Сметанин Ю.Г.[1], Ульянов М.В.[2]

[1]Вычислительный центр им. А.А. Дородницына, Российская академия наук, г. Москва, д.ф.-м.н., главный научный сотрудник, smetanin.iury2011@yandex.ru

[2] Институт проблем управления им. В.А. Трапезникова, Российская академия наук, г. Москва, д.т.н., ведущий научный сотрудник, профессор кафедры алгоритмических языков факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова, muljanov@mail.ru

Энтропийные характеристики разнообразия в символьном представлении временных рядов[1]

КЛЮЧЕВЫЕ СЛОВА:

Временные ряды, обобщенные характеристики, символьные описания, оценка энтропии слов, характеристики символьного разнообразия.

АННОТАЦИЯ:

В статье предложены энтропийные характеристики временных рядов, отражающие разнообразие их символьного представления. Применение подхода символьного кодирования позволяет получить представление временных рядов в пространстве слов 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 пиковая характеристика символьного разнообразия временного ряда. Вторая предлагаемая характеристика представляет собой сумму значений функции энтропии сдвигов, превышающих определенный порог. Предложенные обобщенные характеристики могут быть использованы для последующего выявления характерных особенностей временных рядов на основе их кластерного анализа.

Введение

В настоящее время рассматриваются разнообразные подходы к исследованию временных рядов в аспекте их прогнозирования. По мнению авторов, интерес представляет подход кластерного анализа [1, 2], в котором объектом исследования является множество временных рядов, порожденных различными источниками. Пространство кластеризации строится на основе обобщенных универсальных характеристик временных рядов, которые являются координатами этого пространства. Одному временному ряду в таком пространстве соответствует точка в координатах универсальных характеристик.

Последующий кластерный анализ приводит к выделению кластеров, элементами которых являются временные ряды, близкие в смысле выбранной метрики пространства. Для полученных кластеров может быть решена задача о назначении методов прогнозирования — такой подход может способствовать повышению точности прогнозов за счет выбора метода, учитывающего специфику временных рядов, принадлежащих данному кластеру. В аспекте построения пространства кластеризации авторами в предыдущих работах и докладах [2, 3, 4]

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

были введены некоторые координаты такого пространства — сложность временного ряда по Колмогорову и гармоническая сложность временного ряда. Настоящая статья продолжает исследования авторов по данной проблематике и посвящена обобщенным характеристикам временных рядов, связанным с символьным разнообразием слов над конечным алфавитом.

Применение аппарата 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. Именно эта функция и является базой для построения предлагаемых обобщенных характеристик. Мы хотим определить как границу наблюдаемого разнообразия подслов, так и суммарную нормированную оценку энтропии сдвигов.

Содержательно первая предлагаемая обобщенная характеристика отражает границу разнообразия подслов над фиксированным алфавитом в слове, представляющим собой символьный код рассматриваемого временного ряда, а вторая — среднее значение функции энтропии сдвигов до достижения определенного минимального порога.

Предложенные обобщенные характеристики могут быть использованы для последующего выявления специфических особенностей временных рядов, в частности, они могут использоваться, наряду с другим, как оси пространства кластеризации.

Символическая динамика и комбинаторика слов

В данной статье для исследования временных рядов используется подход, основанный на методах комбинаторики слов и анализе энтропии, в 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 научным дисциплинам в математике и информатике, от теории чисел и теории динамических систем до анализа естественных языков и биологических последовательностей. Предметом исследований в комбинаторике слов является их внутренняя структура. Плодотворность этого подхода проявилась в эффективном применении методов этой научной дисциплины в других предметных 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, символьных кодов 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. Траекториям изучаемой системы соответствуют последовательности символов — слова в определенном алфавите. Наиболее эффективными методы символической динамики оказываются в

описании и 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 статье рассматривается задача анализа бесконечных последовательностей (временных рядов) на основе символьного кодирования их достаточно длинных конечных отрезков. Целью является получения оценок энтропийных характеристик, полезных для выделения специфичных признаков в целях последующей кластеризации временных рядов.

Постановка задачи

Объектом исследования является временной ряд (произвольной природы)

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

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

Предметом исследования являются обобщенные универсальные характеристики ряда, отражающие разнообразие наблюдаемых значений. В этой постановке мы формулируем следующие задачи относительно ряда T:

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

Дальнейшее изложение посвящено описанию предложенных авторами решений для сформулированных выше задач.

