Метод уменьшения неопределенности при оценке взаимовлияния процессов

Е. А. Чернецова
Российский государственный гидрометеорологический университет
chernetsova@list.ru

Н. И. Куракина[1], А. Д. Кузьмина[2]
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]NKurakina@gmail.com, [2]97-kuzya@mail.ru

***Аннотация*. Предложен метод уменьшения неопределенности при оценке взаимовлияния процессов с использованием информационных мер. Рассмотрены возможности применения различных информационных мер для оценки связности сигналов. Показано, что наиболее мощной мерой по критерию связности сигналов, имеющих априорное равномерное распределение и апостериорное нормальное распределение, является информационная мера Кульбака. Оценка размера базы данных временных рядов, необходимого для принятия решения о взаимовлиянии процессов производится с помощью последовательного критерия отношения вероятностей Вальда.**

Ключевые слова: случайный процесс; временной ряд; связность; информационная мера; критерий Вальда

I. Введение

Поскольку одномерная модель временного ряда обычно не может решить задачу классификации сложного объекта или явления с требуемой точностью, для этой цели широко применяются методы слияния данных нескольких временных рядов [1–3]. Меры расстояния играют решающую роль в классификации временных рядов. Однако различные меры расстояния охватывают различные аспекты связности временных рядов: поскольку соответствующие аспекты связности варьируются от приложения к приложению, ни одна из когда либо предложенных мер не может рассматриваться как наиболее оптимальная мера расстояния между временными рядами.

Применение для определения связности временных рядов корреляционного анализа корректно только в том случае, если данные наблюдений однородны и распределены по нормальному закону распределения. Но если определяется связь величин, то необходимо помнить, что отношение независимых нормально распределенных случайных величин распределено по закону Коши [4]. Проблемой распределения Коши является тот факт, что выборочное среднее не оценивает математическое ожидание. Кроме этого, элементы выборки сильно выбрасываются в хвосты распределений (принимают большие по модулю значения), как и само выборочное среднее [5], что затрудняет применение Байесовского решающего правила. Попытки же найти параметры распределений обычно приводят к тому, что одно и то же эмпирическое распределение может быть описано различными гипотетическими распределениями с тем же или лучшим согласием.

Кроме этого, частный коэффициент корреляции является мерой линейной связи между факторами. Последнее обстоятельство в случае оценки парных связей привело к тому, что в практику статистических исследований было эвристически введено большое количество мер зависимости, позволяющих учитывать нелинейный характер стохастической зависимости между случайными величинами [6].

II. Методы

Поскольку для оценки взаимовлияния процессов, представленных в виде временных рядов необходимо получить результаты сравнения сигналов, полученных от датчиков, имеющих , в общем случае, разную физическую природу, то необходимо применение таких мер связности временных рядов, в которых бы не учитывались единицы измерения записанных данных. При применении информационных мер связности временных рядов нет необходимости знать априорные параметры распределения исходных данных. Для применения информационных мер связности необходимо только синтезировать из записанных временных рядов одномерные и взаимные плотности распределения вероятности (ПРВ) сравниваемых временных рядов.

Для построения взаимных ПРВ необходимо также, чтобы два сравниваемых временных ряда были сопоставимы по какому-либо параметру, не зависящему от единиц измерения. Таким параметром может служить огибающая каждого временного ряда, нормированная к своему максимуму. После построения таких огибающих для каждого из сравниваемых сигналов, получаются два временных ряда, вся энергия которых лежит в положительной области, и они отличаются друг от друга только формой.

В процессе построение взаимной ПРВ временных рядов, выборка должна быть представлена в форме гистограммы, состоящей из столбцов с определенной протяженностью соответствующих им интервалов. Если

эти интервалы будут одинаковыми, то количество попаданий элементов выборок в разные интервалы будут сравнимы. Но также существует оптимальное количество класс-интервалов группирования, при котором ступенчатая огибающая гистограммы наиболее близка к плавной кривой распределения генеральной совокупности [7].

