УДК 519.16, 51-76

ОБ ОДНОМ ХАРАКТЕРИСТИЧЕСКОМ ФУНКЦИОНАЛЕ СЛОВ НАД КОНЕЧНЫМ АЛФАВИТОМ

Ульянов М. В., д-р. техн. наук, проф.,
вед. науч. сотрудник ИПУ РАН им. В.А. Трапезникова, проф. ВМК МГУ им. М.В. Ломоносова,

Сметанин Ю. Г., д-р. физ.-мат. наук, гл. науч. сотрудник
ФИЦ ИУ РАН.

Аннотация

Исходными объектами данной статьи являются конечные слова над некоторым конечным алфавитом. Эти слова представляют собой символьные коды исследуемых объектов и процессов, которые и являются предметом последующего анализа. Предложена формализация функции энтропии слов на основе энтропии дискретных распределений. На основе анализа множества всех слов фиксированной длины над заданным алфавитом введено понятие мажоранты функции энтропии. Функция энтропии слов и мажоранта являются основой для предлагаемых в статье двух форм характеристического функционала, дающих количественную оценку близости исследуемого слова конечной длины к подслову такой же длины, случайно выбранному из случайного бесконечного слова над этим алфавитом. Предложенные формы функционала могут быть использованы при решении различных задач анализа данных, в том числе задач кластеризации и распознавания.

Ключевые слова: информационная энтропия, слова над конечным алфавитом, функция энтропии слов, мажоранта функции энтропии, характеристические функционалы.

1. Введение

В данной статье предполагается, что анализируемая информация представлена в виде символьных кодов, то есть образы исследуемых объектов или процессов кодированы словами над некоторым конечным алфавитом. Достоинством такого представления является отсутствие ложных связей и отношений между объектами, возникающего при числовом кодировании при решении задач классификации. Такие символьные представления получили широкое распространение в различных задачах, от комбинаторной теории групп и символической динамики [1] до различных прикладных задач анализа информации.

Здесь будет рассмотрен случай достаточно длинных, но конечных слов, возникающих в задачах исследования объектов и процессов, заданных символьным кодированием, в частности, задачах классификации и кластерного анализа [2]. В этих задачах возникает естественный вопрос о характеризации классов (кластеров). При этом требуется выполнение некоторых естественных требований. Первое из них — обеспечение большей близости объектов из одного класса по сравнению с объектами из других классов, то есть выбор адекватного пространства кластеризации и метрики. Второе требование связано с обеспечением независимости от конкретного выбора кода: желательно, чтобы при выборе другого способа кодирования классы

не изменялись. Использование характеристических функционалов на множестве слов является одним из способов обеспечения выполнения этих требований.

Характерным примером использования таких функционалов является энтропия пространства сдвигов в символической динамике [1]. Предположим, что заданы два различных кода, и в каждом задано пространство сдвигов, то есть некоторое подмножество конечных слов, определяемое языком (или, что то же самое, набором запрещенных подслов). Требуется ответить на вопрос, не являются ли эти пространства по сути одинаковыми в том смысле, что переводятся одно в другое простым перекодированием (кодом скользящего блока). Доказано [1], что необходимым условием существования такого перекодирования является равенство энтропий этих пространств сдвигов.

К настоящему времени предложено много различных характеристических функционалов, адаптированных к различным типам данных и задач. Так, энтропия Реньи используется в анализе случайности и неопределенности процессов, энтропия Колмогорова – Синая — в исследовании динамических систем, взаимная информация характеризует количество информации, которое можно получить об одной случайной переменной с помощью анализа другой случайной переменной, негэнтропия дает оценку расстояния распределения от нормального, квантовая относительная энтропия есть мера неразличимости двух квантовых состояний, и т. д.

В задачах классификации и кластерного анализа чрезвычайно полезными являются так же и пиковые характеристики [2]: максимальные значения некоторого характеристического функционала и значения аргументов, при которых они достигаются, дают возможность выбирать наилучшие признаки и добиваться высокой точности и робастности решений. Эти характеристики оказываются полезными также в решении важной задачи идентификации слов, близких к случайным. Такая идентификация позволяет в задачах прогнозирования с представлением в виде символьных кодов выявлять плохо прогнозируемые процессы.

