Мотренко А.П., Стрижов В.В. Построение агрегированных прогнозов объемов железнодорожных грузоперевозок с использованием расстояния Кульбака-Лейблера // Информатика и ее применения, 2014, 8(2) : 86-97.

Построение агрегированных прогнозов объемов железнодорожных грузоперевозок с использованием расстояния Кульбака–Лейблера*

А. П. Мотренко[1], *В. В. Стрижов*[2]

Данное исследование посвящено проблеме построения агрегированных прогнозов объемов железнодорожных грузоперевозок. Для получения агрегированных прогнозов требуется кластеризовать временные ряды таким образом, чтобы распределения временных рядов внутри кластера совпадали. При решении задачи кластеризации требуется оценить близость между временными рядами, исходя из их эмпирических распределений. Вводится критерий принадлежности временных рядов одному распределению, основанный на расстоянии Кульбака–Лейблера между гистограммами временных рядов. Приводится теоретическое и практическое исследование предложенного критерия. Решается задача кластеризации временных рядов на основе матрицы парных расстояний между ними.

Ключевые слова: *эмпирическая функция распределения; расстояние между гистограммами; расстояние Кульбака–Лейблера; задача двух выборок*

1 Введение

Особенностью задачи прогнозирования объема погрузок по историческим данным о загруженности железнодорожной сети различными группами грузов является необходимость определить оптимальный уровень детализации [1, 2]: по видам перевозимых грузов, по наборам станций, по кодам вагонов. Требуется получить прогноз как для объема погрузок в целом, так и для отдельных групп грузов. При этом спрогнозированный объем погрузок в целом может не совпадать с суммой прогнозов по отдельным группам. Для повышения согласованности полученных прогнозов предлагается [3] вместо прогноза «в целом» объединять ряды только в том случае, если их распределения совпадают, чтобы агрегированные данные имели тот же статистический смысл, что и исходные ряды. Для решения задачи агрегации временных рядов необходимо определить расстояние между временными рядами таким образом, чтобы оно отражало близость эмпирических распределений между рядами.

В литературе по математической статистике вводится множество коэффициентов, показывающих, что некоторые два распределения P и Q близки друг к другу. Такие коэффициенты в различных источниках называются расстоянием между распределениями [4], мерами разделяющей информации [5], мерами статистического расстояния [6]. В работе [7] описан метод порождения коэффициентов $d(P, Q)$ «непохожести» двух распределений, обладающих некоторыми стандартными свойствами, например:

1) коэффициент $d(P, Q)$ должен быть определен на всех парах распределений с одним носителем;
2) значение $d(P, Q)$ должно быть минимально при $P = Q$;
3) при любом измеримом преобразовании носителя распределений P и Q расстояние между ними не увеличивается.

* Работа выполнена при поддержке РФФИ (грант 13-07-13139).

1 Московский физико-технический институт, anastasia.motrenko@gmai.com

2 Вычислительный центр Российской академии наук им. А. А. Дородницына, strijov@ccas.com

Идея метода [7] заключается в том, чтобы рассмотреть различные выпуклые функции случайной величины Q/P. С точки зрения распределения P матожидание Q/P независимо от Q, а дисперсия стремится к нулю при $Q \to P$. Также в [7] показано, что многие известные функции расстояния могут быть получены этим методом. В частности, им могут быть порождены все f-дивергенции [8] и в том числе расстояние Кульбака–Лейблера [4]. В работе [9] приведено сравнение многих известных расстояний с точки зрения скорости сходимости эмпирического распределения к истинному, а также качественного поведения функции расстояния при сходимости. При решении задачи кластеризации в обработке изображений были введены меры [10, 11], основанные на метрике Вассерштейна.

В работе [1] для оценки близости распределений используется расстояние Кульбака–Лейблера между гистограммами, построенными по временным рядам. В данной работе показано, что распределение расстояние Кульбака–Лейблера между гистограммами из одного распределения в пределе ограничено сверху распределением χ^2.

Предложен критерий для решения задачи двух выборок, основанный на расстоянии Кульбака–Лейблера между гистограммами временных рядов. Продемонстрировано применение критерия к решению задачи двух выборок для различных пар распределений и показана его состоятельность. Для набора временных рядов о железнодорожных грузоперевозках решается задача кластеризации с помощью алгоритма кратчайшего незамкнутого пути [12] на основе матрицы парных расстояний [13, 14] между рядами. При решении задачи кластеризации ряды группируются по типу груза.

2 Постановка задачи

Задан набор временных рядов $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_S\}$, где каждый ряд $\mathbf{x}_j = \{x_j(i) \in \mathbb{R}\}_{i=1}^{m_j}$ — это последовательность реализаций некоторого стационарного случайного процесса. Требуется кластеризовать набор

$$\mathbf{X} = \bigsqcup_{k=1}^{K} \mathbf{X}_k, \quad \mathbf{X}_k = \{\mathbf{x}_j,\, j \in \mathcal{A}_k\}, \quad \{1, \ldots S\} = \bigsqcup_{k=1}^{K} \mathcal{A}_k, \tag{1}$$

разбив набор $\mathbf{X}$ на K наборов $\mathbf{X}_k$ временных рядов, таких что все ряды в $\mathbf{X}_k$ принадлежат одному и тому же распределению. Здесь $\mathcal{A}_k$ — множество индексов временных рядов k-го кластера.

В силу стационарности случайного процесса $\mathbf{x}$ пренебрежем последовательностью значений ряда $\mathbf{x}$. Представим временной ряд $\mathbf{x}$ как выборку X реализаций некоторой случайной величины с распределением P:

$$X = \{x \in \mathbb{R} | \text{ для некоторого } i \in \{1, \ldots, m\} : x(i) = x\}. \tag{2}$$

Для решения задачи об агрегировании временных рядов $\mathbf{x}$ и $\mathbf{x}'$ будем сравнивать гистограммы, построенные по выборкам X и X', сопоставленным каждому из рядов в соответствии с (2). Опишем подробнее процедуру построения гистограммы.

Пусть объемы выборок X и X' равны m и m' соответственно. Разобьем область значений случайной величины из P на N промежутков $[a_i, a_{i+1}]$ и обозначим $p_i = P(a_i < x \leqslant a_{i+1})$ вероятность случайной величине с распределением P принять значение из i-го промежутка; n_i и n'_i — количество объектов выборок X и X', попавших в i-й промежуток. Обозначим $\hat{P}_m$ гистограмму, построенную по выборке X объема m из распределения P. Гистограмма $\hat{P}_m$ задается набором оценок

$$\hat{P}_m(a_i < x \leqslant a_{i+1}) = \frac{n_i}{m} = \hat{p}_i, \quad i = 1, \ldots, N-1, \tag{3}$$

вероятности p_i.