Поскольку оптимальное число интервалов гистограммы зависит от вида закона распределения экспериментальных данных, который в данном случае неизвестен, то рекомендуется строить равноинтервальную гистограмму, для выбора количества интервалов которой воспользоваться алгоритмом выбора подходящей расчетной формулы на основе определения энтропийного коэффициента, предложенным в работе [8].

После построения двумерной и двух одномерных ПРВ определяем взаимовлияние двух процессов на основе информационных мер связности, показывающих, насколько взаимная ПРВ двух временных рядов удалена от произведения их отдельно взятых ПРВ.

В качестве информационных мер связности двух временных рядов могут быть, в частности, приняты

1. Критерий информативности для расстояния Кульбака–Лейблера:

$$\rho_1\ (x,y\ ) = \int \log \frac{w(x,y\ )}{w(x)w(y)}\, w(x,y\ )\mu(dxdy), \qquad (1)$$

где *w(x,y)* – совместная ПРВ двух случайных временных рядов *x* и *y; w(x), w(y)* – одномерные (независимые) ПРВ двух случайных временных рядов *x* и *y; μ(dx,dy)* – доминирующая мера;

который есть ни что иное, как количество информации по Шеннону в параметре *y* о сообщении *x* [9]. При применении в формуле (1) логарифма по основанию 2 полученные численные значения меры Кульбака достаточно легко интерпретировать: при полном совпадении временных рядов *x* и *y* мера Кульбака равна 2 бита (поскольку это соответствует разрешению информационной неопределенности по двум случайным величинам. Если мера Кульбака больше 1 бит, то это значит, что величины x и y связаны между собой.

2. Критерий информативности для расстояния χ^2:

$$\rho_2\,(x,y) = \int \frac{\left(w(x,y_i) - w(x)w(y)\right)^2}{w(x,y)}\mu(dxdy) \qquad (2)$$

3. Критерий информативности для расстояния Хеллингера:

$$\rho_3(x,y) = \int \left(\sqrt{w(x,y)} - \sqrt{w(x)w(y)}\right)^2 \mu(dxdy) \qquad (3)$$

4. Критерий информативности для квадрата расстояния между плотностями вероятности:

$$\rho_4(x,y) = \int (w(x,y) - w(x)w(y))^2 \mu(dxdy) \qquad (4)$$

Расстояние Кульбака–Лейблера играет особую роль в теории информации и находит естественное применение в байесовской теории. Однако ни расстояние Кульбака–Лейблера, ни расстояние хи-квадрат не являются симметричными, а также не могут быть определены для всех точек параметрического множества в случае, когда носитель плотности зависит от параметра. Расстояние Хеллингера можно использовать для определения близости между мерами из одного семейства, индексированного различными параметрами. Это расстояние не зависит от выбора доминирующей меры и определено для всех точек параметрического множества [10].

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 работе [11] однако, было 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.

В работе [12] была установлена связь расстояния Кульбака–Лейблера с концепцией информационной меры Шеннона и Винера, а также с информационной мерой Фишера о неизвестном параметре данных. Кроме того, в этой работе было показано, что информационная мера Кульбака–Лейблера является наиболее мощной мерой по критерию связности двух временных рядов для случая, когда априорное распределение данных является равномерным, а апостериорное – нормальным. Поэтому применение информационной меры Кульбака–Лейблера в качестве оценки меры связности временных рядов оправдано при байесовском подходе в условиях полного отсутствия априорной информации [13].

Можно показать, что для гауссовских плотностей вероятности на выборке, стремящейся к бесконечности, мера Кульбака является функцией только коэффициента корреляции ρ и изменяется от 0 до ∞, когда ρ изменяется от 0 до 1.

$$\log \frac{1}{(1-\rho^2)^{1/2}} = -\log(1-\rho^2)^{1/2} = -\frac{1}{2}\log(1-\rho^2) \qquad (5)$$

После выбора информационной меры связности для определения коэффициента связности временных рядов и расчета численного значения выбранной меры необходимо определить число реализаций случайного процесса, достаточных для усреднения по ансамблю реализаций.

