МЕРА СИМВОЛЬНОГО РАЗНООБРАЗИЯ: ПОДХОД КОМБИНАТОРИКИ СЛОВ К ОПРЕДЕЛЕНИЮ ОБОБЩЕННЫХ ХАРАКТЕРИСТИК ВРЕМЕННЫХ РЯДОВ[1]

Ю.Г. СМЕТАНИН
доктор физико-математических наук, главный научный сотрудник,
Вычислительный центр им. А.А. Дородницына, Российская академия наук
Адрес: 119333, г. Москва, ул. Вавилова, д. 40
E-mail: smetanin.iury2011@yandex.ru

М.В. УЛЬЯНОВ
доктор технических наук, профессор кафедры прикладной математики и моделирования систем, Институт коммуникаций и медиабизнеса, Московский государственный университет печати им. Ивана Федорова;
профессор кафедры управления разработкой программного обеспечения, департамент программной инженерии, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: muljanov@mail.ru

В настоящее время рассматриваются разнообразные подходы к исследованию временных рядов в аспекте их прогнозирования. По мнению авторов, интерес представляет подход кластерного анализа, в котором объектом исследования является множество временных рядов, порожденных различными источниками. Пространство кластеризации строится на основе обобщенных универсальных характеристик временных рядов, каждая из которых является координатой этого пространства. Одному временному ряду в таком пространстве соответствует точка в координатах универсальных характеристик. Применение методов кластерного анализа позволяет выделить временные ряды, близкие по метрике пространства, а для полученных кластеров возможно решение задачи о выборе рационального метода прогнозирования. Построение специального метрического пространства для анализа временных рядов является объектом исследования данной статьи. Предметом исследования являются координаты этого пространства – обобщенные характеристики временных рядов. Авторами в ряде предыдущих работ уже были введены две координаты такого пространства – сложность временного ряда по Колмогорову и гармоническая сложность временного ряда. Настоящая статья посвящена разработке новой обобщенной характеристики временного ряда с использованием аппарата комбинаторики слов – мере символьного разнообразия. Применение подхода символьного кодирования позволяет получить представление временных рядов в пространстве слов некоторого выбранного алфавита. Исследование полученного представления методами комбинаторики слов позволяет получить оценку энтропии сдвигов как функцию длины скользящего окна. На основе исследования особенностей первой конечной разности этой функции предлагается мера символьного разнообразия временного ряда. Предложенная обобщенная характеристика может быть использована для последующего выявления характерных особенностей временных рядов, в частности, как одна из осей пространства кластеризации.

Ключевые слова: временные ряды, обобщенные характеристики, символические описания, оценка энтропии слов, мера символьного разнообразия.

[1] Работа выполнена при поддержке гранта РФФИ 13-07-00516.

1. Введение

Одним из перспективных подходов к исследованию временных рядов в аспекте их прогнозирования является подход кластерного анализа [1, 2], приводящий к выявлению групп временных рядов, обладающих близкими свойствами. Кластерный анализ базируется на построении специального метрического пространства, оси координат которого соответствуют обобщенным универсальным характеристикам временных рядов. В таком пространстве конкретному временному ряду соответствует кортеж – точка данного пространства. Последующий кластерный анализ приводит к выделению кластеров, элементами которых являются временные ряды, близкие по выбранной метрике. Для полученных кластеров может быть решена задача о назначениях методов прогнозирования: такой подход может способствовать повышению точности прогнозов за счет выбора метода, учитывающего специфику временных рядов, принадлежащих данному кластеру.

С точки зрения авторов наибольший интерес представляет именно построение специального метрического пространства, координатами которого являются обобщенные универсальные характеристики временных рядов. В аспекте построения пространства кластеризации авторами в предыдущих работах и докладах [2, 3, 4] были введены некоторые координаты такого пространства – сложность временного ряда по Колмогорову и гармоническая сложность временного ряда. Настоящая статья посвящена новой обобщенной характеристике временного ряда – мере символьного разнообразия.

Содержательно предлагаемая обобщенная характеристика отражает границу наблюдаемого разнообразия подслов над фиксированным алфавитом в слове, представляющим собой символьный код рассматриваемого временного ряда. Используемый при этом аппарат основан на методах символической динамики и символическом кодировании значений временного ряда.

2. Символическая динамика и комбинаторика слов

В данной статье для исследования временных рядов предложено использовать подход, основанный на методах комбинаторики слов и анализе энтропии, в анализе информации, представленной в виде слов над конечным алфавитом.