Для решения задачи кластеризации (1) воспользуемся алгоритмом нахождения кратчайшего незамкнутого пути между временными рядами. Результатом применения алгоритма является минимальное остовное дерево: граф с $n-1$ ребрами, покрывающий все n вершин, ребра которого обладают минимальной суммарной длиной. Удалив из минимального остовного дерева $K-1$ самых длинных ребер, получим кластеризацию вершин графа на K кластеров $\mathbf{X}_k$. Вершинами графа являются исследуемые временные ряды $\mathbf{x}_j$; длина ребра, соединяющего две вершины, равна расстоянию между соответствующими временными рядами. Найдя расстояние между всеми парами рядов, получим матрицу парных расстояний D. В качестве расстояний между рядами $\mathbf{x}_r$ и $\mathbf{x}_s$ будем использовать симметризованное расстояние Кульбака–Лейблера между гистограммами $\hat{P}_{m_r}$ и $\hat{P}_{m_s}$, построенными по временным рядам:

$$D(r,s) = \frac{2m_r m_s}{m_r + m_s}\left(D_{\mathrm{KL}}(\hat{P}_{m_r}||\hat{P}_{m_s}) + D_{\mathrm{KL}}(\hat{P}_{m_s}||\hat{P}_{m_r})\right). \tag{4}$$

Первый сомножитель в правой части снимает зависимость от объема выборки. Необходимость его введения будет объяснена в следующем разделе.

Конечной целью кластеризации временных рядов с учетом расстояний между ними является повышение согласованности агрегированных прогнозов. Для оценки качества кластеризации будем рассматривать несогласованность

$$\delta(i) = \left|\sum_k^K \hat{\mathbf{X}}_k(i) - \sum_j^n \hat{\mathbf{x}}_j(i)\right| \tag{5}$$

при прогнозировании по наборам $\mathbf{X}_k$ временных рядов и отдельным временным рядам $\mathbf{x}_j$. Здесь $\hat{\mathbf{X}}_k(i)$ — прогноз агрегированного ряда в момент i, $\hat{x}_j(i)$ — прогноз j-го ряда в момент времени i. Чем меньше несогласованность, тем качественнее выполнена кластеризация. Очевидно, что наименьшее значение $\delta_{\min} = 0$ выражения (5) достигается при $K = S$, поэтому предлагается ограничить число K или ввести в (5) штраф $h(K)$ за его повышение:

$$K = \arg\min_K \left(\sum_{i=1}^m \delta(i) + h(K)\right).$$

В данной работе ограничимся рассмотрением предложенного критерия принадлежности временных рядов к одному распределению и кластеризации временных рядов на основе расстояния между ними.

3 Статистическая значимость расстояния Кульбака–Лейблера

Чтобы показать, что результаты кластеризации временных рядов на основе расстояния Кульбака–Лейблера между ними статистически значимы, необходимо исследовать распределение расстояния Кульбака–Лейблера между гистограммами, построенными по выборкам X и X' из одного распределения P. В данном разделе будет показано, что, хотя расстояние Кульбака–Лейблера не имеет предельного распределения, для него можно получить предельные оценки сверху.

Пусть пока выборки X и X' имеют одинаковый объем m. Рассмотрим расстояние $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$ между гистограммами $\hat{P}_m$ и $\hat{P}'_m$. По определению расстояние Кульбака–Лейблера $D_{\mathrm{KL}}(Q||P)$ между распределениями Q и P равно

$$D_{\mathrm{KL}}(Q||P) = \int P \cdot f\left(\frac{Q}{P}\right), \tag{6}$$

где $f(t) = t \ln t$. Функция f строго выпукла и дважды дифференцируема в единице, и, повторяя рассуждения из [8], разложим подынтегральное выражение из правой части (6) по f в окрестности единицы:

$$P(x)f\left(\frac{Q(x)}{P(x)}\right) = f(1)+f'(1)(Q(x)-P(x))+\frac{f''(1)}{2}\frac{(Q(x)-P(x))^2}{P(x)}+P(x)o\left(\left(\frac{Q(x)}{P(x)}-1\right)^3\right),$$

где $f(1) = 0,\ \ f''(1) = 1$. Подставив вместо Q распределение $\hat{P}_m$, определяемое (3), и просуммировав по i, получим соотношение

$$D_{\text{KL}}(\hat{P}_m||P) = \sum_{i=1}^{N} p_i f\left(\frac{\hat{p}_i}{p_i}\right) =$$

$$= \frac{1}{2}\sum_{i}^{N}\frac{(\hat{p}_i - p_i)^2}{p_i} + \sum_{i=1}^{N} p_i \cdot \varepsilon\left(\left(\frac{\hat{p}_i}{p_i}-1\right)^3\right) \sim \frac{1}{2m}\sum_{i}^{N}\frac{(n_i - mp_i)^2}{mp_i}$$

и следующий предельный переход:

$$2m \cdot D_{\text{KL}}(\hat{P}_m||Q) \sim m\sum_{i=1}^{N}\frac{(\hat{p}_i - p_i)^2}{p_i)} = \sum_{i=1}^{N}\frac{(n_i - mp_i)^2}{mp_i} \to \chi^2_N \ \text{ при } m \to \infty. \tag{7}$$

Докажем следующую теорему:

Теорема 1. *Случайная величина $2m \cdot D_{KL}(Q||\hat{P}_m) \to \chi^2_N$ по распределению при $m \to \infty$.*

Доказательство. Аналогично доказательству предельного перехода (7) разложим $D_{\text{KL}}(Q||\hat{P}_m)$ по степеням $f(t)$ вблизи единицы и получим

$$D_{\text{KL}}(Q||\hat{P}_m) \sim \frac{1}{2}\sum_{i=1}^{N}\frac{(\hat{P}_m(\xi_i) - Q(\xi_i))^2}{\hat{P}_m(\xi)} = \frac{1}{2m}\sum_{i=1}^{N}\frac{(n_i - mp_i)^2}{n_i}.$$

Пусть $G_m(x)$ — функция распределения случайной величины $\sum_{i=1}^{N}\frac{(n_i-mp_i)^2}{mp_i}$, $F_m(x)$ — случайной величины $\sum_{i=1}^{N}\frac{(n_i-mp_i)^2}{n_i}$. Так как $G_m(x)$ сходится поточечно к $F_{\chi^2_{N-1}}$ при $m \to \infty$, имеем

$$|G_m(x) - F_{\chi^2_{N-1}}| < \frac{\varepsilon}{2} \quad \forall m > m'.$$

Докажем, что $|G_m(x) - F_m(x)| \to 0$ при $m \to \infty$. Для этого покажем, что $\forall \varepsilon > 0$ найдется объем выборки m_0 такой, что для всех $m > m_0$ выполняется

$$\text{P}\left(\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \leqslant \frac{\varepsilon}{N}\right) > 1 - \varepsilon. \tag{8}$$

Согласно центральной предельной теореме

$$\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}} \to \mathcal{N}(0,1) \ \text{ по распределению при } m \to \infty,$$

