ПОДХОД К ОПРЕДЕЛЕНИЮ ХАРАКТЕРИСТИК КОЛМОГОРОВСКОЙ СЛОЖНОСТИ ВРЕМЕННЫХ РЯДОВ НА ОСНОВЕ СИМВОЛЬНЫХ ОПИСАНИЙ[1]

Ю.Г. Сметанин,
доктор физико-математических наук, главный научный сотрудник Вычислительного центра им. А.А. Дородницына Российской академии наук

М.В. Ульянов,
доктор технических наук, профессор кафедры управления разработкой программного обеспечения Национального исследовательского университета «Высшая школа экономики»,
профессор кафедры прикладной математики и моделирования систем Московского государственного университета печати имени Ивана Федорова

E-mail: smetanin.iury2011@yandex.ru, muljanov@mail.ru
Адрес: г. Москва, ул. Кирпичная, д. 33/5

В статье предложен подход к исследованию временных рядов, основанный на определении сложности по Колмогорову строк символов, являющихся представлением временных рядов в пространстве слов некоторого выбранного алфавита. В рамках данного подхода описаны методики символьного описания временных рядов по уровням и по тенденциям. В основу описания по уровням положен разработанный при участии одного из авторов бикритериальный метод построения гистограмм. На основе оценок колмогоровской сложности строк, полученных с помощью общеизвестных алгоритмов сжатия, построены характеристики сложности временных рядов, которые могут быть использованы для выявления их характерных особенностей на основе последующей кластеризации.

Ключевые слова: временные ряды, символические описания, сложность по Колмогорову, бикритериальный метод, гистограммы, сжатие данных, кластеризация.

1. Введение

Основные задачи исследования как одномерных, так и многомерных временных рядов преследуют, прежде всего, цель повышения точности прогнозирования их поведения и адекватности соответствующих прогностических математических моделей. В этом аспекте исследуются структуры временных рядов, вводятся различные классификации, отражающие особенности порождающих эти ряды процессов, предлагаются разнообразные методы прогнозирования и математические аппараты [1]. Тем не менее, предлагаемые

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

классификации временных рядов, как правило, являются классификациями по одному признаку, причем, в основном, с качественным, а не количественным классификатором.

Одной из альтернатив является построение специального метрического пространства, координатами которого являются обобщенные универсальные характеристики временных рядов. Кластерный анализ в таком пространстве приводит к выделению кластеров, элементами которых являются временные ряды, близкие по особенностям в метрике данного пространства. Дальнейшее исследование особенностей полученных кластеров в аспекте выбора рациональных методов прогнозирования может способствовать повышению точности прогнозов за счет выбора метода, учитывающего специфику временных рядов в данном кластере.

В рамках данной статьи авторы вводят координаты такого пространства, основанные на сложности временного ряда по Колмогорову. Содержательно колмогоровская сложность есть характеристика строки символов, отражающая сложность (в смысле длины записи) алгоритма и его входа, генерирующих данную строку, иными словами длину формального описания строки. В теории колмогоровской сложности такой алгоритм носит название декомпрессора, а сама сложность определяется как минимальная длина оптимального способа описания строки, где минимум берется по всем описаниям [2]. Отметим, что колмогоровская сложность определена с точностью до константы [2].

При фиксированном алгоритме сжатия строк и при фиксированной длине исходных строк, оценка верхней границы колмогоровской сложности может быть получена через измерение длин сжатых строк. В теории сложности строк по Колмогорову известно, что существуют строки «не сжимаемые на 1» [2] – в аспекте временных рядов это означает существенную случайность значений и значительные трудности при их прогнозировании. Если длины полученных сжатых строк существенно меньше исходной длины, то можно говорить о возможности хорошего прогноза, например, такая ситуация характерна для чисто периодических временных рядов, наблюдаемых на протяжении многих периодов.