Комбинаторика слов – термин, введенный в широкое обращение группой математиков, публикующих результаты своих исследований под псевдонимом M.Lothaire [5]. Этим термином объединяются направления исследований, связанные общими подходами, которые ранее оставались разбросанными по различным областям математики и информатики, от теории чисел и теории динамических систем до анализа естественных языков или биологических последовательностей. В комбинаторике слов исследуется внутренняя структура слов. Плодотворность этого подхода проявилась в эффективном применении методов из одной предметной области в других областях. Типичными примерами являются применения в теории формальных языков и автоматов [6], теории групп [7], теории хаоса [8], фрактальном анализе [9], символической динамике [10] и анализе временных рядов, биоинформатике [11, 12], лингвистике и некоторых других областях.

Анализ временных рядов или очень длинных последовательностей (например, ДНК [11, 13]) также тесно связан с задачами символической динамики [14].

Символической динамикой называют раздел теории динамических систем, в котором для описания последовательностей измерений состояния системы используются символы из некоторого алфавита. Последовательности соответствуют траекториям изучаемой системы. Наиболее эффективными методы символической динамики оказываются в исследовании детерминированных систем, в которых из-за ограниченных возможностей измерения возникает сходство со случайными процессами. Описание динамики получается в терминах топологических аналогов марковских процессов – матриц возможных переходов между состояниями системы. Для построения такого описания необходимо выбрать алфавит, наиболее подходящий для представления разбиения пространства состояний системы на области, соответствующие измеряемым значениям. Сложность систем естественно оценивать числом различных конечных слов, входящих в допустимые бесконечные последовательности. Для количественной оценки целесообразно применять метрическую энтропию по Колмогорову или топологическую энтропию. Для оценки сложности индивидуальных траекторий системы можно строить оценки на основе сложности по Колмогорову. С помощью теста Колмогорова-Мартина-Лефа можно принимать решения о том, является ли индивидуальная траектория случайной.

В качестве примера можно привести задачу распознавания вторичной структуры белков [15], которая заключается в следующем. Белок можно представлять как одномерную последовательность аминокислот или как одномерную последовательность характерных локальных конфигураций. В настоящее время общепринятым является допущение, что первичная структура однозначно определяет вторичную. При этом задача определения вторичной структуры (структуры локальных конфигураций) формулируется как задача преобразования слов в алфавите имен аминокислот в слова над алфавитом локальных конфигураций с помощью кодов скользящего блока.

В данной работе рассмотрена задача анализа бесконечных последовательностей (временных рядов) на основе их достаточно длинных конечных отрезков. Предложен метод оценки энтропийных характеристик, полезный для выделения признаков в целях последующей классификации временных рядов.

3. Постановка задачи

Объектом исследования является временной ряд (произвольной природы)

$$V = \left\{ \left(f_i, t_i\right),\ f_i \in R^1, i = 1,\ldots,n \right\}, \qquad (1)$$

где f_i – значение характеристики наблюдаемого процесса в момент t_i, n – число наблюдений (отсчетов).

Предметом исследования является построение обобщенной характеристики ряда, отражающей разнообразие наблюдаемых значений. В этой постановке мы формулируем следующие задачи относительно ряда V:

- ✦ задачу символьного кодирования значений временного ряда;
- ✦ задачу построения функции оценки энтропии сдвигов;
- ✦ задачу определения меры символьного разнообразия временного ряда.

Дальнейшее изложение посвящено описанию решений сформулированных задач, предложенных авторами.

4. Символьное кодирование временного ряда по значениям

Требование универсальности пространства кластеризации налагает, очевидно, и требования к обобщенным характеристикам временных рядов, конкретные значения которых интерпретируются как координаты точки, представляющей данный временной ряд в осях этого пространства. Проблема универсализации связана с тем, что различные временные ряды имеют различную точность измерений (число значащих цифр в значении f_i) и различный масштаб по значениям. Решение проблемы авторы видят в едином масштабировании значений наблюдаемой функции процесса и построении на этой основе строки символов, отражающей динамику числовых значений исследуемого ряда.