причем для скорости сходимости имеет место неравенство Берри–Эссеена:

$$|Q_m(x) - \Phi(x)| \leqslant \frac{A}{\sqrt{m}},$$

где $Q_m(x)$ — функция распределения величины $\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}}$, $\Phi(x)$ — функция стандартного нормального распределения, A — некоторая константа. Тогда вероятность

$$\mathrm{P}\left(\left|\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}}\right| < C\right) = Q_m(C) - Q_m(-C) \geqslant 2\Phi(C) - 1 - \frac{2A}{\sqrt{m}}. \tag{9}$$

Пусть, кроме того, выполняется $0 < 1 - p \leqslant p_i \leqslant p < 1$. Тогда с вероятностью $\mathrm{P}_C \geqslant 2\Phi(C) - 1$ выполняется

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| = \frac{|n_i - mp_i|^3}{mn_ip_i} \leqslant \frac{C^3(1-p_i)^3p_i^2}{n_i}\sqrt{m} \leqslant \frac{C^3(1-p)^3p}{\sqrt{m} - C(1-p)}.$$

Обозначим $m_1 = [4C^2(1-p)^2]$, тогда при $m > m_1$ имеет место $\sqrt{m} - C(1-p_i) > \frac{1}{2}\sqrt{m}$ и

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \leqslant \frac{2C^3(1-p_i)^3p_i}{\sqrt{m}}.$$

Тогда для фиксированного ε определим

$$C_\varepsilon = \frac{\varepsilon^{1/3}m^{1/6}}{(1-p_i)(2p_iN)^{1/3}},\ \mathrm{P}_m(\varepsilon) = 2\Phi(C_\varepsilon) - 1 - \frac{2A}{\sqrt{m}}.$$

При заданном ε вероятность $\mathrm{P}_m(\varepsilon) \to 1$ при $m \to \infty$, поэтому найдется m_2 такое, что для любого $m > m_2$ выполнено $\mathrm{P}_m(\varepsilon) > 1 - \varepsilon$. Выбрав $m_0 = \max(m_1, m_2)$, получим утверждение (8). Тогда

$$\left|\sum_{i=1}^{N}\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| \leqslant \sum_{i=1}^{N}\left|\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| < \varepsilon \quad \text{при } m > m_0.$$

Из только что доказанного следует, что $|F_m(x) - G_m(x)| \to 0$ при $m \to 0$. Тогда $\forall\varepsilon > 0\ \exists m'' :$ при $m > m''$ выполняется

$$|F_m(x) - F_{\chi^2_{N-1}}| < |F_m(x) - G_m(x)| + |G_m(x) - F_{\chi^2_{N-1}}| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}.$$

■

Доказанная теорема и утверждение (7) позволяют получить оценки распределения случайных величин $2m \cdot D_{\mathrm{KL}}(\hat{P}_m||Q)$ и $2m \cdot D_{\mathrm{KL}}(Q||\hat{P}_m)$ при больших m. Для решения задачи кластеризации (1) потребуется также исследовать поведение расстояния Кульбака–Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l)$ между гистограммами, построенными по выборкам X, X' различных длин m и l. Воспользовавшись неравенством треугольника

$$D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leqslant D_{\mathrm{KL}}(\hat{P}_m||Q) + D_{\mathrm{KL}}(Q||\hat{P}_l),$$

получим следствия из теоремы 1.

Следствие 1. $2m \cdot D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m) \leqslant \chi^2_{2N}$ в пределе при $m \to \infty$.

Следствие 2. Пусть выборки X, X' растут таким образом, что $m/l \to \rho$, $0 < \rho < \infty$. Тогда

$$2\frac{ml}{m+l} \cdot D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leqslant \chi^2_{2N}$$

в пределе при $m, l \to \infty$.

Доказательство. Действительно, при выполнении условия $m/l \to \rho$, $0 < \rho < \infty$, имеем

$$\frac{l}{m+l} \to \frac{1}{1+\rho}, \quad \frac{m}{m+l} \to \frac{\rho}{1+\rho}$$

и

$$2\frac{ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leqslant \frac{l}{m+l} 2mD_{\mathrm{KL}}(\hat{P}_m||Q) + \frac{m}{m+l} 2lD_{\mathrm{KL}}(Q||\hat{P}_l) \to \chi^2_{2N}.$$

■

Обозначим величину $\frac{2ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$ через $\xi_{m,l}$. Следствие 2 дает верхнюю оценку поведения случайной величины $\xi_{m,l}$ при больших m и l, а именно: пусть $\eta \sim \chi^2_{2N}$, тогда при достаточно больших m и l для любого элементарного исхода w из вероятностного пространства Ω выполнено $\xi_{m,l}(w) < \eta(w)$. Следовательно, для любого $x \in \mathbb{R}$ верно

$$\mathrm{P}(\xi_{m,l} < x) \geqslant \mathrm{P}(\eta < x). \tag{10}$$

В следующем разделе покажем, как этот факт будет использоваться для проверки принадлежности временных рядов к одному распределению.