Необходимо отметить, что характеристические функционалы, как правило, не дают полного ответа на поставленные вопросы. Так, в приведенном выше примере об энтропии пространства сдвигов получено только необходимое условие сводимости; достаточных условий при этом не известно [1]. Если удастся определить несколько независимых функционалов с аналогичными свойствами, ситуация станет намного яснее: неравенство значений хотя бы одного из них означает различие пространств сдвигов, а равенство всех значений делает весьма правдоподобным предположение о совпадении.

Сходная ситуация возникает во многих задачах анализа данных. Отсюда вытекает важность задачи расширения множества характеристических функционалов. При анализе функционалов указанного вида полезной оказывается также информация об элементах, на которых достигаются экстремумы функционалов. Такие элементы могут играть роль

своеобразных эталонов для сравнения с остальными элементами, выделения классов элементов и т. д.

Объектом исследования в данной статье являются конечные слова над конечным алфавитом, порожденные символьным кодированием образов исследуемых объектов или процессов.

Предмет исследования — характеристические функционалы на множестве слов, построенные на основе функции энтропии.

Постановка задачи — предложить для исследуемого слова конечной длины над конечным алфавитом характеристический функционал, отражающий близость этого слова (в некотором метрическом пространстве) к случайно выбранному (из случайного бесконечного слова) подслову такой же длины.

2. Символьное кодирование как средство унификации описаний объектов

При исследовании разнообразных объектов или процессов определенный интерес представляет подход к анализу информации, при котором образы исследуемых объектов или процессов представляются словами над некоторым конечным алфавитом. Преобразование информации в этом случае происходит с использованием символьного кодирования.

Данный подход не требует числового кодирования образов объектов, зачастую приводящего к появлению ложной информации, вызванной исключительно особенностями использованного кода. Вместо этого предлагается проводить анализ символьных строк с целью выявления их эффективных характеристик в аспекте решаемых задач, например, в целях их классификации или распознавания. Заметим, что в пользу символьного кодирования говорит и тот факт, что переход от числовых значений к символам, кодирующим числовой сегмент, приводит к меньшей чувствительности полученных слов по отношению к шумам, сопутствующим измерительной аппаратуре.

Изучением таких символьных представлений занимается раздел современной дискретной математики — комбинаторика слов [3]. Отличительной особенностью этого подхода является тот факт, что аппарат комбинаторики слов применяется к символьным кодам объектов или процессов из разнообразных областей исследований, тем самым символьное кодирование может рассматриваться как средство унификации описаний разнородных объектов и процессов, что позволяет в дальнейшем выявлять их общности или общие особенности.

Результатом символьного кодирования является слово в выбранном алфавите, которое затем изучается методами комбинаторики слов. Выбор алфавита и способа символьного кодирования, очевидно, обусловливают и качественные результаты последующего анализа. В ряде случаев сама специфика проблемной области четко обусловливает алфавит и код.

Например, в молекулярной биологии и биоинформатике при исследовании геномов очевидным является четырехсимвольный алфавит нуклеотидов ДНК — $\{A, G, T, C\}$.

В случае, когда исходными являются числовые характеристики, символьное кодирование требует определенной аккуратности. Мы продемонстрируем это на примере символьного кодирования временных рядов. Пусть текущим объектом исследования является временной ряд произвольной природы

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

где f_i — значение характеристики наблюдаемого процесса в момент t_i, n — число наблюдений (отсчетов), а предметом исследования является построение символьного кода для значений временного ряда T. При символьном кодировании временных рядов возникает требование универсальности метода кодирования, поскольку различные временные ряды имеют различную точность измерений (число значащих цифр в значениях f_i) и различный диапазон значений. Возможное решение может состоять в едином масштабировании значений наблюдаемой функции процесса и построении на этой основе строки символов, отражающей динамику числовых значений исследуемого ряда [4].