2. Постановка задачи

Рассмотрим временной ряд

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

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

Для указанного ряда мы формулируем следующие задачи:

- ✦ задачу символьного кодирования значений временного ряда по уровням, включая подзадачу разбиения размаха варьирования значений на полусегменты;
- ✦ задачу символьного кодирования значений временного ряда по тенденциям;
- ✦ задачу оценки колмогоровской сложности полученных строк символов;
- ✦ задачу определения характеристик колмогоровской сложности временного ряда.

Изложению предлагаемых авторами решений сформулированных задач и посвящена настоящая статья.

3. Символьное кодирование временного ряда по уровням

Поскольку сложность по Колмогорову определена для строк над некоторым алфавитом Σ, возникает задача представления временного ряда T строкой символов над данным алфавитом. Возникающие на этом пути проблемы связаны с тем, что различные временные ряды имеют различную точность измерений (число значащих цифр в значениях элементов ряда) и различный масштаб по значениям, что не позволяет использовать непосредственное сжатие исходного ряда для оценки его колмогоровской сложности. В качестве решения авторы предлагают ввести единое (по методике) масштабирование значений наблюдаемой функции процесса и построение на этой основе строки символов (для которой и определено понятие сложности по Колмогорову), отражающей числовые значения исследуемого ряда.

В целях такого масштабирования на диапазоне размаха варьирования значений функции процесса (значений ряда) мы вводим разбиение на полусегменты (что равноположено первому шагу интегрирования по Лебегу), определение числа которых также представляет отдельную задачу, примитивное решение которой доставляется разбиением размаха варьирования на фиксированное число полусегментов равной длины. Число полусегментов определяет мощность алфавита, каждый полусегмент кодируется символом этого алфавита, и проходом по временному ряду мы получаем его кодирование (представление) строкой символов.

При этом числовое значение f_i кодируется именем (символом) полусегмента, в котором оно находится. Отметим, что кодирование значений по именам полусегментов отражает и подход интервального анализа, поскольку истинные значения временного ряда, за исключением некоторых финансовых рядов (типа рядов курсов валют), очевидно, находятся в некотором доверительном интервале. Для решения обозначенной выше задачи масштабирования диапазона значений временных рядов могут быть предложены разнообразные подходы — от равномерного разбиения до подхода к решению задачи определения числа и длины полусегментов на основе аппарата математической статистики.

Еще один вопрос связан с масштабированием исследуемого множества временных рядов по числу наблюдений. Очевидно, что различные исследуемые временные ряды содержат не равное число наблюденных значений. В рамках принятого подхода символьного кодирования это приводит к появлению строк различной длины в фиксированном алфавите. Поэтому очевидным является решение о переходе от оценки абсолютной сложности строки по Колмогорову в виде длины сжатой строки к относительной оценке – коэффициенту сжатия. В связи с этим именно значение коэффициента сжатия авторы и предлагают использовать как основу для одной из обобщенных универсальных характеристик временного ряда.

Дополнительное исследование полученной строки символов может быть проведено и аппаратом символической динамики с целью выявления запрещенных подслов и описания пространства сдвигов, к которому принадлежит данная строка [3]. Пусть, например, кодирование значений временного ряда осуществляется в алфавите $\Sigma = (A,B,C,D,E,F)$, символами которого обозначаются полусегменты значений наблюдаемой величины в порядке их возрастания: A – имя полусегмента наименьших значений, F – наибольших. Если наблюдения ведутся в дискретном времени, то описание значений временного ряда по именам полусегментов есть слово над алфавитом имен полусегментов. В случае, если наблюдаемый процесс характеризуется резкими выбросами значений наблюдаемой величины (до уровня) относительно базального уровня (*A, B*) за один дискрет времени, равно как и резкими спадами (от *F* до *B*), то получаемые кодовые слова временного ряда не будут содержать подслов *CDE* и *EDC*. Тем самым язык символьного кодирования такого временного ряда есть язык над указанным выше алфавитом, с запретами подслов *CDE* и *EDC*, определяющими пространство сдвигов, при рассмотрении порожденных слов со все более возрастающей длиной. Обратно, гладким периодическим временным рядам с плавно изменяющимися значениями соответствует язык символического кодирования, содержащий запреты подслов *AF, BF, FB, FA.*

4. Разбиение множества значений временного ряда на полусегменты

Рациональное разбиение размаха варьирования временного ряда на полусегменты в целях последующего символического кодирования является самостоятельной и достаточно сложной задачей. Для ее решения авторы предлагают применить бикритериальный метод построения гистограмм, предложенный одним из авторов и В.Н. Петрушиным в [4] и считают необходимым привести здесь его краткое изложение.