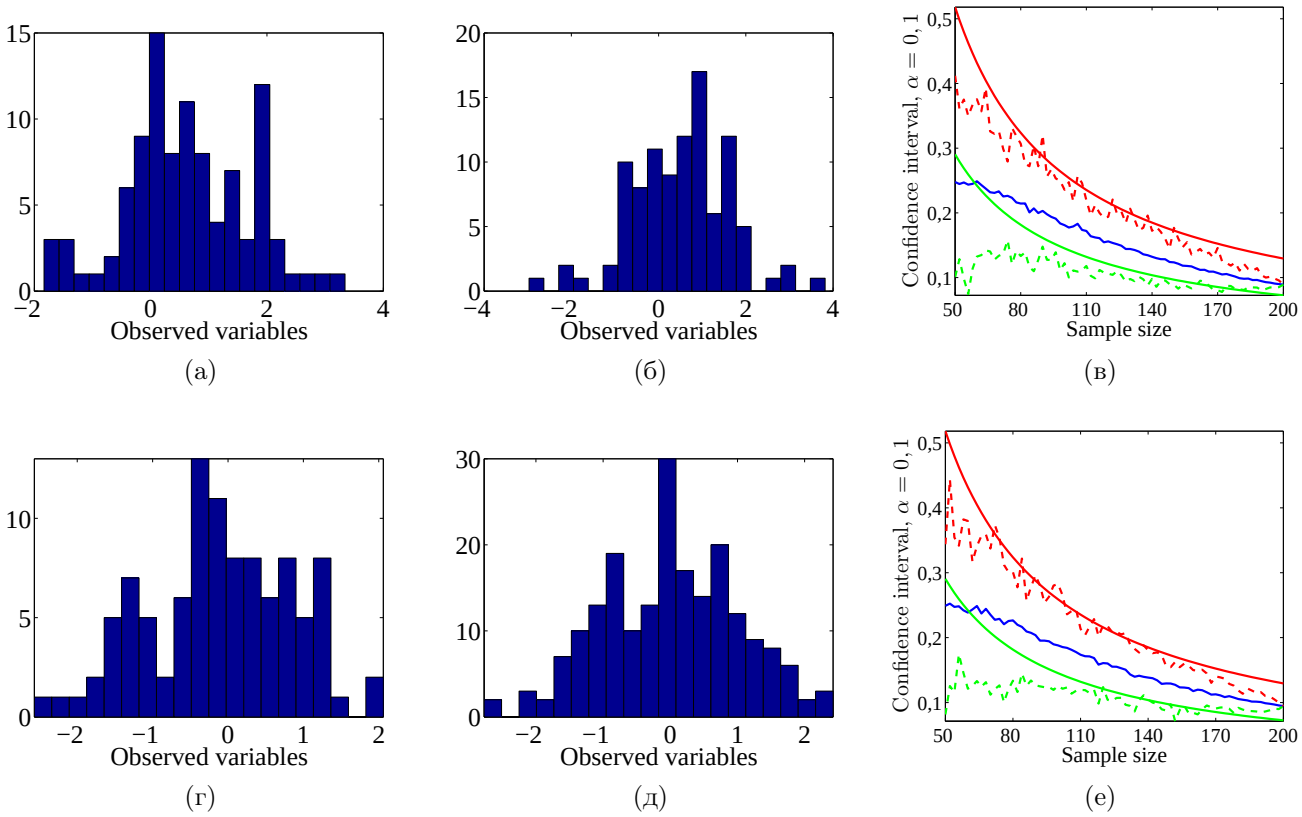


Рис. 1: Гистограммы, построенные по двум выборкам из нормального распределения (а, б), и зависимость статистики t_m от объема выборки (в, е). Красным и зеленым отмечены границы доверительного интервала для t_m при $\alpha = 0,1$. Выборки (г, д) были зашумлены

4 Проверка принадлежности временных рядов к одному распределению

Для решения задачи агрегирования временных рядов $\mathbf{x}$ и $\mathbf{x}'$ необходимо уметь принимать решение о принадлежности временных рядов к одному распределению. Опишем

процедуру проверки гипотезы о принадлежности выборок X и X', составленных (2) из временных рядов $\mathbf{x}$ и $\mathbf{x}'$. Пусть нулевая гипотеза H_0 состоит в принадлежности выборок X и X' к одному распределению:

$$H_0 : P(x) = P'(x).$$

Сформулируем критерий проверки гипотезы H_0 при альтернативе $H_1 : P(x) \neq P'(x)$. Для этого определим критическую область $U(\alpha)$ для статистики $t_{m,l}$ с уровнем значимости α:

$$U(\alpha) = \{t : \bar{t}_{1-\alpha} > t \text{ или } t > \bar{t}_\alpha\},$$

где критическое значение $\bar{t}_\alpha$ определяется соотношением

$$\mathrm{P}(t > \bar{t}_\alpha | H_0) = \alpha. \tag{11}$$

Так как предельное распределение величины $\xi_{m,l}$ неизвестно, будем использовать критическую область, задаваемую распределением χ^2_{2N}. Будем говорить, что данные отвергают гипотезу H_0 в случае, если статистика $t_{m,l}$ принадлежит критической области

$$U^{\chi^2}(\alpha) = \{t : \bar{t}^{\chi^2}_{1-\alpha} > t \text{ или } t > \bar{t}^{\chi^2}_\alpha\}, \tag{12}$$

где $\bar{t}^{\chi^2}$ — критическое значение величины χ^2_{2N}:

$$\mathrm{P}(t > \bar{t}^{\chi^2}_\alpha | t \sim \chi^2_{2N}) = \alpha.$$

Из неравенства (10) и определения (11) критических значений следует, что критические области U и U^{χ^2} не сравнимы, то есть

$$\bar{t}_{1-\alpha} < \bar{t}^{\chi^2}_{1-\alpha}, \quad \bar{t}_\alpha < \bar{t}^{\chi^2}_\alpha.$$

Это означает, что возможны следующие ситуации:

1. Случай $\bar{t}^{\chi^2}_{1-\alpha} < t_{m,l} < \bar{t}_\alpha$, когда статистика $t_{m,l}$ одновременно принадлежит истинной, но неизвестной критической области U и вычислимой критической области U^{χ^2}.
2. Случай $\bar{t}_{1-\alpha} < t_{m,l} < \bar{t}^{\chi^2}_{1-\alpha}$, когда статистика $t_{m,l}$ принадлежит истинной, но неизвестной критической области U и не принадлежит U^{χ^2}. Так как $t_{m,l} \in U$, то с высокой вероятностью гипотеза H_0 неверна, и есть риск принять неверное решение об истинности гипотезы H_0. То есть зазор между $\bar{t}_{1-\alpha}$ и $\bar{t}^{\chi^2}_{1-\alpha}$ повышает вероятность ошибки второго рода.
3. Случай $\bar{t}_\alpha < t_{m,l} < \bar{t}^{\chi^2}_\alpha$, когда статистика $t_{m,l}$ попадает в U^{χ^2}, хотя на самом деле $t_{m,l}$ не принадлежит U. В этом случае велика вероятность, что H_0 верна, но решение будет принято в пользу H_1. Таким образом, зазор между $\bar{t}_\alpha$ и $\bar{t}^{\chi^2}_\alpha$ повышает вероятность ошибки первого рода.

Второй случай разрешается следующим образом: использование симметризованного расстояния позволяет перейти от двусторонних критериев U и U^{χ^2} вида (12) к односторонним критериям

$$U_1(\alpha) = \{t : t > \bar{t}_\alpha\}, \quad U^{\chi^2}_1(\alpha) = \{t : t > \bar{t}^{\chi^2}_\alpha\}.$$

В этом случае $U^{\chi^2}_1 \subseteq U_1$ и справедливо следствие $t_{m,l} \in U^{\chi^2}_1 \Rightarrow t_{m,l} \in U_1$. Кроме того, далее будет показано (теорема 2), что при увеличении объема выборки m вероятность

отклонить гипотезу H_0 с помощью критерия (12) в случае, если гипотеза H_0 неверна, стремится к единице. Влияние третьего случая на возможность применения критерия (12) для принятия нулевой гипотезы исследуется экспериментально. Эксперименты, приведенные ниже и в разделе 5, показывают, что при истинности нулевой гипотезы области U и U^{χ^2} достаточно близки для принятия верного решения.

Пример применения критерия (12) **при истинности** H_0**.** На рис. 1 изображены гистограммы для двух выборок из стандартного нормального распределения (рис. 1, а, б) и стандартного нормального распределения с шумом $\varepsilon \sim 0.1 \cdot R[0,1]$ (рис. 1, г, д), а также зависимость расстояния Кульбака–Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$ между дискретными распределениями, задаваемыми гистограммами $\hat{P}_m$ и $\hat{P}'_m$, между выборками одинакового объема m от объема выборки m и область допустимых значений с точки зрения критерия (12) (рис. 1, в, е). Здесь вместо критических значений $\bar{t}^{\chi^2}_{1-\alpha}$, $\bar{t}^{\chi^2}_{\alpha}$ и $t_m = 2m \cdot D_{\mathrm{KL}}$ отложены величины $\bar{t}^{\chi^2}_{1-\alpha}/2m$, $\bar{t}^{\chi^2}_{\alpha}/2m$ и $t_m/2m$, чтобы продемонстрировать масштаб расстояния Кульбака–Лейблера и наличие сходимости. Рисунки показывают, что в данном случае использование распределения χ^2_{2N} в качестве оценки предельного распределения статистики t_m позволяет принять верное решение о принадлежности рядов к одному распределению. Пунктирная линия показывает границу области, в которую вошло $1-\alpha = 90\%$ выборки, и задает оценку критической области для $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$. Более подробно результаты описаны в разделе 5.