В целях такого масштабирования вводится разбиение $y_i, i = 1, \ldots, m$ диапазона размаха варьирования значений f_i, где

$$y_1 = \min_{k=1,n} f_k\ ,\ y_m = \max_{k=1,n} f_k, \quad y_{i+1} = y_i + \Delta y_i, i = \overline{1, m-1}.$$

Поскольку значения временного ряда могут попасть в точки разбиения, рассматриваются множества $[y_i, y_{i+1}) = \{y \,|\, y_i \leq y < y_{i+1}\}$, которые далее будем называть полусегментами, при этом определение числа (m) и границ полусегментов $[y_i, y_{i+1})$ является самостоятельной интересной задачей. Один из вариантов ее решения — применение бикритериального метода построения гистограмм [5], при этом число полученных полусегментов гистограммы определяет мощность алфавита, а сами полусегменты кодируются символами выбранного алфавита. Подробное изложение этого метода по отношению к символьному кодированию приведено в [6]. Выбор символов алфавита, по сути, не принципиален, иллюстративно далее будут использоваться строчные символы латинского алфавита. Далее каждому полусегменту ставится в соответствие уникальный символ алфавита, а числовое значение ряда кодируется символом полусегмента, в котором оно находится. В результате мы получаем представление временного ряда в виде строки символов. Для временного ряда, содержащего n наблюдений, мы получаем его представление в виде строки из n символов над алфавитом Σ.

Отметим еще одно преимущество подхода символьного кодирования. За редким исключением значения в отсчетах временных рядов не являются точными. Одним из таких исключений являются, например, ряды курсов валют. Для значений, имеющих погрешность

измерений, в математической статистике принято строить доверительные интервалы. Используемый бикритериальный метод построения гистограмм как раз и определяет ширину полусегмента гистограммы, а, следовательно, и «ширину» полосы значений для кодирующего этот полусегмент символа, на основе доверительной вероятности для среднего значения [5]. Таким образом, подход символьного кодирования более достоверно отражает исследуемый процесс с точки зрения математической статистики, нивелируя возможные шумы наблюдений.

3. Функция энтропии конечных слов

Реализация поставленной задачи представляется нам в виде обобщения функции энтропии конечных слов, детальную формализацию которой мы даем ниже. Отметим, что в символической динамике функции энтропии рассматривается для бесконечных слов [1], но для конечного случая подход, предусматривающий переход к пределу в бесконечности, неприемлем.

Терминология и обозначения. Далее будет использоваться терминология, и применяться обозначения, как общепринятые в комбинаторике слов и символической динамике [1, 3], так и специальные авторские обозначения, связанные с особенностью данной задачи. Введем следующие обозначения:

$\Sigma = \{s_1, s_2, \ldots, s_m\}$ — конечный алфавит, s — произвольный символ алфавита;

Σ^k — k-ая декартова степень множества Σ;

$c \in \Sigma^k$ — k-элементный кортеж, $c = \left(s^{(1)}, s^{(2)}, \ldots, s^{(k)}\right)$, $s^{(i)} \in \Sigma$;

w — слово (над алфавитом Σ) — последовательность символов алфавита;

$WD(c) = w$ — оператор порождения слова w из кортежа c путем последовательной, в порядке кортежа, конкатенации символов

$$WD\left(\left(s^{(1)}, s^{(2)}, \ldots, s^{(k)}\right)\right) = s^{(1)}s^{(2)}\ldots s^{(k)};$$

$|w| = k$ — длина слова, понимаемая как мощность порождающего кортежа;