На данном этапе вычислений можно применить оценку Ходжеса–Лемана, относительно которой было установлено, что она мало отличается от широко распространенной для решения подобных задач статистики Вилкоксона, но уменьшает количество необходимых вычислительных затрат [14].

Оценка сдвига распределений Ходжеса–Лемана. представляет собой медиану всех возможных пар разностей элементов одной и другой групп:

$$r_{ij} = med(x_{ui} - x_{vj})\ ;\ u = 1,...n_i\ ;\ v = 1,...n_j\ . \qquad (6)$$

где x_{ui} и x_{vj} – параметры выборок по координате *i* и *j*.

Положительное свойство этой медианы состоит в том, что насколько первая группа «больше» второй, настолько вторая «меньше» первой, то есть: r_{ij}=-r_{ij}. Существенным недостатком медианы Ходжеса–Лемана является нетранзитивность. Если *x* больше *y* на a, а *s* больше *y* на b, то желательно, чтобы *s* было больше *x* на величину $a+b$. Медиана этим свойством не обладает [15].

Однако, поскольку в нашем случае имеется только две альтернативы, то статистику Ходжеса–Лемана использовать правомерно.

В качестве инструмента используем последовательный критерий отношения вероятностей Вальда (п. к. о. в.) [16], основанный на последовательных рангах, расставив замеры сравниваемых выборок так, чтобы они чередовались: $x_1, y_1, x_2, y_2, ..., x_n, y_n$. Обозначим объединенные замеры на *k*-том шаге вектором $V(k) = [v_1, v_2, ..., v_k]$, где $v_1 = x_1, v_2 = y_1$ и т.д. Пусть $S(k) = [S_1, S_2, ..., S_k]$ есть вектор последовательных рангов для $V(k)$, а

$$\lambda_k = \frac{P_k\left(S(k)/H_1\right)}{P_k\left(S(k)/H_0\right)} \qquad (7)$$

представляет собой последовательное отношение вероятностей на *k*-том шаге процесса. Если верна гипотеза H_0, то для произвольного вектора S из $S(k)$ имеем $P_k(S(k) = S/H_0) = 1/k!$ и, следовательно, можно вычислить $P_k(S(k) = S/H_1)$, учитывая, что каждый полученный вектор S соответствует взаимно однозначным образом определенному порядку объединенных замеров x_i и y_i.

Таким образом, достаточно вычислить

$$P\left(\nu_1 \le \nu_2 \le ... \le \nu_k / H_1\right) = \int\limits_{-\infty < t_1 \le t_2 \le ... \le t_k < \infty} ... \int \prod_{t=1}^{k} df_i(P(t_i))\ , \quad (8)$$

где $f_i(P(t_i)) = P(t_i)$, когда v_i есть *x* и $f_i\left(P(t_i)\right) = f\left(P(t_i)\right)$, когда v_i есть *y*.

В случае альтернатив Лемана имеем две гипотезы.

Гипотеза о связи двух временных рядов, $H_0 : G = P(X)$, против гипотезы об их независимости

$$H_1 : G = f(P(X)) = P^r(X);\ \ r > 0\ .$$

При этом последовательное отношение вероятностей на *k*-том шаге

$$\lambda_k = \frac{P_k(S(k)/H_1)}{P_k(S(k)/H_0)} = \frac{k!r^{k/2}}{\prod_{i=1}^{k}(\sum_{j=1}^{i} A_j)} \qquad (9)$$

для четных значений *k*

и

$$\lambda_k = \frac{P_k(S(k)/H_1)}{P_k(S(k)/H_0)} = \frac{k!r^{(k-1)/2}}{\prod_{i=1}^{k}(\sum_{j=1}^{i} A_j)} \qquad (10)$$