В дальнейшем изложении этой части статьи в согласии с обозначениями математической статистики мы понимаем под выборкой значения временного ряда f_i, обозначая вариационный (сортированный по возрастанию) ряд этих значений через $\tilde{x}_i$. Метод в целом основан на построении системы из двух критериев, приводящих к обоснованному выбору как числа полусегментов гистограммы, так и их длин.

Первый из них основан на применении критериев согласия. Полученная некотором методом гистограмма может рассматриваться как аппроксимация неизвестного закона распределения кусочно-равномерными функциями плотностей (по полусегментам). Обозначим полученную интегрированием гистограммы на полном размахе варьирования кусочно-линейную аппроксимацию эмпирической функции распределения вероятностей через $F_G(x), x \in [\tilde{x}_1, \tilde{x}_n]$. Таким образом, возникает частная задача проверки гипотезы о соответствии эмпирической функции распределения $F_V(\tilde{x}_i)$, построенной по значениям временного ряда, рассматриваемой как эталонная, и гистограммной функции $F_G(x)$, вычисленной в точках вариационного ряда $F_G(\tilde{x}_i)$. Для решения этой задачи метод использует критерий Колмогорова. В рассматриваемой ситуации статистикой критерия является величина

$$D_n = \max_{i=1,n} \left| F_V(\tilde{x}_i) - F_G(\tilde{x}_i) \right|,$$

которая подчиняется следующему интегральному закону распределения вероятностей

$$\lim_{n\to\infty} P\left(\sqrt{n} D_n \le x\right) = K(x) = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2x^2}.$$

В [4] предложено в качестве первого критерия использовать значение вероятности ошибки первого рода α в точке наблюдаемого значения статистики критерия Колмогорова, т.е. в точке $x = \sqrt{n}D_n$. Эта вероятность в [4] обозначена через $\alpha(V, G)$, поскольку аппроксимация фиксированной выборки V различными гистограммами G приведет к изменению наблюдаемого значения критерия D_n, а следовательно и вероятности $\alpha(V, G)$. Аналитическая формула для вычисления $\alpha(V, G)$ имеет вид

$$\alpha(V,G) = \int_{\sqrt{n}D_n}^{\infty} K'(x)dx = 1 - K\left(\sqrt{n}D_n\right).$$

Увеличение числа полусегментов гистограммы приведет, очевидно, к лучшей аппроксимации эмпирической функции распределения, тем самым наблюдаемое значение критерия Колмогорова D_n (при фиксированной выборке $n = const$) будет уменьшаться, нижний предел интеграла будет смещаться влево, что приведет к увеличению значения $\alpha(V, G)$.

Второй компонент критерия представляет собой показатель надежности оценки среднегруппового значения в полусегменте [4]. Из математической статистики известно, что интервальная оценка средней групповой формируется на основе распределения Стьюдента. Пусть $\overline{x}_j$ — выборочная групповая средняя в j-ом полусегменте, а $\overline{X}_j$ — математическое ожидание групповой средней. Тогда при заданной надежности (доверительной вероятности) γ_j доверительный интервал для $\overline{X}_j$ определяется в виде:

$$\overline{X}_j \in \left(\overline{x}_j - \delta_j, \overline{x}_j + \delta_j\right), \delta_j = \frac{t\left(\gamma_j, n_j\right)\cdot S_j}{\sqrt{n_j}},$$