$L(\cdot)$ — оператор порождения множества слов, действующий на множество кортежей посредством оператора $WD(\cdot)$. Пусть $C \subseteq \Sigma^*$ — множество кортежей, W — множество слов, тогда

$$L(C) = W = \{w \mid \forall c \in C\ w = WD(c)\}.$$

Будем говорить, что оператор $L(\cdot)$ порождает некоторый язык L над алфавитом Σ, например, L_k — язык всех слов длины k над алфавитом Σ,

$$L_k = L\left(\Sigma^k\right) = \{w \mid |w| = k\};$$

$SW(w, i, l)$ — оператор выделения подслова v длины l в слове w, начиная с символа в позиции i. Пусть $|w| = k$, тогда оператор определен при $i + l - 1 \le k$:

$$SW(s_1 s_2 \ldots s_k, i, l) = v = s_i s_{i+1} \ldots s_{i+l-1};$$

$SH1(w,k)$ — оператор сдвига 1, действующий на слово w окном ширины k. Определенный при $|w|>k$ оператор порождает (в общем случае) мультимножество подслов длины k, с мощностью $|w|-k+1$, выполняя сдвиг на единицу окна ширины k по слову w, начиная с крайней левой позиции:

$$SH1(w,k)=\left\{v_i \mid v_i = SW(w,i,k),\ i=\overline{1,|w|-k+1}\right\},$$

где подслова v_i могут, очевидно, порождать мультимножество

$$SH1(w,k)=\left\{v_i^{(c_i)} \mid i=\overline{1,l}\right\}, \sum_{i=1}^{l} c_i = |w|-k+1.$$

Например, для алфавита $\Sigma=\{a,b\}$ и слова $bbababa$ при окне ширины 4 имеем:

$$SH1(bbababa,4)=\{bbab, baba, abab, baba\}=\left\{bbab^{(1)}, baba^{(2)}, abab^{(1)}\right\}.$$

Энтропия дискретных распределений с конечным носителем. Рассмотрим классическую вероятностную модель с конечным носителем $M=\langle\Omega,P(\cdot)\rangle$, в которой вероятностное пространство Ω конечно — $\Omega=\{\omega_1,\omega_2,\ldots,\omega_k\}$, а вероятностная мера $P(\cdot)$ распределяет суммарную единицу вероятности между случайными событиями ω_i — $P(\omega_i)=p_i$. Тогда по определению [7] информационная энтропия такого дискретного распределения определяется в виде

$$H_P \overset{def}{=\!=} -\sum_{i=1}^{k} p_i \log p_i. \qquad (1)$$

Индекс P у H_P подчеркивает, что энтропия порождена вероятностной мерой $P(\cdot)$. Заметим, что энтропия инвариантна по отношению к элементам вероятностного пространства ω_i. По сути, H_P есть функционал, действующий в классе дискретных распределений с конечным носителем. Важно, что для этого класса распределений максимум энтропии достигается для равномерного распределения [8], т.е. когда $\forall i\ P(\omega_i)=1/|\Omega|=1/k$, характеризуя тем самым наибольшую непредсказуемость событий в такой вероятностной модели. Выбрав в качестве основания логарифма мощность вероятностного пространства, мы можем нормировать максимальное значение энтропии (1) в единицу, т.е. при $P(\omega_i)=1/|\Omega|=1/k$ имеем

$$H_P=-\sum_{i=1}^{k} p_i \log_{|\Omega|} p_i = -\sum_{i=1}^{k} \frac{1}{k}\log_k \frac{1}{k}=1.$$

Отметим, что при любом другом распределении, отличном от равномерного, значение $0\le H_P<1$, а для вырожденного распределения $\exists\omega_i: P(\omega_i)=1,\ \forall\omega_j: j\ne i\ P(\omega_j)=0$ энтропия равна нулю.

Функция энтропии слова. Предлагается выполнить построение этой функции в два этапа — на первом этапе вычисляется значение функции энтропии для слова w по подсловам

фиксированной длины k, а на втором этапе на этой основе вычисляется функция энтропии слова w для всех возможных длин подслов.