Покажем теперь, что критерий (12) также можно использовать для отвержения гипотезы H_0.

Теорема 2. *Критерий* (12) *состоятелен:*

$$\lim_{m\to\infty} \mathrm{P}(t_m \in U|H_1) = 1,$$

то есть вероятность отвергнуть гипотезу H_0, если распределения временных рядов X и X' не совпадают, с увеличением выборки стремится к единице.

Доказательство. Пусть функции распределения P и P' временных рядов не совпадают. Тогда найдется $x^* \in \mathbb{R}$, при котором значения этих функций различны: $P(x^*) \neq P'(x^*)$. Следовательно, найдется такой способ разбиения пространства $\mathbb{R}$, что для некоторого i вероятность попадания в i-й промежуток не одинакова для рассматриваемых случайных величин:

$$P(a_i < x \leqslant a_{i+1}) = p_i \neq p'_i = P'(a_i < x \leqslant a_{i+1}).$$

Пусть $p_i > p'_i$. Согласно (9) при больших m с вероятностью $\mathrm{P} > (2\Phi(C_1) - 1)(2\Phi(C_2) - 1)$ выполнено

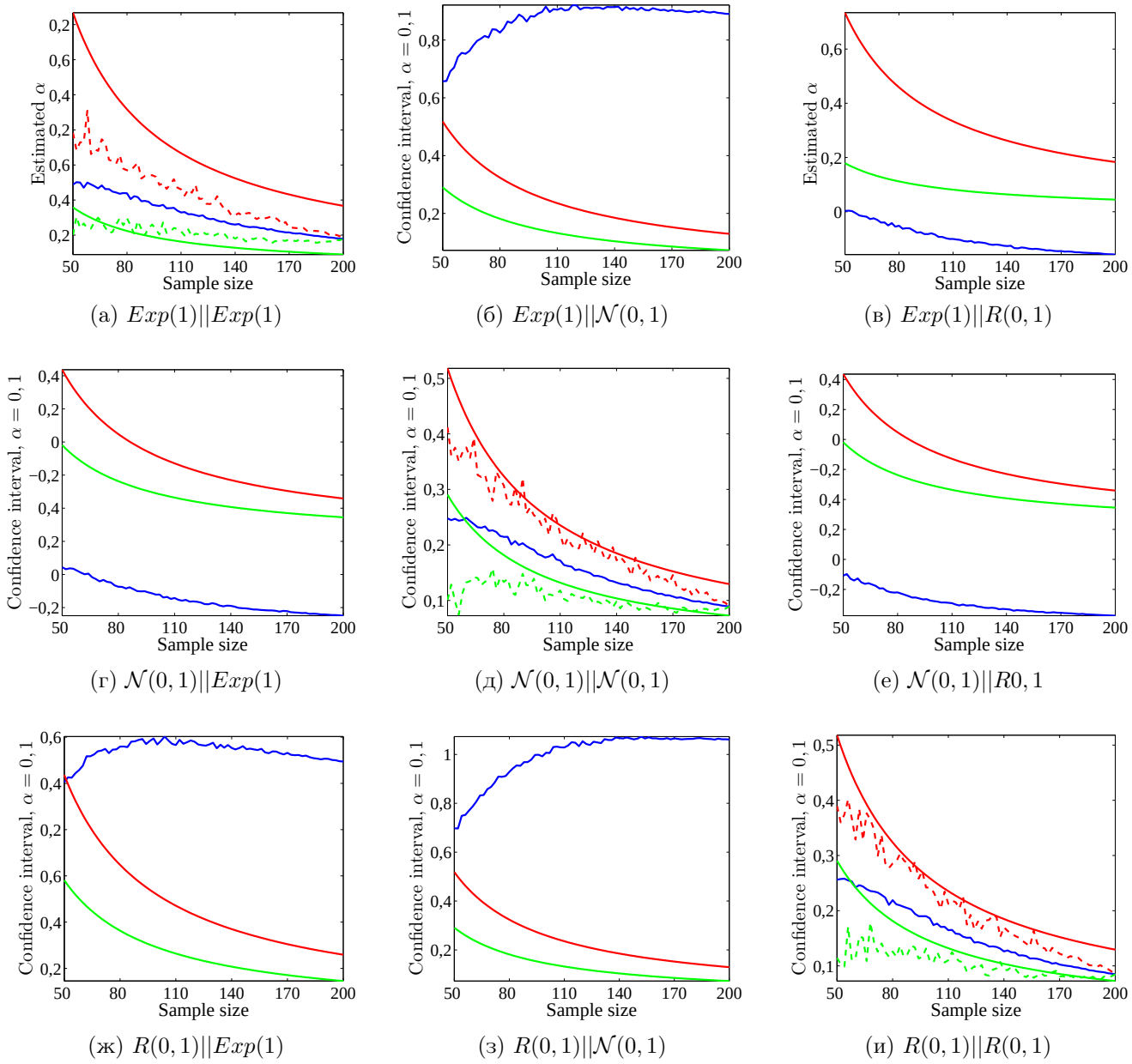
$$|n_i - mp_i| < C_1\sqrt{m}, \quad |n'_i - mp'_i| < C_2\sqrt{m}.$$

Для любого $\varepsilon > 0$ найдется константа $C_\varepsilon : \mathrm{P} > (2\Phi(C_\varepsilon)-1)^2 > 1-\varepsilon$. Выберем $C_1 = C_2 = C_\varepsilon$. Тогда $(n_i - n'_i) > m(p_i - p'_i) + O(\sqrt{m})$ и

$$\frac{(n_i - n'_i)^2}{n_i} > m\frac{(p_i - p'_i)^2}{p_i} + O(\sqrt{m}) > Cm. \tag{13}$$

Следовательно, для любого $\alpha \in (0,1)$ при достаточно больших m

$$t_m = 2mD_{\mathrm{KL}}(\hat{P}^1_m||\hat{P}^2_m) \sim \sum_{i=1} \frac{(n_i - n'_i)^2}{n_i} > Cm > \bar{t}_\alpha$$



(а) $Exp(1)||Exp(1)$ (б) $Exp(1)||\mathcal{N}(0,1)$ (в) $Exp(1)||R(0,1)$

(г) $\mathcal{N}(0,1)||Exp(1)$ (д) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$ (е) $\mathcal{N}(0,1)||R0,1$