В целях такого масштабирования мы вводим разбиение x_i, $i = 1, \ldots, m$ диапазона размаха варьирования значений f_i от $x_1 = \min\limits_{i=1,n} f_i$ до $x_m = \max\limits_{i=1,n} f_i$. Поскольку значения временного ряда могут попасть в точки разбиения, мы рассматриваем множества $[x_i, x_{i+1}) = \left\{x \mid x_i \le x < x_{i+1}\right\}$, которые далее будем называть полусегментами. Определение числа и границ полусегментов доставляется бикритериальным методом построения гистограмм [16]. Подробное изложение этого метода по отношению к символьному кодированию приведено в [3]. Число полученных этим методом полусегментов определяет мощность алфавита кодирования. Заметим, что последний элемент разбиения является, очевидно, сегментом. Выбор символов алфавита, по сути, не принципиален, но мы в дальнейшем будем использовать прописные символы латинского алфавита. Далее каждый полусегмент кодируется соответствующим символом алфавита, и мы получаем представление временного ряда в виде строки символов, например (для алфавита $\Sigma = \{A, B, C, D\}$): «*BABCDEEEDDCCCBBAABB*.....» При этом числовое значение ряда кодируется символом полусегмента, в котором оно находится. Для временного ряда, содержащего n наблюдений, мы получаем его представление в виде строки из n символов над алфавитом Σ. Полученная строка символьного кода значений может быть основой и для решения задачи символьного кодирования временного ряда по тенденциям, более подробно см. в [3].

Отметим еще одно преимущество предлагаемого подхода символьного кодирования. За редким исключением значения в отсчетах временных рядов не являются точными. Одними из таких исключений являются, например, ряды курсов валют. Для значений, имеющих погрешность измерений, в математической статистике принято строить доверительные интервалы. Используемый авторами бикритериальный метод построения гистограмм как раз и определяет ширину полусегмента гисто-

граммы, а, следовательно, и «ширину» значений для кодирующего этот полусегмент символа, на основе доверительной вероятности для среднего значения [16]. Таким образом, подход символьного кодирования более достоверно отражает исследуемый процесс с точки зрения математической статистики.

5. Функция оценки энтропии сдвигов

С целью построения предлагаемой меры разнообразия полученная символьным кодированием временного ряда строка подвергается обработке, первым этапом которой является оценка энтропии слов. Эта оценка используется как в символической динамике [10], так и в биоинформатике для оценки сложности нуклеотидных геномных последовательностей [17].

Оценка энтропии слов строится следующим образом [18]. Фиксируется длина подслова m и алфавит Σ, и далее рассматриваются все подслова длины m над алфавитом Σ. Множество различных подслов есть Σ^m, мощность этого множества $M=|\Sigma^m|$ есть общее число подслов, очевидно, что $M=k^m$. Для рассматриваемого значения m вводится произвольная нумерация подслов $i=\overline{1,M}$ и вводятся счетчики числа подслов c_i, которые изначально обнуляются. Изначально позиционированное в начале анализируемого слова длины n, окно ширины m сдвигается каждый раз на один символ. Таким образом, мы имеем $n-m+1$ позиций окна, и для каждого его положения распознается подслово, полученное в окне. Если мы наблюдаем в текущей позиции окна ширины m подслово, которое имеет номер i в принятой нумерации, то значение счетчика c_i увеличивается на единицу. По полученным значениям c_i, $i=\overline{1,M}$ и рассчитывается оценка энтропии слов C_m по следующей формуле:

$$C_m=-\sum_{i=1}^{M}\left(\frac{c_i}{n-m+1}\right)log_M\left(\frac{c_i}{n-m+1}\right). \qquad (2)$$

Заметим, что применение основания M у логарифма приводит автоматически к нормировке значений $C(m)$ – значение 0 означает, что все подслова длины m одинаковы и состоит из одного и того же подслова (фундаментальное отсутствие разнообразия). Просто показать, что значение $C(m)=1$ соответствует равночастотности всех возможных подслов в исходном слове. На основании оценки энтропии слов мы строим функцию $C(m)=C_m$, аргументом которой является длина подслова m, с областью определения: $1\le m\le n$. Функция $C(m)$ вычисляется при фиксированном m сдвигом окна ширины m по исходному слову по формуле (2) и увеличением на единицу ширины окна при изменении m от 1 до n. В соответствии с принятой в символической динамике терминологией [10] будем называть $C(m)$ функцией оценки энтропии сдвигов.

6. Монотонность функции оценки энтропии сдвигов

Для дальнейшего построения предлагаемой меры временного ряда рассмотрим поведение функции оценки энтропии сдвигов $C(m)$ как функции длины подслова m. Очевидно, что при $m=n$ мы наблюдаем всего одно подслово, совпадающее с исходным словом, и, в соответствии с (2) $C(n)=0$. При $m=1$ максимум $C(m)$ будет равен единице в случае, если частота символов алфавита в символьном представлении временного ряда одинакова.