На первом этапе мы фиксируем длину подслова $k, 1 \le k \le n$. Пусть $U = L\left(\Sigma^k\right) = L_k$ — множество всех слов длины k над алфавитом Σ. Введем в рассмотрение формальное мультимножество

$$\widetilde{U} = \left\{ u_i^{(0)} \mid i = \overline{1, |\Sigma|^k} \right\},$$

элементы которого (все возможные слова длины k над алфавитом Σ) имеют нулевую кратность. Далее применим к исследуемому слову w оператор $SH1(w,k)$ и получим мультимножество $\widetilde{V}$

$$\widetilde{V} = SH1(w,k) = \left\{ v_i^{(c_i)} \mid i = \overline{1, l} \right\}.$$

Пусть m есть число позиций окна ширины k на слове w, и $|w| = n$, тогда сумма кратностей

$$\sum_{i=1}^{l} c_i = |w| - k + 1 = n - k + 1 = m.$$

Построим объединенное мультимножество $\widetilde{V} \cup \widetilde{U}$. Поскольку $\widetilde{U}$ содержит все возможные слова длины k, то $\widetilde{V} \cup \widetilde{U}$ не будет содержать новых по отношению к множеству $\widetilde{U}$ элементов. Тем самым объединение мультимножеств приведет только к изменению кратностей некоторых, или быть может всех элементов из $\widetilde{U}$. На этой основе построим вероятностную модель

$$M_k = \left\langle \Omega_k = \widetilde{V} \cup \widetilde{U},\ P_k(\cdot) \right\rangle, \tag{2}$$

где мощность $|\Omega_k| = |\Sigma|^k$, а вероятностная мера отражает частоту появления различных слов длины k в слове w, и вычисляется на основе кратности элементов мультимножества $\widetilde{V}$

$$P_k(\cdot): \begin{cases} p_i = p(\omega_i) = \dfrac{c_i}{m}, & \omega_i \in \widetilde{V}; \\ p_i = p(\omega_i) = 0, & \omega_i \in \widetilde{U} \setminus \widetilde{V}. \end{cases} \tag{3}$$

Для полученной вероятностной модели с вероятностной мерой $P_k(\cdot)$ мы определяем нормированную энтропию, используя $|\Omega_k| = |\Sigma|^k$ в качестве основания логарифма в (1)

$$H_{P_k} = -\sum_{i=1}^{|\Sigma|^k} p_i \log_{|\Sigma|^k} p_i = -\sum_{i=1}^{l} \left(\frac{c_i}{m}\right) \log_{|\Sigma|^k} \left(\frac{c_i}{m}\right). \tag{4}$$

Отметим, что значение $H_{P_k} = 0$ при $k \ne n$ означает, что все подслова, порожденные оператором $SH1(w,k)$ одинаковы и, следовательно, состоят из одного и того же символа — мы констатируем фундаментальное отсутствие разнообразия в исследуемом слове w. При $k = n$

мы получаем одно подслово, совпадающее с w, и $H_{P_n} = 0$. Просто показать, что значение $H_{P_k} = 1$ может быть получено только в случае совпадения множеств $\widetilde{V}$ и $\widetilde{U}$, при этом все слова в $\widetilde{U}$ имеют одинаковую кратность, т.е. при равночастотности и полноте всех возможных подслов длины k в исследуемом слове w.

Построение функции энтропии слова. На втором этапе мы используем естественное расширение энтропии (4) путем введения функции $H(w,k) = H_{P_k}$, аргументами которой является слово w и длина подслова k, с областью определения: $1 \le k \le n$. Значения функции $H(w,k)$ при фиксированном значении аргумента k вычисляются по формуле (4) на основе вероятностной модели (3), полученной в результате применения оператора $SH1(w,k)$ к исходному слову w. Последовательный перебор значений ширины окна k от 1 до n и дает искомые значения $H(w,k)$. Таким образом, функция $H(w,k)$ для фиксированного слова w является действительнозначной функцией ограниченного целочисленного аргумента k.

В работе [4] авторы показали, что в области значений аргумента k от $\log_{|\Sigma|} n$ до $\sqrt{n}$ имеет место соотношение $H(w,k+1) \le H(w,k)$, и на этом сегменте $H(w,k)$ является монотонно убывающей функцией. В целом на сегменте значений аргумента от 1 до n можно говорить, что $H(w,k)$ является функцией, «убывающей по совокупности».