где $t\left(\gamma_j, n_j\right)$ — значение критерия Стьюдента при выбранной доверительной вероятности γ_j и объеме группы, а $S_j = \sqrt{S_j^{\,2}}$, где $S_j^{\,2}$ — несмещенная оценка внутригрупповой дисперсии в j-ом полусегменте. Обращением данной формулы в случае уже имеющейся гистограммы можно вычислить оценку доверительной вероятности (надежности) γ_j:

$$\gamma_j = t^{-1}\left(\frac{\delta_j \cdot \sqrt{n_j}}{S_j}, n_j\right),$$

Очевидно, что принимая гипотезу о независимости групповых средних, надежность гистограммы в целом $\gamma(G)$ будет представлять собой произведение надежности всех групповых средних γ_j, таким образом второй компонент критерия оценки качества гистограммы представим в виде

$$\gamma(G) = \prod_{j=1}^{k}\gamma_j = \prod_{j=1}^{k} t^{-1}\left(\frac{\delta_j \cdot \sqrt{n_j}}{S_j}, n_j\right).$$

Описываемый метод использует следующую бикритериальную оценку качества гистограммы

$$Q(V,G) = Q\left(\alpha(V, G), \gamma(G)\right) = \alpha(V,G)\cdot\gamma(G),$$

которая позволяет строить гистограммы, определяя как число полусегментов, так и их длину. Именно этот метод авторы и предлагают применить для получения рационального решения задачи разбиения размаха варьирования временного ряда в целях его символьного кодирования.

Со ссылкой на [4] приведем пример применения этого метода к тестовой выборке, на которой получено улучшение значение критерия по сравнению с равномерным разбиением с $Q(V,G) = 0,392$ до $Q\left(V,G^*\right) = 0,491$. Значения компонент критерия приведены *в табл. 1.* При этом рациональное значение числа полусегментов осталось равным 11.

Таблица 1.

Значения $Q(V, G)$ для равномерного разбиения и бикритериального метода

k	$\gamma(G)$	$\alpha(V, G)$	$Q(V,G)$
11 (равномерно)	0,963	0,407	0,392
11 (бикритериальный метод)	0,989	0,496	0,491

Полученная бикритериальным методом гистограмма приведена на *рис. 1а.*

На рис. 1б для сравнения показана гистограмма с полусегментами равной длины. Отметим качественные отличия гистограмм: бикритериальный метод позволил выявить бимодальный характер выборки, в то время, как гистограмма с равномерным разбиением в окрестности моды имеет унимодальный характер.

5. Символьное описание временного ряда по тенденциям

В ряде случаев интерес представляет не реальное изменение значения исследуемого процесса в следующий момент дискретного времени, а изменение его тенденции. Отметим, что целый ряд методов прогнозирования временных рядов, особенно экономического характера, ориентирован специально на прогноз тенденций. Возникающая при этом задача определения рациональных порогов идентификации смены тенденции является достаточно сложной. Действительно: увеличение значения на 1% — это уже положительная тенденция или еще отсутствие таковой?