для нечетных значений *k*.

$$\text{где } A_j = \begin{cases} 1; & \text{если } v_j \text{ есть } x, \\ r; & \text{если } v_j \text{ есть } y. \end{cases} \qquad (11)$$

Таким образом, непараметрическая процедура определения количества реализации случайного процесса в банке данных, необходимого для установления связи между двумя процессами с использованием последовательного критерия отношения вероятностей Вальда сводится к следующим шагам:

Шаг 1: Получить последовательный ранг $(k+1)$-й выборки параметра классификации.

Шаг 2: Образовать вектор $A(k+1)$ из $A(k)$ и S_{k+1} …

Шаг 3. Вычислить последовательное отношение вероятностей по формуле (8) или (9) и сравнить с останавливающими границами.

Обозначим e_{ij} – вероятность принятия гипотезы H_i, тогда как в действительности верна гипотеза H_j; $i, j = 0,1$. Тогда останавливающие границы (пороги) в п.к.о.в. Вальда приближенно равны

$$A = \frac{1-e_{01}}{e_{10}},\ \ B = \frac{e_{01}}{1-e_{10}} \qquad (12)$$

А связь между средним числом измерений $E(k)$ и параметром r приближенно описывается зависимостью

$$E_r(k) \cong \frac{\log[(1-e_{10})/e_{01}]}{\log\frac{1}{2}(r^{-1/2} + r^{1/2})} \qquad (13)$$

III. Полученные результаты

Вычисление информационных мер связности сигналов достаточно легко алгоритмизируется. В пакете Матлаб была разработана программа, позволившая вычислить их для гауссовских и негаусовских сигналов, зарегистрированных датчиками мониторинговой сети от различных объектов.

Результаты тестирования информационных мер (1–4) на совокупности экспериментально полученных временных рядов при известной априорной информации о связности или несвязности сигналов датчиков приведены на рис. 1 и 2.

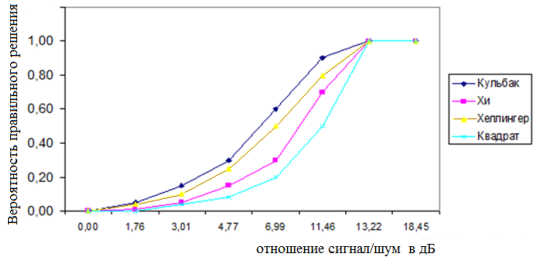
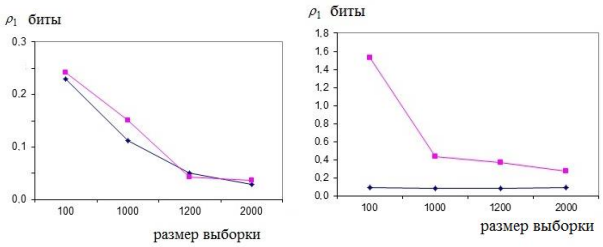


Рис. 1. Мощность информационных критериев связности при вероятности ложной тревоги p=0,1



а б

Рис. 2. Значение меры Кульбака, вычисленной по формуле (1) (квадрат) и по формуле (5) (ромб) для белого шума (а) и для сигнала от объекта (б)

Из рис. 1 можно видеть, что для экспериментально полученного банка данных сигналов наиболее мощной информационной мерой по критерию определения связности временных рядов является информационная мера Кульбака, которая есть ни что иное, как количество информации по Шеннону в параметре y о сообщении x. При полном совпадении временных рядов x и y мера Кульбака равна 2 бита (поскольку это соответствует разрешению информационной неопределенности по двум случайным величинам). Если мера Кульбака больше 1 бит, то это значит, что величины x и y связаны между собой, если мера Кульбака меньше 1, то временные ряды не связаны между собой.