Покажем, что при малых значениях m (длины подслова) функция $C(m)$ является монотонно убывающей. Далее мы сформулируем условие на «малость» m. Пусть алфавит $\Sigma=\{s_1,s_2,\ldots,s_k\}$, где k – мощность алфавита, c_i – кратность вхождения подслова $w_i=(\alpha_1\alpha_2\ldots\alpha_m)$, где $\alpha_i\in\Sigma$, в рассматриваемое слово. При длине подслова m функция оценки энтропии сдвигов в соответствии с (2) может быть записана в виде

$$H_i(m)=\frac{c_i}{n-m+1}log_M\left(\frac{n-m+1}{c_i}\right),\ C(m)=\sum_{i-1}^{M}H_i(m). \qquad (3)$$

При увеличении длины подслова на 1, $m\to m+1$, из каждого подслова длины $m-w_i=(\alpha_1\alpha_2\ldots\alpha_m)$ образуются подслова длины $m+1-w_{i0}=(\alpha_1\alpha_2\ldots\alpha_m s_1)$, $w_{i1}=(\alpha_1\alpha_2\ldots\alpha_m s_2)$, ..., $w_{i,k-1}=(\alpha_1\alpha_2\ldots\alpha_m s_k)$, при этом, возможно, некоторые из них будут иметь нулевую кратность, поэтому в функцию $C(m)$ при длине подслова $m+1$ вместо i-го слагаемого в (3)

$$H_i(m)=\frac{c_i}{n-m+1}log_M\left(\frac{n-m+1}{c_i}\right)$$

войдет слагаемое

$$H_i(m+1)=\sum_{j=0}^{k-1}\frac{c_{ij}}{n-m}1og_{(Mk)}\left(\frac{n-m}{c_{ij}}\right),$$

где через c_{ij} мы обозначили кратность вхождения подслова w_{ij}.

Это слагаемое достигает максимума, когда кратности всех образовавшихся подслов w_{ij} одинаковы, $c_{ij}=c_i/k$, $j=0,1,\ldots,(k-1)$, то есть его значение не превышает величины

$$H_i(m+1)_{max} = \frac{c_i}{n-m} log_{(Mk)}\left(\frac{k(n-m)}{c_i}\right).$$

При $M \cdot k = k^{m+1} < \frac{n-m}{c_i}$ выполняется неравенство

$$log_{(Mk)}k < log_{(Mk)}\left(\frac{n-m}{c_i}\right)^{1/(m+1)},$$

следовательно,

$$H_i(m+1)_{max} < \frac{c_i}{n-m} log_{(Mk)}\left(\frac{(n-m)}{c_i}\right)^{1+1/(m+1)} =$$

$$= \frac{c_i}{n-m} \frac{(1+\frac{1}{m+1}) log_M\left(\frac{(n-m)}{c_i}\right)}{log_M(Mk)} =$$

$$= \frac{\left(1+\frac{1}{m+1}\right)}{1+\frac{1}{m}} \frac{c_i}{n-m} log_M\left(\frac{(n-m)}{c_i}\right) =$$

$$\frac{m(m+2)}{(m+1)^2} \cdot \frac{n-m+1}{n-m}\left(\frac{c_i}{n-m+1} log_M\left(\frac{n-m+1}{c_i}\right)\right).$$

Можно показать, что при $n > m^2 + 3m$

$$\frac{m(m+2)}{(m+1)^2} \cdot \frac{n-m+1}{n-m} < 1,$$

следовательно,

$$H_i(m+1)_{max} < H_i(m).$$

Таким образом, при достаточно малой длине подслова m по отношению к длине слова n (т.е. при $m^2 + 3m < n$) значение функции оценки энтропии сдвигов уменьшается при переходе от m к $m+1$, следовательно, $C(m)$ как функция целочисленного аргумента является функцией, спадающей от начального значения $C(1)$, которое при больших n и близких частотах символов алфавита, как правило, близка к единице, до $C(n) = 0$.

7. Мера символьного разнообразия

Интерес представляет изучение характера убывания значений $C(m)$ с ростом аргумента. Поскольку функция $C(m)$ — «убывающая по совокупности», рассмотрим инверсную конечную разность функции $C(m)$:

$$\Delta C(m) = C(m) - C(m-1),\ m = \overline{2,n}. \qquad (4)$$

По определению $C(m)$ значения $\Delta C(m)$ ограничены, и $0 \le \Delta C(m) \le 1$, но поведение $\Delta C(m)$ может быть достаточно сложным. Предположим, что начальное значение $C(1) \approx 1$, т.е. символы алфавита кодирования временного ряда имеют слабо отличающуюся частотную встречаемость. Тогда близкие к нулю начальные значения $\Delta C(m)$, характеризует нашу символьную последовательность как последовательность, обладающую достаточно богатым разнообразием подслов. Однако функция $C(m)$ не может долго «держать единицу». Определим пороговое значение $\hat{m}$, при котором теоретически функция оценки энтропии сдвигов еще может быть равной единице. Поскольку в сдвигающемся окне ширины m при мощности алфавита кодирования $k = |\Sigma|$ может наблюдаться максимально $M = |\Sigma^m| = k^m$ различных подслов, а всего в слове длины n мы имеем $n - m + 1$ позиций окна, то максимально возможная длина подслова при котором еще можно наблюдать полное разнообразие подслов, определяется из уравнения $M = k^{\hat{m}} = n - \hat{m} + 1$, что с учетом целочисленности $\hat{m}$ приводит к уравнению $\hat{m} = \lfloor log_k(n - \hat{m} + 1) \rfloor$. В предположении, что $m << n$, значение $\hat{m} \approx \lfloor log_k n \rfloor$. В окне ширины $\hat{m} + 1$ максимально наблюдаемое разнообразие слов в k раз меньше полного разнообразия в алфавите мощности k. Поэтому мы ожидаем резкого падения значения функции $C(m)$ при $m = \hat{m} + 1$, и, следовательно, резкого скачка $\Delta C(m)$ даже для псевдослучайной последовательности символов исходного слова, обладающего конечной длиной. Таким образом, наличие ярко выраженного максимума у функции $\Delta C(m)$ при $m < \hat{m}$ означает, что начиная с данного значения m разнообразие подслов резко уменьшилось, и исходное слово обладает определенной регулярностью или периодичностью.

На основе этих рассуждений авторы и вводят меру символьного разнообразия временного ряда $\mu_s(V)$ в виде отношения значения аргумента функции $\Delta C(m)$, доставляющего ее максимум к максимально возможной ширине окна, сохраняющей полное разнообразие подслов. В этих целях определим максимум функции $\Delta C(m)$, и обозначим через m^* аргумент этого максимума

$$m^* = \underset{1 \le m \le n}{arg\ max}\ \Delta C(m).$$

Тем самым значение m^* определяет положение скачка конечной разности. Рассмотрим отношение $\mu_s(V) = m^*/\hat{m}$. Оно нормировано в [0,1], и малые значения свидетельствуют о раннем наступлении потери разнообразия и большей «простоте» исследуемого слова. Учитывая принцип построения меры, мы окончательно получаем меру символьного разнообразия временного ряда $\mu_s(V)$ в виде

$$\mu_s(V) = \frac{m^*}{\hat{m}} = \frac{arg\,max\,\Delta C(m)}{\lfloor log_k n \rfloor}. \quad (5)$$

Здесь авторы понимают меру символьного разнообразия как функционал, т.е. как индивидуальную меру на одноэлементных подмножествах. Заметим, что в соответствии с (4) для периодических слов с малым периодом с ростом длины слова (числа отсчетов исходного временного ряда) значение $\mu_s(V)$ будет уменьшаться, что соответствует логике введенной меры — для длинного периодического слова с малым периодом символьное разнообразие очевидно мало.

8. Модельный пример

Приведем модельный пример вычисления предложенной меры символьного разнообразия для бесконечной периодической строки $(ABBAAB)_n$ в алфавите $\Sigma = \{A, B\}$. Вычисленные по формулам (3) и (4) значения функций $C(m)$ и $\Delta C(m)$ приведены в *табл. 1.*

Таблица 1.

Значения функции оценки энтропии сдвигов и ее конечной разности для модельной строки

m	$C(m)$	$\Delta C(m)$
1	1,000	
2	0,959	0,041
3	0,862	0,097
4	0,646	0,215
5	0,517	0,129
6	0,431	0,086
7	0,369	0,062
8	0,323	0,046

Соответствующие графики приведены на *рис. 1* и *2.* Обе функции являются функциями целочисленного аргумента, но мы показываем их как кусочно-линейные для наглядности отображения тенденций.

Заметим, что поскольку модельное слово имеет период 6, то, начиная с окна ширины 3, мы наблюдаем всего 6 различных подслов, а поскольку $\Sigma = \{A, B\}$, то полное разнообразие подслов увеличивается вдвое при увеличении ширины окна на единицу. Таким образом, при $m = 3$ мы наблюдаем 6 подслов из 8 возможных, а при $m = 4$ — всего 6