Символьное кодирование временного ряда по значениям

Требование универсальности пространства кластеризации разнородных временных рядов налагает, очевидно, и требования к их обобщенным

универсальным характеристикам, конкретные значения которых интерпретируются как координаты точки, представляющей данный временной ряд в таком пространстве. Проблема универсализации связана с тем, что различные временные ряды имеют различную точность измерений (число значащих цифр в значении f_i) и различную вариацию по значениям. Решение проблемы авторы видят в едином масштабировании значений наблюдаемой функции процесса и построении на этой основе строки символов, отражающей динамику их числовых значений.

В целях такого масштабирования мы определяем размах варьирования значений исследуемого временного ряда: $V = y_{\max} - y_{\min}$, где

$$y_{\min} = \min_{i=1,n} f_i,\ y_{\max} = \max_{i=1,n} f_i,$$

на котором вводим разбиение $y_i, i = 1,\ldots,m$ диапазона $[y_1, y_m]$, при этом $y_1 = y_{\min}, y_m = y_{\max}$. Поскольку значения временного ряда f_i могут попасть и в точки разбиения, мы рассматриваем множества

$$[y_i, y_{i+1}) = \{y \mid y_i \le y < y_{i+1}, i = 1,...,m-1\},$$

которые далее будем называть полусегментами. Определение числа $k = m-1$ и внутренних границ $y_i, i = 2,\ldots,m-1$ таких полусегментов представляет собой самостоятельную и интересную задачу, одним из вариантов решения которой является применение бикритериального метода построения гистограмм [16]. Подробное изложение этого метода по отношению к символьному кодированию временных рядов приведено в [3]. Число полученных этим методом полусегментов k и определяет мощность используемого 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.

Каждый полусегмент в порядке разбиения кодируется соответствующим в порядке алфавита символом, и мы получаем представление временного ряда в виде строки символов, например для алфавита $\Sigma = \{A, B, C, D\}$) и некоторого временного ряда получаем строку: «BABCDEEDCCBBA…..». 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. В качестве такого исключения приведем, например, ряды курсов валют. Для значений, имеющих погрешность измерений, в математической статистике принято строить доверительные интервалы. Используемый авторами бикритериальный метод построения гистограмм определяет ширину полусегмента гистограммы на основе построения доверительного интервала для математического ожидания относительно среднего значения в полусегменте [16], а, следовательно, и «ширину» диапазона значений для символа, кодирующего этот полусегмент. Таким образом, подход символьного кодирования более достоверно отражает

исследуемый процесс с точки зрения математической статистики для значений, имеющих погрешности измерений.

Функция оценки энтропии сдвигов

С целью построения предлагаемых характеристик 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 [10, 18]. Фиксируется длина подслова m и алфавит Σ, и далее рассматриваются все возможные подслова длины m над этим алфавитом. Множество различных подслов есть Σ^m, а его мощность — $M = \left|\Sigma^m\right| = k^m$ есть общее число различных подслов длины m в алфавите Σ. Для фиксированного значения m вводится произвольная (например лексикографическая) нумерация подслов $s_i, i = \overline{1,M}$ в Σ^m.

Для вычисления оценки введем счетчики числа подслов c_i, начальные значения которых полагаются равными нулю. Введем также в рассмотрение функцию окна $w(m, j) = s$, значением которой является подслово s длины m, расположенное в исходном слове, начиная с позиции j.

Изначально позиционированное в начало анализируемого слова длины n, окно ширины m сдвигается каждый раз на один символ вплоть до достижения конца слова. Таким образом, мы имеем $n - m + 1$ позиций окна и соответствующие значения $w(m, j), j = 1,...,n - m + 1$. Для каждого положения окна j распознается подслово $w(m, j)$, полученное в окне, для которого по введенной нумерации подслов определяется номер i, такой, что:

$$s_i = w(m, j).$$

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. Полученные значения счетчиков $c_i, i = \overline{1,M}$ являются основой для расчета оценки 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 состоят из одного и того же символа (фундаментальное отсутствие разнообразия). Просто показать, что значение $C_m = 1$ соответствует одинаковой частоте встречаемости всех возможных подслов из Σ^m 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 по формуле (2) сдвигом окна ширины m по исходному слову. В соответствии с принятой в

символической динамике терминологией [10] будем называть $C(m)$ функцией оценки энтропии сдвигов.

Особенности функции оценки энтропии сдвигов

Для дальнейшего построения предлагаемых характеристик временного ряда рассмотрим в общем случае поведение функции оценки энтропии сдвигов $C(m)$ как функции длины подслова m. Очевидно, что для произвольного слова при $m = n$ мы наблюдаем всего одно подслово, совпадающее с исходным словом, и, в соответствии с (2) $C(n) = 0$. На основании свойств энтропии, при $m = 1$ максимум $C(m)$ будет равен единице в случае, если частота символов алфавита в исходном слове одинакова.