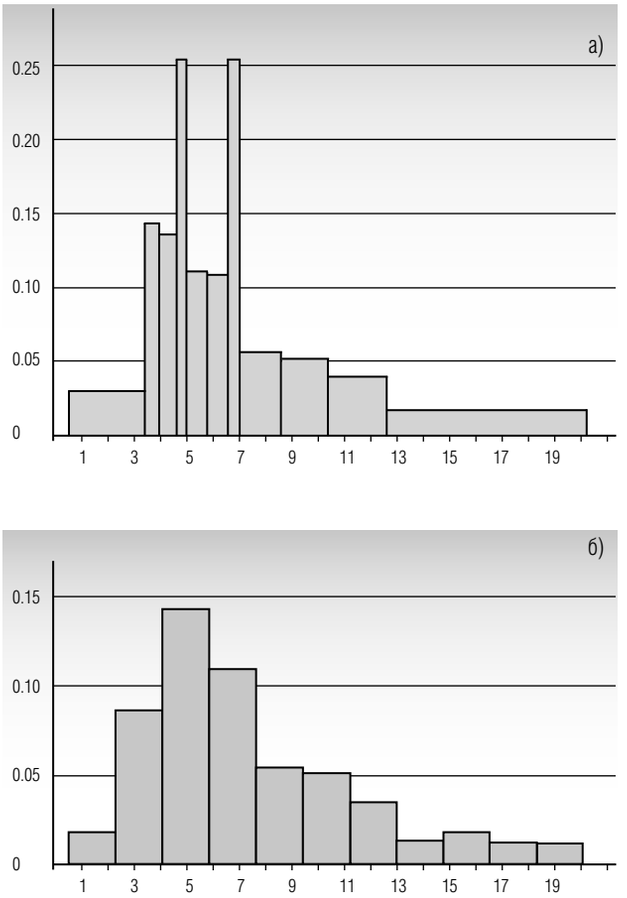

Рис. 1. Гистограмма по предложенному методу (а), и по полусегментам равной длины (б).

Возможные решения этой задачи, как правило, опираются на специальную предварительную обработку исходных значений или на применение метода экспертных оценок. В последнем случае решение не является математически обоснованным и отражает специфику проблемной области временного ряда с точки зрения данной группы экспертов.

Для решения этой задачи авторы предлагают использовать уже полученную информацию о символьном кодировании по значениям, которую в данном случае мы интерпретируем как данные предварительной обработки. Поскольку бикритериальный метод построения полусегментов гарантирует, что доверительный интервал для выборочного среднего в полусегменте не шире самого полусегмента, то локализация значений, кодируемых одним символом алфавита является статистически достоверной. С другой стороны адекватность (в смысле критерия согласия Колмогорова) эмпирической функции распределения и гистограммы, построенной на полусегментах гарантирует адекватность мощности алфавита символьного описания. Таким образом, мы предлагаем считать, что в рамках символьного кодирования изменение символа в следующий дискрет времени есть квалификация тенденции, а изменение, не выводящее значение за данный полусегмент — отсутствие такового.

Пусть кодирование по тенденциям осуществляется в алфавите $\Sigma_r = \{-, 0, +\}$, где символом 0 обозначается отсутствие тенденции в значении для следующего дискрета времени. Если мы кодируем значения временного ряда в алфавите $\Sigma_y = (A, B, C, D, E, F)$, то, например, слово кода значений

$$BBACDECCCABDDDE$$

будет кодировано с использованием предложенного метода в алфавите тенденций следующим словом:

$$00-+++-00-++00+,$$

где мы по умолчанию предполагаем, что первый символ кода тенденций — всегда «0».

6. Оценка колмогоровской сложности строки символов

Описание временного ряда, полученное на основе символического кодирования полусегментов, или кодирования тенденций и представляет собой то слово, для которого путем вычисления коэффициента сжатия и будет определяться оценка верхней границы колмогоровской сложности временного ряда. Отметим. что речь идет именно об оценке колмогоровской сложности, поскольку мы предполагаем использование любого широко распространенного алгоритма сжатия, а точнее — некоторой его программной реализации.

Таким образом, пусть $S(T, \Sigma)$ есть функция кодирования временного ряда T символами алфавита Σ, значением которой является строка s:

$$s = S(T, \Sigma), \qquad (2)$$

пусть также $C(\cdot)$ есть оператор сжатия строки, которая является его аргументом, реализуемый любым, но фиксированным, алгоритмом сжатия. Результатом применения оператора $C(\cdot)$ к строке s является строка w:

$$w = C(s). \qquad (3)$$

Именно длина этой строки и является классически [3] оценкой колмогоровской сложности. Отметим, что в теории колмогоровской сложности обратный оператор $s = C^{-1}(w)$ называется декомпрессором [3]. Переход к относительным единицам очевиден: в этих обозначениях коэффициент сжатия строки s определяется как:

$$\mu(s,C)=\frac{l(s)}{l(w)}=\frac{l\left(S(T,\Sigma)\right)}{l\left(C\left(S(T,\Sigma)\right)\right)}, \qquad (4)$$

где $l(\cdot)$— длина строки.