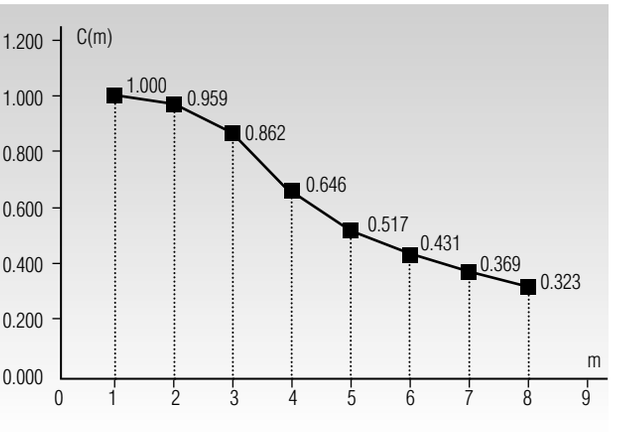

Рис. 1. График функции оценки энтропии сдвигов $C(m)$ для модельной строки

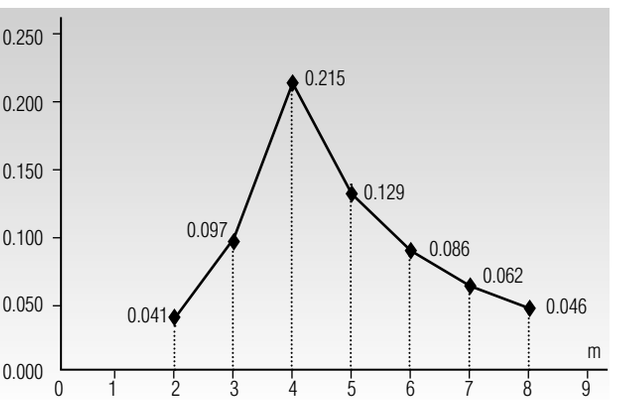

Рис. 2. График функции конечной разности $\Delta C(m)$ для модельной строки

подслов из 16 возможных, и максимум функции $\Delta C(m)$ фиксирует потерю разнообразия в окне ширины 4. Значение предложенной меры зависит от длины строки в соответствии с (5). Значения в *табл. 1* рассчитаны для бесконечного слова, однако при больших длинах, например, при $n = 1033$ эти значения изменятся незначительно. Для такой строки $m^* = 4$, а $\hat{m} = 10$ и значение меры $\mu_s(V)$=0,40.

9. Заключение

Представление временного ряда, полученное на основе символического кодирования полусегментов с использованием бикритериального метода построения гистограмм, является основной для построения функции оценки энтропии сдвигов $C(m)$, аргументом которой является ширина окна. Построение конечной разности $\Delta C(m)$ позволяет изучить особенности разнообразия подслов в исследуемом слове, а максимум этой разности сви-

детельствует о падении разнообразия как в смысле отклонения от равномерности частот подслов, так и в смысле собственно наблюдаемого разнообразия подслов. На основе исследования поведения функции $\Delta C(m)$ авторы вводят меру символьного разнообразия временного ряда $\mu_s(V)$ как отношение значения аргумента функции $\Delta C(m)$, доставляющего ее максимум, к максимально возможной ширине окна, сохраняющей полное разнообразие подслов. По принципам построения, малые значения $\mu_s(V)$ соответствуют «простым» временным рядам с вероятно хорошей возможностью их прогнозирования, а большие, близкие к единице, значения – рядам с выраженной хаотичностью. ■