Можно показать, что при достаточно малой длине подслова m по отношению к длине слова n, а именно при выполнении условия: $m^2 + 3m < n$, значение функции оценки энтропии сдвигов уменьшается при переходе от аргумента m к аргументу $m+1$, 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, близко к единице, до значения $C(n) = 0$. Таким образом, мы можем характеризовать функцию $C(m)$ как функцию «убывающую по совокупности».

Пиковая характеристика 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 $C(m)$:

$$\Delta C(m) = C(m) - C(m-1),\ m = \overline{2,n}. \qquad (3)$$

В силу определения функции $C(m)$ 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 — символьное разнообразие исходного слова достаточно богато. 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 $w(m, j)$, 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 значению порога

$$\hat{m} = \lfloor \log_k (n - \hat{m} + 1) \rfloor.$$

В предположении, что $n >> n$, 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 для алфавита мощности k. Поэтому мы ожидаем падения значения функции $C(m)$ при $m = \hat{m} + 1$, и, следовательно, скачка значения $\Delta C(m)$ даже для псевдослучайной последовательности символов в исходном слове, обладающим 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 пиковую характеристику символьного разнообразия временного ряда $\mu_p(T)$ 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

$$m^* = \arg\max_{2 \le m \le n} \Delta C(m).$$

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

$$\mu_p(T) = \frac{m^*}{\hat{m}} = \frac{\arg\max_{2 \le m \le n} \Delta C(m)}{\lfloor \log_k n \rfloor}. \qquad (4)$$

Заметим, что в соответствии с (4) для периодических слов с малым периодом с ростом длины слова (числа отсчетов исходного временного ряда) значение $\mu_p(T)$ будет уменьшаться, что соответствует логике введенной характеристики — для длинного периодического слова с малым периодом символьное разнообразие очевидно мало.

Кумулятивная характеристика символьного разнообразия

Наряду с пиковой характеристикой авторы предлагают рассмотреть и кумулятивную характеристику, т.е. накопленную с увеличением длины окна энтропию сдвигов. По сути это интегральная характеристика, по поскольку функция $C(m)$ является функцией целочисленного аргумента, то мы вправе лишь суммировать полученные значения.

В теории такая сумма должна рассматриваться до полной потери разнообразия, т.е. до значения $m = n$, при котором $C(n) = 0$. Однако заметим, что

если мы уже прошли пиковое значение обратной конечной разности $\Delta C(m^*)$, то значения функции энтропии сдвигов при существенно больших значениях аргумента $m >> m^*$ близки к нулю и мало информативны. На основании этих рассуждений авторы предлагают рассмотреть пороговое значение для функции $C(m)$, при достижении которого прекращается существенное накопление энтропии.

Введем в рассмотрение пороговое значение ε и определим значение аргумента $\tilde{m}$, при котором функция $C(m)$ после точки потери разнообразия станет меньше заданного порога:

$$\tilde{m}:\ m > m^*, C(\tilde{m}) < \varepsilon,\ C(\tilde{m}-1) \geq \varepsilon\ .$$

На этой основе введем нормированную кумулятивную характеристику символьного разнообразия временного ряда $\mu_s(T)$ в виде

$$\mu_s(T) = \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} C(i). \tag{5}$$

Заметим, что в силу (2) значения $C(i)$ нормированы в $[0,1]$, а следовательно в этот же сегмент нормированы и значения $\mu_s(T)$ в соответствии с (5). По принципу построения мы ожидаем, что «сложные» временные ряды, долго сохраняющие символьное разнообразие, будут иметь большие (близкие к единице) значения предложенной характеристики, чем «простые» ряды (например периодические), быстро теряющие такое разнообразие.

Модельный пример

Приведем модельный пример вычисления предложенных характеристик символьного разнообразия для бесконечной периодической строки $(ABBAAB)_\infty$ в алфавите $\Sigma = \{A, B\}$.

Предварительные вычисления. Полученные по формулам (2) и (3) значения функций энтропии сдвигов $C(m)$ и ее обратной конечной разности $\Delta C(m)$ для модельной строки приведены в таблице 1, при этом результаты проводятся с тремя значащими цифрами после запятой.

Табл. 1. Значения функции оценки энтропии сдвигов и ее конечной разности.

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 рисунках 1 и 2. Очевидно, что обе функции являются функциями целочисленного аргумента, но мы показываем их как кусочно-линейные для наглядности отображения тенденций.