(ж) $R(0,1)||Exp(1)$ (з) $R(0,1)||\mathcal{N}(0,1)$ (и) $R(0,1)||R(0,1)$

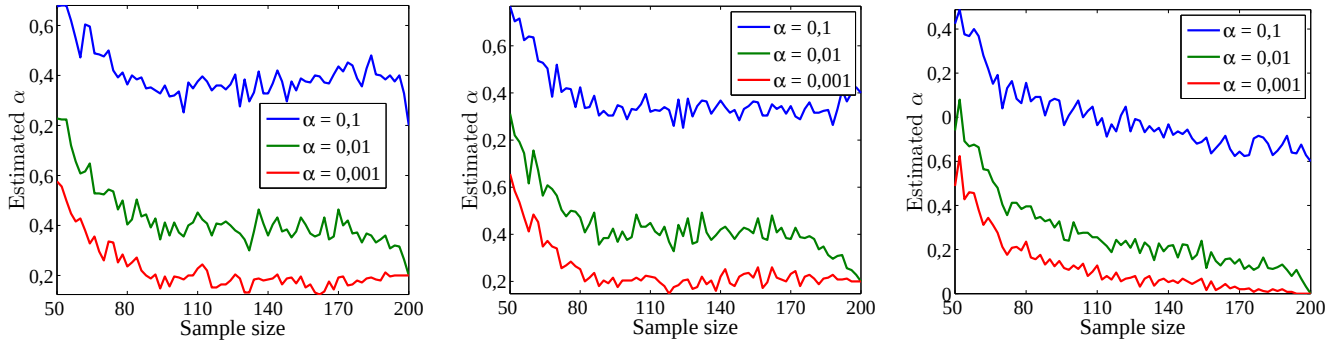
Рис. 2: Зависимость статистики t_m от объема выборки для различных пар распределений. Красным и зеленым отмечены границы доверительного интервала для t_m при $\alpha = 0,1$

с вероятностью $\mathrm{P} > 1 - \varepsilon$, то есть вероятность $\mathrm{P}(t_m > \bar{t}_\alpha) \to 1$ при $m \to \infty$. ■

5 Вычислительный эксперимент

Работа критерия была рассмотрена на различных парах распределений. Для выбранной пары распределений повторялась следующая процедура:

1) генерировались выборки X и X' одинакового объема m;
2) по выборкам строились гистограммы $\hat{P}_m$ и $\hat{P'}_m$ с фиксированным числом разбиений $N = 20$ и вычислялись расстояния Кульбака–Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P'}_m)$;
3) расстояния усреднялись по 1000 генерациям выборок;
4) объем m выборки увеличивался.



(а) $Exp(1)||Exp(1)$ (б) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$ (в) $R(0,1)||R(0,1)$

Рис. 3: Зависимость фактического уровня значимости от объема выборки при различных уровнях значимости критерия χ^2_{2N}

На каждом из графиков на рис. 2 отложены расстояния $D_{\text{KL}}(\hat{P}_m|\hat{P'}_m)$ в зависимости от объема выборки и критические значения $\bar{t}^{\chi^2}/2m$ при заданном уровне значимости α. Заметим, что в случае различных распределений расстояние $D_{\text{KL}}(\hat{P}_m|\hat{P'}_m)$ быстро попадает в критическую область и характер его зависимости от m согласуется с оценкой (13). Отрицательные значения, не характерные для расстояния Кульбака–Лейблера, возникают при численном приближении интеграла (6), когда распределение в знаменателе под знаком логарифма имеет большую область определения. Именно из-за отрицательных значений был использован двусторонний критерий. В дальнейших экспериментах было использовано симметризованное расстояние Кульбака–Лейблера (4), что позволило использовать односторонний критерий. На графиках, иллюстрирующих применение критерия к выборкам из одного распределения (рис. 2, а, д, и), также отложены пунктиром значения $\bar{t}/2m$, где $\bar{t}$ — оценки критических значений, полученные экспериментально

$$\bar{t}_\alpha = \min\{\bar{t} : \frac{1}{M}\sum_{t\in T}[t > \bar{t}] < \alpha\}.$$

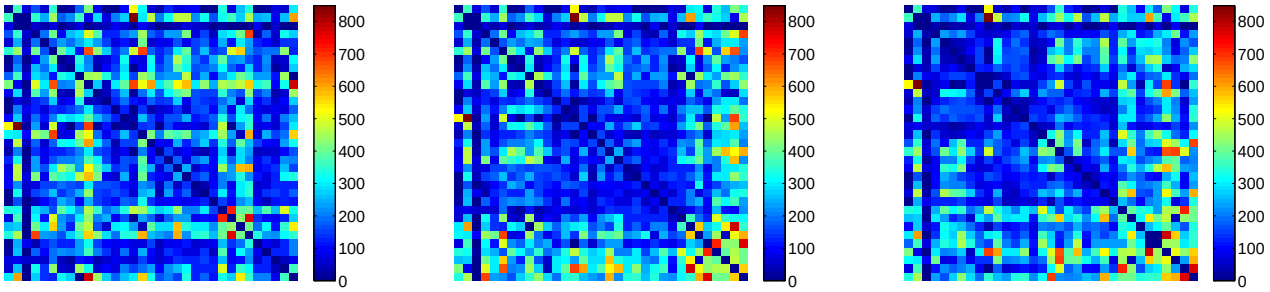
Здесь суммирование индикаторной функции $[t > \bar{t}]$ ведется по всей выборке T статистик t, полученных по M генерациям пар выборок X и X' (в данном эксперименте $M = 1000$). Видно, что, хотя критические области U^{χ^2} и U не совпадают, при истинности гипотезы H_0 статистика t_m не попадает ни в U^{χ^2}, ни в U.

Оценка фактического значения α. Так как распределение статистики t_m не совпадает с распределением χ^2, уровень значимости α, при котором определяется критическая область U^{χ^2}, не соответствует реальному уровню значимости критерия. Чтобы оценить реальный уровень значимости решения о принятии или отвержении гипотезы H_0, необходимо подсчитать долю объектов выборки T, попавших в U^{χ^2} при заданном α:

$$\hat{\alpha} = \frac{1}{M}\sum_{t\in T}[t > \bar{t}^{\chi^2}_\alpha].$$

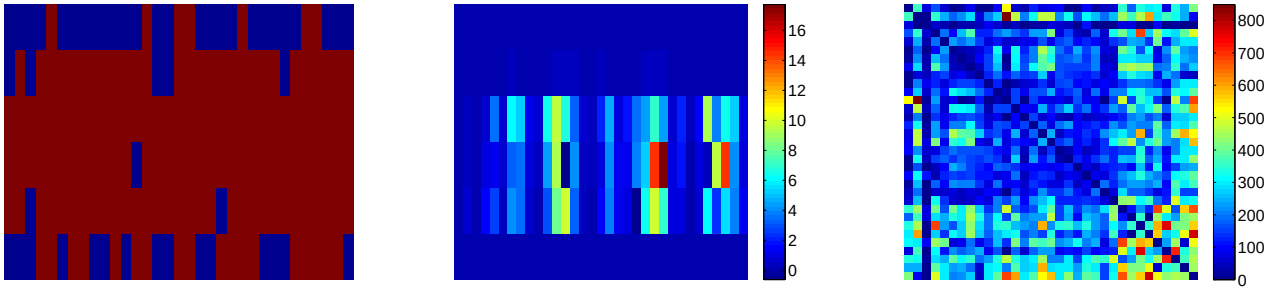
Результаты отражены на рис. 3. Из рисунков следует, что для достижения уровня значимости $\alpha = 0,1$ нужно использовать в качестве оценки U критическую область U^{χ^2} с уровнем значимости $\alpha = 0,001$.