Литература

1. Грабуст П. Способы оценок сходства временных рядов // Научные труды Международной конференции «Теория вероятностей, случайные процессы, математическая статистика и приложения», Минск, БГУ, 15-19 сентября 2008 г. Минск: Белорусский государственный университет, 2008. С. 23–24.
2. Ульянов М.В., Сметанин Ю.Г. Об одном подходе к построению кластерного пространства временных рядов: колмогоровская и гармоническая сложность // Proceedings of the International scientific-practical conference «Information Control Systems and Technologies» (ICST 2013). Odessa, 2013. С. 30-36.
3. Ульянов М.В., Сметанин Ю.Г. Подход к определению характеристик колмогоровской сложности временных рядов на основе символьных описаний // Бизнес-информатика. 2013. №2 (24). С. 49-54.
4. Сметанин Ю.Г., Ульянов М.В. Пространство обобщенных характеристик для классификации временных рядов: характеристика гармонической сложности // Проблемы автоматизации и управления в технических системах: Сборник статей Международной научно-технической конференции / Под ред. д.т.н., проф. М.А.Щербакова. Пенза: Изд. ПГУ, 2013. С. 125-128.
5. Lothaire M. Algebraic combinatorics on words. Cambridge, UK: Cambridge University Press, 2002. 455 pp.
6. Хопкрофт Д., Мотвани Р., Ульман Дж. Введение в теорию автоматов, языков и вычислений. М.: Издательский дом «Вильямс», 2008. 528 с.
7. Morse M., Hedlund G. Unending chess, symbolic dynamics and a problem in semigroups // Duke Mathematical Journal. 1944. No.11. P. 1-7.
8. Симиу Э. Хаотические переходы в детерминированных и стохастических системах. М.: Физматлит, 2007. 208 с.
9. Афраймович В., Угальде Э., Уриас Х. Фрактальные размерности для времен возвращения Пуанкаре. Москва, Ижевск: Институт компьютерных исследований, R&C Dynamics, 2011. 292 с.
10. Lind D., Marcus B. An introduction to symbolic dynamics and coding. Cambridge, UK: Cambridge University Press, 1995. 495 pp.
11. Математические методы для анализа последовательностей ДНК. М.: Мир, 1999. 349 с.
12. Гамов Г. Комбинаторные принципы в генетике // Прикладная комбинаторная математика / Под ред. Э.Беккенбаха. М.: Мир. 1968. С. 288-308.
13. Гасфилд Д. Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология / Пер с англ. СПб.: Невский диалект; БХВ-Петербург, 2003. 654 с.
14. Боуэн Р. Методы символической динамики. М.: Мир, 1979. 245 с.
15. Рудаков К.В., Торшин И.Ю. Об отборе информативных значений признаков на базе критериев разрешимости в задаче распознавания вторичной структуры белка // ДАН. 2011. Т. 441, № 1. С .1–5.
16. Петрушин В.Н., Ульянов М.В. Бикритериальный метод построения гистограмм // Информационные технологии и вычислительные системы. 2012. № 4. С. 22–31.
17. Орлов Ю.Л. Анализ регуляторных геномных последовательностей с помощью компьютерных методов оценок сложности генетических текстов // Дисс. на соискание уч. ст. канд. биол. наук. Новосибирск, 2004. 148 с.
18. Орлов Ю.Л. Компьютерная реализация оценок сложности текстов // Материалы Российской конференции «Дискретный анализ и исследование операций» (ДАОР), Новосибирск, Институт математики СО РАН, 28 июня – 2 июля 2004 г. Новосибирск: Изд-во Института математики СО РАН, 2004. С. 225.

MEASURE OF SYMBOLICAL DIVERSITY: COMBINATORICS ON WORDS AS AN APPROACH TO IDENTIFY GENERALIZED CHARACTERISTICS OF TIME SERIES

Yuri SMETANIN
Chief Researcher, Dorodnitsyn Computing Centre, Russian Academy of Sciences
Address: 40, Vavilova street, Moscow, 119333, Russian Federation
E-mail: smetanin.iury2011@yandex.ru

Mikhail ULYANOV
Professor, Department of Applied Mathematics and Systems Modeling,
Institute of Communications and Media Business,
Moscow State University of Printing Arts;
Professor, Software Management Department, School of Software Engineering,
Faculty of Computer Science, National Research University Higher School of Economics
Address: 20, Myasnitskaya street, Moscow, 101000, Russian Federation
E-mail: muljanov@mail.ru

Currently various approaches to time series analysis are being investigated in terms of their forecasting. In the authors' opinion, an approach to cluster analysis, which research object constitutes sets of time series generated by various sources, is of particular interest. The clusterization space is constructed by using generalized universal characteristics of time series each of which is a coordinate in this space. In such space for each time series there is a corresponding point in the coordinates of universal characteristics. Application of cluster analysis methods enables to identify time series that are space metric, and for the obtained clusters it is possible to solve the problem of choosing an efficient method of forecasting.

Construction of a special metric space to analyze time series constitutes the research object of this article. The research subject is this space coordinates – generalized characteristics of time series. In their previous articles, the authors have already defined two coordinates of such space: the Kolmogorov complexity of the time series and its harmonic complexity. This paper focuses on elaboration of a new generalized characteristic of time series by using combinatorics on words technique: a measure of symbolic diversity. The application of the symbolic coding approach enables to represent time series in a space of words in a selected alphabet. Investigation of the representation generated by combinatorics on words methods enables to estimate the entropy of shifts as a function of the length of the sliding window. A measure of symbolic diversity of time series has been proposed based on investigation of specifics of the first finite difference of this function. The proposed generalized characteristic may be applied for further identification of specific features of time series; in particular as one of the axes in the clusterization space.

Key words: time series, generalized characteristics, symbolic descriptions, words entropy assessment, measure of symbolic diversity.

