinomial distributions // *Journal of the American Statistical Association.* — 1987. — Vol. 398, no. 82. — Pp. 624–629.