Кластеризация временных рядов, отражающих железнодорожные грузоперевозки. Продемонстрировав таким образом статистическую значимость расстояния



(а) (б) (в)

Рис. 4: Симметризованные матрицы попарных расстояний Кульбака–Лейблера между временными рядами для различных групп грузов: до кластеризации (а) и после нее для $K = 5$ (б) и $K = 10$ кластеров (в)



(а) (б) (в)

Рис. 5: Результат проверки гипотезы о принадлежности исходного временного ряда и временного ряда, получаемого присоединением к нему некоторого более короткого ряда, к одному распределению (а). Значения статистики $t_{m,l}$ при проверке гипотезы H_0 (б). Матрица парных расстояний после объединения временных рядов и их кластеризации (в)

Кульбака–Лейблера, построим кластеризацию набора временных рядов РЖД. Данные о перевозках включают даты отправления, даты прибытия, станции внутри железнодорожной ветки и группы грузов. Для исследования были выбраны временные ряды с весами вагонов, нагруженных различными группами грузов, агрегированные по станциям. Вначале из рассмотрения были исключены ряды, содержащие менее 50 отсчетов времени. Матрицы (4) D симметризованных расстояний Кульбака–Лейблера для набора исследуемых временных рядов изображены на рис. 4. Решая задачу кластеризации (1), необходимо стремиться привести матрицу D к блочному виду. В левом столбце таблицы показано, как именно перевозимые группы грузов были разбиты на кластеры для случая $K = 5$.

Ряды, содержащие менее 50 отсчетов, последовательно присоединялись к каждому из рядов, содержавших более 50 отсчетов. Затем для объединенного ряда и исходного ряда длиной более 50 отсчетов проверялась гипотеза о принадлежности рядов к одному распределению. Результаты проверки гипотезы и значения статистики $t_{m,l}$ изображены на рисунке 5, а, б. Строки соответствуют временным рядам длиной менее 50 отсчетов, столбцы — длиной более 50 отсчетов. Результат проверки гипотезы H_0 отмечен синим цветом, если гипотеза отвергается, красным — если гипотеза принимается. Было выполнено слияние

рядов «Поваренная соль» и «Продукты промышленного потребления», а также «Флюсы» и «Шлаки гранулированные» с временным рядом «Нефть и нефтепродукты». Затем снова была выполнена кластеризация, результаты которой занесены в правый столбец таблицы. Вид матрицы парных расстояний после слияния временных рядов и их кластеризации представлен на рис. 5, в.

Таблица Группы грузов

1 — Каменный уголь, 2 — Кокс, 3 — Нефть и нефтепродукты, 7 — Руда железная и марганцевая, 8 — Руда цветная и серное сырье, 9 — Черные металлы, 10 — Метизы и оборудование, 11 — Металлические конструкции, 14 — Сельскохозяйственные машины, 15 — Автомобили, 16 — Цветные металлы, изделия из них и лом цв. металлов, 17 — Химические и минеральные удобрения, 18 — Химикаты и сода, 20 — Промышленное сырье и формовочные материалы, 22 — Огнеупоры, 23 — Цемент, 25 — Сахар, 26 — Мясо и масло животное, 27 — Рыба, 28 — Картофель, овощи и фрукты, 36 — Комбикорма, 38 — Жмыхи, 39 — Бумага, 42 — Грузы в контейнерах	1 — Каменный уголь, 2 — Кокс, 3 — Нефть и нефтепродукты, **6 — Флюсы**, **21 — Шлаки гранулированные**, 7 — Руда железная и марганцевая, 8 — Руда цветная и серное сырье, 10 — Метизы и оборудование, 11 — Металлические конструкции, 12 — Метизы, 14 — Сельскохозяйственные машины, 15 — Автомобили, 16 — Цветные металлы, изделия из них и лом цв. металлов, 17 — Химические и минеральные удобрения, 18 — Химикаты и сода, 20 — Промышленное сырье и формовочные материалы, 22 — Огнеупоры, 23 — Цемент, 25 — Сахар, 26 — Мясо и масло животное, 27 — Рыба, 28 — Картофель, овощи и фрукты, 36 — Комбикорма, 38 — Жмыхи, 39 — Бумага, 42 — Грузы в контейнерах
43 — Остальные и сборные грузы, 19 — Строительные грузы	9 — Черные металлы, 12 — Метизы
31 — Промышленные товары народного потребления, 34 — Зерно	43 — Остальные и сборные грузы, 19 — Строительные грузы, 35 — Продукты перемола
13 — Лом черных металлов	31 — Промышленные товары народного потребления, **29 — Поваренная соль**, 34 — Зерно
35 — Продукты перемола	13 — Лом черных металлов

6 Заключение

Расстояние Кульбака–Лейблера широко применяется для сравнения распределений, однако считается непригодным для использования в статистических целях из-за того, что не имеет предельного распределения. В данной работе показано, что распределение расстояния Кульбака–Лейблера в пределе ограничено сверху хи-квадратом. Это дает, пусть и ограниченную, возможность использования расстояния Кульбака–Лейблера между распределениями в качестве статистики для проверки гипотезы о принадлежности двух выборок одному распределению и позволяет говорить о статистической значимости расстояния Кульбака–Лейблера. Код, позволяющий выполнить процедуру проверки нулевой гипотезы о принадлежности выборок к одному распределению на основе расстояния Кульбака–Лейблера между их гистограммами, находится в свободном доступе [15]. В работе продемонстрировано использование предлагаемого критерия, а также приведен пример кластеризации временных рядов на основе расстояния Кульбака–Лейблера.

Литература

[1] *Вальков А. С., Кожанов Е. М., Медведникова М. М., Хусаинов Ф. И.* Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным // Машинное обучение и анализ данных, 2012. Т. 1. № 4. С. 448–465.

[2] *Вальков А. С., Кожанов Е. М., Мотренко А. П., Хусаинов Ф. И.* Построение кросс-корреляционных зависимостей при прогнозе загруженности железнодорожного узла // Машинное обучение и анализ данных, 2013. Т. 1. № 5. С. 505–518.

[3] *Медведникова М. М.* Согласование агрегированных непараметрических прогнозов временных рядов // Машинное обучение и анализ данных, 2014. Т. 1. № 8 (в печати).