References

1. Grabust P. (2008) Sposoby otsenok skhodstva vremennykh riadov [Methods of time series estimation], Nauchnye trudy Mezhdunarodnoi konferentsii «Teoria veroiatnostei, sluchainye protsessy, matematicheskaia statistika i prilozhenia» [Proceedings of the *Probability Theory, Random Processes, Mathematical Statistics and Applications: International conference* (Minsk, Belarus, September 15-19, 2008), Minsk, Belarus State University, pp. 23–24. (in Russian)
2. Ulyanov M., Smetanin Y. (2013) Ob odnom podkhode k postroeniiu klasternogo prostranstva vremennykh riadov: kolmogorovskaia i garmonicheskaia slozhnost' [On an approach to constructing a cluster space of time series: Kolmogorov and harmonic complexity] Proceedings

of the *Information Control Systems and Technologies ICST 2013: International scientific-practical conference*, Odessa, Ukraine, pp. 30–36. (in Russian)

3. Ulyanov M., Smetanin Y. Podkhod k opredeleniiu kharakteristik kolmogorovskoi slozhnosti vremennykh riadov na osnove simvol'nykh opisanii [Approach to determining of characteristics of Kolmogorov complexity of time series: An approach based on symbolical descriptions]. *Business Informatics*, no. 2 (24), pp. 49–54. (in Russian)
4. Smetanin Y., Ulyanov M. Prostranstvo obobshchennykh kharakteristik dlia klassifikatsii vremennykh riadov: kharakteristika garmonicheskoi slozhnosti [A Space of generalized characteristics for the classification of time Series: A characteristic of harmonic complexity. Proceedings of the *International Scientific and Technical Conference (Penza, Russia, 2013)* (Ed. M.Shcherbakov), Penza: Penza State University, pp. 125–128. (in Russian)
5. Lothaire M. (2002) *Algebraic combinatorics on words*, Cambridge, UK: Cambridge University Press.
6. Hopcroft J.E., Motwani R., Ullman J.D. (2000) *Introduction to Automata theory, languages, and computation,* Pearson Publication.
7. Morse M., Hedlund G. (1944) Unending chess, symbolic dynamics and a problem in semigroups. *Duke Mathematical Journal*, no. 11, pp. 1–7.
8. Simiu E. (2009) *Chaotic transitions in deterministic and stochastic dynamical systems: Applications of Melnikov processes in engineering, physics, and neuroscience*, Princeton University Press.
9. Afraimovich V., Ugalde E., Urias J. (2006) *Fractal dimensions for Poincare recurrences,* Elsevier.
10. Lind D., Marcus B. (1995) *An introduction to symbolic dynamics and coding,* Cambridge University Press.
11. Mir (1999) *Matematicheskie metody dlia analiza DNK* [Mathematical Methods of DNA Analysis]. Moscow: Mir. (in Russian)
12. Gamov G. (1968) Kombinatornye metody v genetike [Combinatorial Methods in Genetics]. *Prikladnaia kombinatornaia matematika* [Applied Combinatorial Mathematics] (Ed. E.Beckenbach), Moscow: Mir, pp. 288–308. (in Russian)
13. Gusfield D. (1997) *Algorithms on strings, trees, and sequences,* Cambridge, UK: Cambridge University Press.
14. Bowen R. (1979) *Metody simvolicheskoi dinamiki* [Methods of symbolical dynamics], Moscow: Mir. (in Russian)
15. Rudakov K., Torshin I. (2011) Ob otbore informativnykh znachenii priznakov na baze kriteriev razreshimosti v zadache raspoznavaniia vtorichnoi struktury belka [On the selection of informative features using the solvability criteria in the problem of recognition of secondary structure of the protein]. *Doklady Mathematics (Doklady Akademii Nauk),* vol. 441, no. 1, pp. 1–5. (in Russian)
16. Petrushin V., Ulyanov M. (2012) Bikriterial'nyi metod postroeniia gistogramm [Bicriterial method of histogram construction]. *Information Technologies and Computer Systems*, no. 4, pp. 22–31. (in Russian)
17. Orlov Y. (2004) *Analiz regulatornykh genomnykh posledovatel'nostei s pomoshchiu komp'iuternykh metodov otsenok slozhnosti geneticheskikh tekstov* [Analysis of regulatory genome sequences using computer methods of genetic texts complexity estimation] (PhD Thesis), Novosibirsk: Institute of Mathematics. (in Russian)
18. Orlov Y. Komp'iuternaia realizatsiia otsenok slozhnosti tekstov [Computer implementation of text complexity estimations]. Proceedings of the *Discrete Analysis and Operation Research: Russian Conference (Novosibirsk, Russia, June 28 – July 2, 2004)*, Novosibirsk: Institute of Mathematics, p. 225. (in Russian)