УДК 004.67, 519.72

Д.ф.-м.н. Сметанин Ю.Г., д.т.н. Ульянов М.В.

ВЦ РАН им. А.А. Дородницына, НИУ ВШЭ

МЕРА СИМВОЛЬНОГО РАЗНООБРАЗИЯ — ХАРАКТЕРИСТИКА ВРЕМЕННЫХ РЯДОВ

Dr. Sci. Smetanin Y.G., Dr.Sci. Ulyanov M.V.

MEASURE OF SYMBOLICAL DIVERSITY — CHARACTERISTIC OF TIME SERIES

В докладе предлагается новая обобщенная характеристика временных рядов — мера символьного разнообразия. Применение подхода символьного кодирования позволяет получить представление временных рядов в пространстве слов некоторого выбранного алфавита [1]. Исследование полученного представления методами комбинаторики слов позволяет получить оценку энтропии сдвигов как функцию длины скользящего окна.

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.

Для оценки энтропии слов фиксируется длина подслова m и алфавит Σ, и далее рассматриваются все подслова длины m над алфавитом Σ. Вводится произвольная нумерация подслов $i = \overline{1, M}$ и счетчики числа подслов c_i, которые изначально обнуляются. Если в текущей позиции окна подслово с номером i, 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 [2]:

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

Просто показать, что значение C_m=1 соответствует равночастотности всех возможных подслов в исходном слове. На основании оценки энтропии слов мы строим функцию $C(m)=C_m$, с областью определения: $1\le m\le n$. В соответствии с принятой в символической динамике терминологией [3] будем называть $C(m)$ функцией оценки энтропии сдвигов.

Интерес представляет изучение характера убывания значений $C(m)$ с ростом аргумента. Рассмотрим инверсную конечную разность функции $C(m)$:

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

По определению $C(m)$ значения $\Delta C(m)$ ограничены, и $0\le\Delta C(m)\le 1$. 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. Можно показать, что значения $C(m)$, равные единице могут держаться до $\hat{m}\approx\lfloor\log_k n\rfloor$. 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 слова, состоящего из псевдослучайной

последовательности символов. 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.

На основе этих рассуждений авторы и вводят меру символьного разнообразия временного ряда

$$\mu_s(V) = \frac{m^*}{\hat{m}} = \frac{\arg\max\limits_{1 \le m \le n} \Delta C(m)}{\lfloor \log_k n \rfloor},$$

где

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

Предложенная обобщенная характеристика может быть использована для последующего выявления характерных особенностей временных рядов, в частности, как одна из осей пространства кластеризации.

Библиографический список

1. 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.

2. 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.

3. Lind, Marcus. Symbolic Dynamics and Coding. Cambridge University Press. 1995. 495 p.