[4] *Kullback S.* Information Theory and Statistics. — New York: Wiley, 1959.

[5] *Chernoff H.* A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations // Annals of Mathematical Statistics, 1952. Vol. 4. No. 23. P. 493–655.

[6] *Kolmogorov A. N.* On the approximation of distributions of sums of independent summands by infinitely divisible distributions // Contributions to statistics. — Oxford: Pergamon Press, 1965. P. 158–174.

[7] *Ali S. M., Silvey S. D.* A general class of coefficients of divergence of a distribution from another // Journal of Royal Statistical Soiety. Series B (Methodoligical), 1966. Vol. 1. No. 28. P. 131–142.

[8] *Csiszar I., Shields P.* Information theory and statistics: A tutorial // Foundations and Trend in Communications and Information Theory, 2004. No. 4. P. 417–528.

[9] *Gibbs A. L., Su F. E.* On Choosing and bounding probability metrics // International Statistical Review, 2002. Vol. 3. No. 70. P. 419–435.

[10] *Mallows C.* A note on asymptotic joint normality // Annals of Mathematical Statistics, 1972. Vol. 42. No. 2. P. 508–515.

[11] *Irpino A., Verde R., Lechevallier Y.* Dynamic clustering of histograms using Wasserstein metric // Advances in computational statistics. — Heidelberg: Physica-Verlag, 2006. P. 869–876.

[12] *Двоенко С. Д.* Неиерархический дивизимный алгоритм кластеризации // Автоматика и телемеханика, 1999. № 4. С. 117–123.

[13] *Двоенко С. Д., Пшеничный Д. О.* О метрической коррекции матриц парных сравнений // Машинное обучение и анализ данных, 2013. Т. 1. № 5. С. 606–620.

[14] *Стрижов В. В., Кузнецов М. П., Рудаков К. В.* Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах // Математическая биология и биоинформатика, 2012. Т. 7. № 1. С. 345–359.

[15] *Мотренко А. П.* Статистический тест для проверки гипотезы о принадлежности двух выборок одному распределению на основе расстония Кульбака–Лейблера // Algorithms of Machine Learning. — Sourceforge, 2014. `http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group874/Motrenko2014KL/code/KLtest.m`

Motrenko A., Strijov V.V. Obtaining an aggregated forecast of railway freight transportation using Kullback–Leibler distance // Informatics and applications, 2014, 8(2) : 86-97.

Obtaining an aggregated forecast of railway freight transportation using Kullback–Leibler distance

A. P. Motrenko[1], V. V. Strijov[2]
1 — Moscow Institute of Physics and Technology, anastasia.motrenko@gmai.com
2 — Dorodnicyn Computing Centre of RAS, strijov@ccas.com

This study addresses the problem of obtaining an aggregated forecast of railway freight transportation. To improve the quality of aggregated forecast, we solve a time series clusterization problem, such that the time series in each cluster belong to the seme distribution. Solving the clusterization problem, we need to estimate the distance between empirical distributions of the time series. We introduce a two-sample test based on the Kullback–Leibler distance between histograms of the time series. We provide theoretical and experimental research of the suggested test. Also, as a demonstration, the clusterization of a set of railway time series based on the Kullback–Leibler distance between time series is obtained.

Keywords: *empirical distribution function, distance between histograms, Kullback-Leibler distance, two-sample problem*

References

[1] Val'kov A.S., Kozhanov E.M., Medvednikova M.M., Khusainov F.I. Neparametricheskoe prognozirovanie zagruzhennosti sistemy zheleznodorozhnykh uzlov po istoricheskim dannym [Non-parametric forecasting of railroad stations occupancy according to historical data]. Mashinnoe obuchenie i analiz dannykh [Journal of Machine Learning and Data Analysis]. 4(1): 448-465.

[2] Val'kov A. S., Kozhanov E. M., Motrenko A. P., Khusainov F. I. 2013. Postroenie kross-korrelyatsionnykh zavisimostey pri prognoze zagruzhennosti zheleznodorozhnogo uzla [Constructing a cross-correlation model to forecast the utilization of a railway junction station]. Mashinnoe obuchenie i analiz dannykh [Journal of Machine Learning and Data Analysis]. 5(1): 505-518.

[3] Medvednikova M. M. 2014. Soglasovanie agregirovannykh neparametricheskikh prognozov vremennykh ryadov [Mathing aggeragared non-parametric forecasts of time series]. Mashinnoe obuchenie i analiz dannykh [Journal of Machine Learning and Data Analysis]. 8(1) (In Russian, unpubl.)

[4] S. Kullback. 1959. Information Theory and Statistics. New York: Wiley.

[5] H. Chernoff. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. Ann. Math. Statist. 4(23):493-655

[6] A. N. Kolmogorov. 1965. On the approximation of distributions of sums of independent summands by infinitely divisible distributions. Contributions to statistics. Oxford : Pergamon P. Pp. 158-174.

[7] S. M. Ali, S. D. Silvey. 1966. A general class of coefficients of divergence of a distribution from another. // Journal of Royal Statistical Soiety. Series B (Methodoligical). 1(28):131-142.

[8] I. Csiszar and P. Shields. 2004. Information theory and statistics: A tutorial. Foundations and Trend in Communications and Information Theory, 4:417–528.

[9] A. L. Gibbs, F. E. Su. 2002. On Choosing and bounding probability metrics // International Statistical Review. 3(70):419–435.

[10] C. Mallows. 1972. A note on asymptotic joint normality. Annals of Mathematical Statistics, 42(2):508–515.

[11] A. Irpino, R. Verde, and Y. Lechevallier. 2006. Dynamic clustering of histograms using Wasserstein metric. // COMPSTAT. 869-876.

[12] S. D. Dvoenko. 1999. Neirarkhicheskiy divizimnyy algoritm klasterizatsii [Non-hierarchical divisible clasterization algorithm]. Avtomatika i telemekhanika [Automation and Remote Control]. 4:117–123.

[13] S. D. Dvoenko, D. O Pshenichnyy. 2013. O metricheskoy korrektsii matrits parnykh sravneniy [On metric correction of matrices of pairwise comparisons]. Mashinnoe obuchenie i analiz dannykh [Journal of Machine Learning and Data Analysis]. 5(1): 606-620.

[14] V. V. Strizhov, M. P. Kuznetsov, K. V. Rudakov. 2012. Metricheskaya klasterizatsiya posledovatel'nostey aminokislotnykh ostatkov v rangovykh shkalakh [Metric clustering of sequences of amino acid residues in rank scales]. Matematicheskaya biologiya i bioinformatika [Mathematical Biology and Bioinformatics]. 7(1): 345–359.

[15] A. P. Motrenko. 2014. Statisticheskiy test dlya proverki gipotezy o prinadlezhnosti dvukh vyborok odnomu raspredeleniyu na osnove rasstoniya Kul'baka-Leyblera [A statistical test for the two-sampe problem based on the Kullback-Leibler distance]. `http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group874/Motrenko2014KL/code/KLtest.m`