4. Модельный пример для функции энтропии слов

В качестве модельных примеров приведем (см. табл. 1 и табл. 2) значения функции $H(w,k)$ вычисленные по формулам (3) и (4) для двух слов длины 10 в алфавите $\Sigma = \{a,b\}$, функция энтропии которых обладает различным качественным поведением. Первое из них — периодическое слово $w_1 = (ab)_5$, второе — слово $w_2 = (\,aaabbbabaa\,)$ построено с использованием предложенного ранее авторами метода реконструкции слов по подсловам фиксированной длины [9] и порождает в окне длины три при сдвиге 1 все возможные восемь подслов с единичной встречаемостью в алфавите $\Sigma = \{a,b\}$, доставляя тем самым значение $H(w_2,3) = 1$. Отметим, что для слова w_2 функция $H(w_2,k)$ не является монотонно убывающей при начальных значениях аргумента.

Соответствующие графики приведены на рисунках 1 и 2, мы показываем функции $H(w,k)$ как кусочно-линейные для наглядности отображения тенденций.

Таблица 1.

Значения функции энтропии для слова $w_1 = (ab)_5$

k	1	2	3	4	5	6	7	8	9	10
$H(k)$	1,000	0,496	0,333	0,246	0,200	0,162	0,143	0,115	0,111	0,000

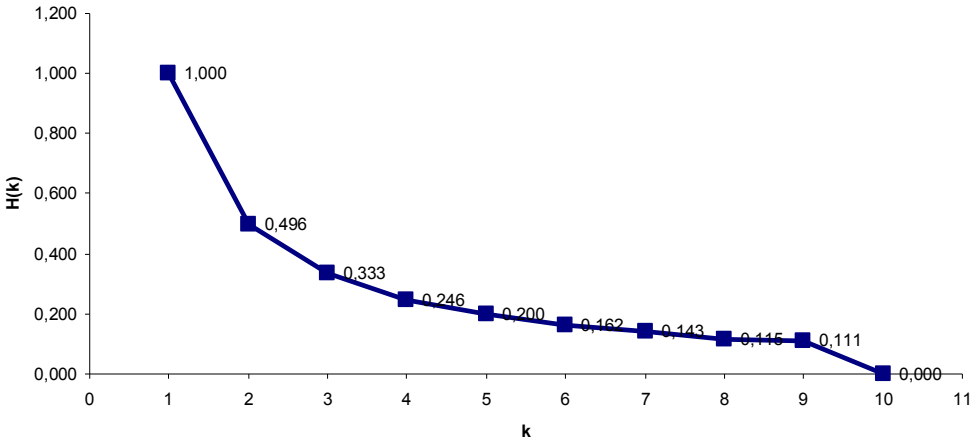

Рис. 1. График функции энтропии $H(w_1,k)$ *для модельного слова* $w_1=(ab)_5$.

Укажем на то, что $H(w_1,1)=1$, поскольку символы алфавита равночастотны в слове $w_1=(ab)_5$, а резкое падение энтропии в точке $k=2$ — $H(w_1,2)=0{,}496$ отражает потерю разнообразия слов в окне длины два, поскольку мы наблюдаем только два подслова (ab) и (ba), а два других возможных подслова длины два — (aa) и (bb) вообще не встречаются в w_1 . При этом функция $H(w_1,k)$ — монотонно убывающая на всей области определения.

Таблица 2.