Именно значение $\mu(s,C)$ авторы и будут использовать в дальнейшем для построения характеристик колмогороввской сложности временного ряда. Напомним, что мы можем получить два, быть может отличающихся по значениям, коэффициента сжатия — один для строки символов, содержащей символьное кодирование значений временного ряда $\mu_v(s,C)$, а второй — для строки символьного кодирования тенденций — $\mu_r(s,C)$.

7. Построение характеристик колмогоровской сложности временных рядов

Могут быть предложены различные варианты преобразования значений коэффициентов сжатия $\mu_v(s,C)$ и $\mu_r(s,C)$ в значения, соответствующее данному временному ряду по координатам колмогоровской сложности значений и тенденций в пространстве кластеризации.

Например, возможен следующий вариант. Коэффициент сжатия есть отношение длины исходной строки к длине сжатой строки, и, по определению, не может быть меньше единицы. Тогда нормировка в значение координаты выполняется вычитанием единицы из значения коэффициента сжатия, и в целях обеспечения наглядности, мы используем значение, обратное к полученному. Обозначим такие характеристики через $D_v(T)$ и $D_r(T)$, тогда

$$D_v(T)=\frac{1}{\mu_v(s,C)-1}, D_r(T)=\frac{1}{\mu_r(s,C)-1},$$

где соответствующие значения $\mu_v(s,C)$ вычисляются по (4) а s и w определяются по временному ряду T на основе (2) и (3).

Полученные значения $D_v(T)$, $D_r(T)$, и есть характеристики временного ряда по координатам колмогоровской сложности значений и тенденций в пространстве кластеризации. При такой нормировке малые положительные значения соответствуют большим коэффициентам сжатия, и, следовательно, временным рядам с простой регулярной структурой. Большие значения характеризуют временные ряды с коэффициентом сжатия близким к единице, т.е. ряды, обладающие выраженной случайностью (в мере колмогоровской сложности, но не в мере случайности по Колмогорову [3]). Временные ряды, обладающие большими значениями характеристики колмогоровской сложности, по мнению авторов, должны обладать плохой предсказуемостью или коротким (по времени) приемлемым результатом прогноза.

8. Заключение

В статье предложен подход к исследованию особенностей временных рядов, основанный на оценке их колмогоровской сложности на основе коэффициента сжатия символьного кода временного ряда. Предлагаемое разбиение размах варьирования значений на полусегменты для символьного кодирования основано на предложенном одним из авторов (совместно с В.Н. Петрушиным) бикритериальном методе построения гистограмм. Полученные оценки относительной сложности временного ряда по Колмогорову служат базой для вычисления меры сложности временного ряда, являющейся одной из осей кластерного пространства временных рядов, при символическом кодировании значений. В статье описан так же переход от символьного кодирования по значениям к символьному кодированию по тенденциям, позволяющему ввести еще одну координату пространства кластеризации временных рядов.

Предполагаемое авторами в дальнейшем исследование особенностей методов прогнозирования по отношению к кластерам временных рядов позволит указать наиболее рациональные методы для выделенных кластерных групп. Очевидно, что наиболее интересной и научно значимой задачей является построение разнообразия координатных осей самого пространства кластеризации, равно как и введение функции расстояния для определения в этом координатном пространстве структуры метрического пространства. ■

Литература

1. Любушин А.А. Анализ данных систем геофизического и экологического мониторинга. – М.: Наука, 2007.
2. Верещагин Н.К., Успенский В.А., Шень А. Колмогоровская сложность и алгоритмическая случайность. – М.: МЦНМО, 2013.
3. Lothaire M. Algebraic Combinatorics on Words. – 2005.
4. Петрушин В.Н., Ульянов М.В. Бикритериальный метод построения гистограмм // Информационные технологии и вычислительные системы. – 2012. – № 4. – С. 22-31.