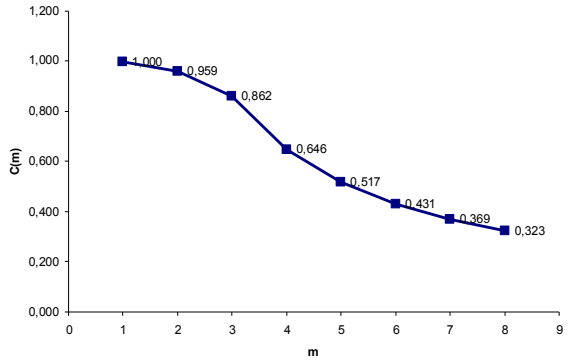

Рис. 1. График функции оценки энтропии сдвигов $C(m)$ для модельной строки

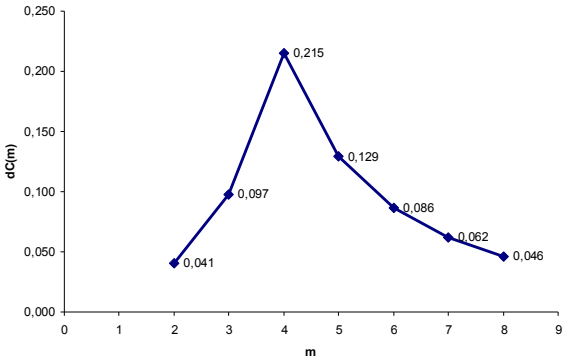

Рис. 2. График функции конечной разности $\Delta C(m)$ для модельной строки

Пиковая характеристика символьного разнообразия. Заметим, что поскольку модельное слово имеет период 6, то, начиная с окна ширины 3, мы наблюдаем всего 6 различных подслов, а поскольку $|\Sigma| = 2$, то мощность полного множества подслов увеличивается вдвое при увеличении ширины окна на единицу. При $m = 3$ мы наблюдаем 6 подслов из 8 возможных, а при $m = 4$ — тоже 6 подслов, но уже из 16 возможных, и максимум функции $\Delta C(m)$ фиксирует потерю разнообразия при ширине окна $m = 4$. Значение предложенной пиковой характеристики зависит от длины строки в соответствии с (4). Значения в таблице 1 рассчитаны для бесконечного слова, однако при больших длинах, например, при $n = 1033$ эти значения изменяются уже незначительно. Для такой строки $m^* = 4$, а $\hat{m} = 10$ и значение характеристики $\mu_p(T) = 0{,}40$. Отметим, что при этом $\Delta C(4) = 0{,}215$, что составляет более, чем $1/5$ от возможного сегмента варьирования функции $C(m)$.

Кумулятивная характеристика символьного разнообразия. Очевидно, что исследователь вправе сам выбирать пороговое значения ε для $C(m)$, необходимым является только требование единого значения порога для всех исследуемых в данный момент временных рядов. В данном случае авторы приняли, что значение конечной разности $\Delta C(m) < 0{,}05$ при $m > m^*$, уже свидетельствует о значительной потере разнообразия. Такой выбор приводит к определению $\varepsilon = 0{,}37$ для $C(m)$ и значения $\tilde{m} = 7$ (см. таблицу 1), что дает сумму в формуле (5) равную 4,784 и значение кумулятивной характеристики $\mu_s(T) = 0{,}6835$.

Заключение

Представление временного ряда, полученное на основе символьного кодирования полусегментов с использованием бикритериального метода построения гистограмм, является основной для построения функции оценки энтропии сдвигов $C(m)$, аргументом которой является ширина окна. Построение обратной конечной разности $\Delta C(m)$ позволяет изучить особенности разнообразия подслов в исследуемом слове, а максимум этой разности свидетельствует 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 пиковую характеристику символьного разнообразия временного ряда $\mu_p(T)$ как отношение значения аргумента функции $\Delta C(m)$, доставляющего ее максимум, к максимально возможной ширине окна, сохраняющей полное разнообразие подслов. Вторая — кумулятивная характеристика $\mu_s(T)$ отражает поведение накопленной энтропии слов, позволяя оценить среднее значение энтропии в символьном коде временного ряда. По предложенным принципам построения, малые значения $\mu_p(T)$ и $\mu_s(T)$ соответствуют «простым» временным рядам с вероятно хорошей возможностью их прогнозирования, а большие, близкие к единице, значения — рядам с ярко выраженной хаотичностью.

Литература

1. Грабуст П. Способы оценок сходства временных рядов // Научные труды Межд. Конф. «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 // Сборник статей МНТК «Проблемы автоматизации и управления в технических системах» / 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 конференции «Дискретный анализ и исследование операций» (ДАОР), Новосибирск, Институт математики СО РАН, Новосибирск: Изд-во Института математики СО РАН, 2004. С. 225.