Для белого шума получено примерное совпадение меры связности Кульбака, вычисленной по формуле (1) и вычисленной по формуле (5) (рис. 2*а*), тогда как для сигналов, обладающих негауссовскими ПРВ наблюдается различие в значениях меры Кульбака вычисленной по (1) и вычисленной по (5) при небольших размерах выборки. При больших размерах выборок сигналов наблюдается тенденция к нормализации процессов, поэтому можно наблюдать стремление вычисленных значений меры Кульбака по (1) и (5) друг к другу при увеличении объема выборки (рис. 2*б*).

Критерий отношения вероятностей Вальда позволяет определить количество реализаций случайного процесса, достаточное для определения связности реализаций с заранее заданной вероятностью ошибки I или II рода.

Список литературы

[1] H. Becker, M. Naaman, and L. Gravano. Learning Similarity Metrics for Event Identification in Social Media. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pages 291-300. ACM, 2010.

[2] M.F. Botsch. Machine Learning Techniques for Time Series Classification. Cuvillier, 2009, 216 p.

[3] K. Buza, A. Nanopoulos, and L. Schmidt-Thieme. Fusion of Similarity Measures for Time Series Classification. In Proceedings of the 6th International Conference on Hybrid Artificial Intelligence Systems, volume 6679 of Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI), pages 253–261, Berlin/Heidelberg, 2011. Springer.

[4] Галкин В.М., Ерофеева Л.Н., Лещева С.В. Оценки параметра распределения Коши //Труды Нижегородского государственного технического университета им. Р.Е. Алексеева. 2014. № 2(104). с. 314-319.

[5] Pisarenko V., Rodkin M. Heavy-Tailed Distributions in Disaster Analysis// Mathematical geosciences. 2011.No 43(4). Pp. 501-502.

[6] Симахин В.А. Условные меры зависимости //Вестник ТюмГУ. Физико-математическое моделирование. Нефть, газ, энергетика.. 2011. № 7. с. 119-122.

[7] Калмыков В.В., Антонюк Ф.И., Зенкин Н.В. Определение оптимального количества классов группирования экспериментальных данных при интервальных оценках// Южно-Сибирский научных вестник. 2014. №3(7), с. 56-58.

[8] Тынынька А.Н. Применение энтропийного коэффициента для оптимизации числа интервалов при интервальных оценках //Технология и конструирование в электронной аппаратуре. 2017. №3. С. 49-54.

[9] Теребиж В.Ю. Восстановление изображений при минимальной априорной информации // Успехи физических наук. 1995. Том 165. № 2, с.143-176.

[10] Шемякин А.Е. Новый подход к построению объективных априорных распределений:информация Хеллингера// Прикладная эконометрика. 2012. №4(28), с. 124-137.

[11] Мотренко А.П., Стрижов В.В. Построение агрегированных прогнозов объемов железнодорожных грузоперевозок с использованием расстояния Кульбака-Лейблера // Информатика и ее применения. 2014. №8(2), с. 86-97.

[12] Cedilnik A., Košmelj K. Relations among Fisher, Shannon-Wiener and Kullback Measures of Information for Continuous Variables. Developments in Statistics. Metodološki zvezki, 17, Ljubljana: FDV, 2002. 8 p.

[13] Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. М.: Финансы и статистика, 1983. 471 с.

[14] Роенко А.А., Лукин В.В., Зеленский А.А. Определение параметра сдвига выборки данных с симметричным негауссовым распределением на основе использования методов адаптивного робастного оценивания//Радіоелектронні і комп'ютерні системи. 2005. No2(10), с.78-87.

[15] Кремер Н.Ш. Математическая статистика. М.: Издательство Юрайт. 2017. 259 с.

[16] Шишкин А.Д., Чернецова Е.А. Последовательное решающее правило классификации аномалий на морской поверхности// Системы управления и информационные технологии. №2(52). 2013, с. 94-97.