Значения функции энтропии для слова $w_2=($ *aaabbbabaa* $)$

k	1	2	3	4	5	6	7	8	9	10
$H(k)$	0,971	0,987	1,000	0,702	0,517	0,387	0,286	0,198	0,111	0,000

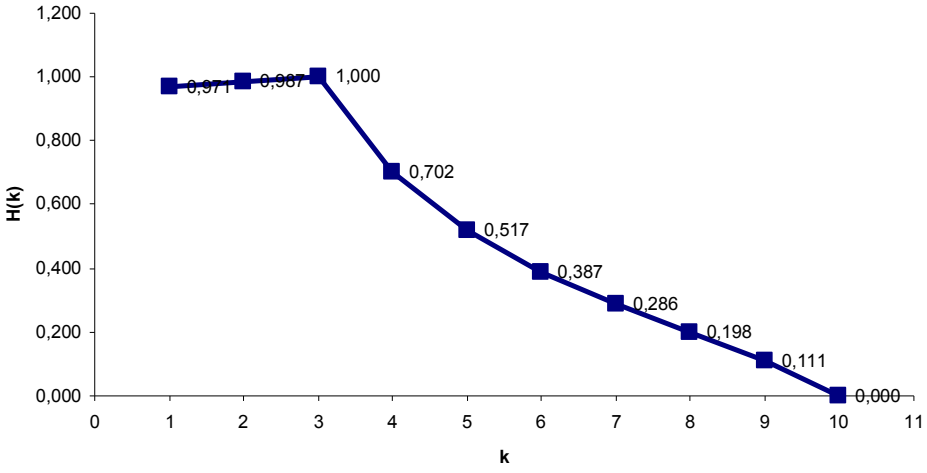

Рис. 2. График функции энтропии $H(w_2,k)$ *для модельного слова* $w_2=($ *aaabbbabaa* $)$.

Для этого слова максимум энтропии приходится на окно длины три, меньшие единицы значения для аргумента $k=1$ и $k=2$ объясняются отсутствием равночастотности символов и подслов длины два в исследуемом слове. Потеря разнообразия происходит при $k=4$ — ниже мы объясним этот факт.

5. Мажоранта функции энтропии для слов фиксированной длины

При изучении качественного поведения функции $H(w,k)$ возникает несколько вопросов, из которых, по нашему мнению, наибольший интерес представляют следующие два:

1. Как долго при начальных значениях аргумента k функция энтропии может принимать значения, близкие или равные единице? Иначе — когда происходит потеря разнообразия подслов в увеличивающемся окне сдвига один?

2. Какие максимальные значения может принимать функция $H(w,k)$ при фиксированной длине слова для различных значений аргумента k?

Порог потери разнообразия. Ответ на первый вопрос можно получить на основе сопоставления числа возможных позиций окна ширины k в слове w и числа возможных подслов длины k в данном алфавите Σ.

Для дальнейшего изложения введем следующее обозначение

$$X = \left\{ x_i^* \right\} = \underset{x \in G}{\text{Arg zero}}\, h(x),$$

где X — множество значений аргумента, при которых исследуемая функция действительного аргумента $h(x)$ обращается в ноль на области $G \subseteq \mathbf{R}^1$. Формально множество X может быть пустым множеством (например, для $h(x) = x^2 + 1, G = \mathbf{R}^1$). Обозначение допускает очевидное расширение на множество комплексных чисел.

Пусть $|w| = n$ при данном алфавите Σ. Рассмотрим функцию действительного аргумента $h(x) = n - x + 1 - |\Sigma|^x$ на сегменте $x \in [1, n]$. Поскольку на этом сегменте производная $h'(x) = -1 - |\Sigma|^x \ln|\Sigma|$ отрицательна, то функция всюду убывает на $[1, n]$. При выполнении условия $n > |\Sigma|$ значение $h(1) > 0$, следовательно, $h(x)$ имеет единственный ноль на сегменте $[1, n]$ (т.е. $|X| = 1$), который мы обозначим через

$$x^* = \underset{x \in [1,n]}{\arg\text{ zero}}\, h(x). \qquad (5)$$

Тогда ответ на первый вопрос сводится к рассмотрению следующих двух случаев:

Случай **а**: Решение уравнения $h(x) = 0$ — целочисленное, $x^* \in \mathbf{N}$. Обозначим $k^* = x^*$, тогда $h(k^*) = n - k^* + 1 - |\Sigma|^{k^*} = 0 \Rightarrow n - k^* + 1 = |\Sigma|^{k^*}$. При этом если для исследуемого слова w наблюдаемые в окне ширины k^* подслова совпадают со всеми возможными словами языка L_k, то значение $H(w, k^*) = 1$. При увеличении ширины окна на единицу $k = k^* + 1$ число возможных позиций окна уменьшается на единицу, а число всевозможных слов в алфавите Σ увеличивается в $|\Sigma|$ раз. Таким образом, мы констатируем потерю возможного разнообразия в

окне ширины k^*+1. Именно этот случай мы и наблюдаем для слова $w_2 = (\ aaabbbabaa\ )$ длины $n=10$ в алфавите мощности два в окне ширины черыре ($k^*=3$ и $10-3+1=2^3$). При этом $H(w_2,3)=1$, а $H(w_2,4)=0{,}702$. Оценка для k^* может быть получена в предположении, что длина слова значительно превышает ширину окна — т.е. при условии $n>>k$, тогда значение $k^* \approx \lfloor \log_{|\Sigma|} n \rfloor$.

Случай **b**: Уравнение $h(x)=0$ имеет решение в действительных числах $x^* \in \mathbf{R}^1 \setminus \mathbf{N}$. В этом случае последнее значение ширины окна, при котором возможно наблюдать полное разнообразие подслов — $k^* = \lfloor x^* \rfloor$. Однако это разнообразие не гарантирует, что $H(w,k^*)=1$, поскольку при этом $n-k^*+1>|\Sigma|^{k^*}$, что может быть препятствием для получения равночастотности. Например при $|\Sigma|=2$ и $n=14$, решая соответствующее трансцендентное уравнение, получаем $x^* \approx 3{,}52$ и значение $k^* = \lfloor 3{,}52 \rfloor = 3$. Тем самым мы получаем 12 позиций окна ширины три при восьми возможных подсловах, что приводит к невозможности получения равных наблюдаемых частот, и, следовательно, $H(w,3)<1$ для любого слова w, имеющего длину 14 в бинарном алфавите. Можно показать, что и в общем случае при $x^* \in \mathbf{R}^1 \setminus \mathbf{N}$ значение $H(w,k^*)<1$. С другой стороны при $k^* = \lfloor x^* \rfloor + 1$ ситуация меняется, что связано с увеличением в $|\Sigma|$ раз полного разнообразия подслов (в нашем примере 11 позиций окна при 16 возможных подсловах) и $H(w,k^*+1)<H(w,k^*)$.

Таким образом, в общем случае, последнее значение k при котором функция $H(w,k)$ может иметь значения, близкие или равные единице определяется значением $k^* = \lfloor x^* \rfloor$, где

$$x^* = \underset{x\in[1,n]}{\arg\operatorname{zero}}\, h(x),\ \text{а}\ h(x)=n-x+1-|\Sigma|^x. \qquad (6)$$

Мажоранта функции энтропии слов. Ответ на вопрос: «Какие максимальные значения может принимать функция $H(w,k)$ при фиксированной длине слова n для различных значений аргумента k?» можно получить на основе следующих рассуждений: Зафиксируем длину слова n и рассмотрим язык L_n — язык всех слов длины n над алфавитом Σ. Для конкретной ширины окна k оператора сдвига 1 (т.е. для аргумента k функции $H(w,k)$) в силу конечности множества L_n, существует, по крайней мере, одно слово, доставляющее максимум функции энтропии. Тем самым, мы переходим от функции $H(w,k)$ к функции $\widetilde{H}(n,k)$, определяемой на всем языке L_n следующим образом:

$$n = const, \forall k = \overline{1,n} \;\; \widetilde{H}(n,k) \overset{def}{=\!=} \max_{w \in L_n} H(w,k). \qquad (7)$$

Будем называть функцию $\widetilde{H}(n,k)$ — мажорантой функции энтропии слов, поскольку по определению (7) $\forall w \in L_n \; \forall k = \overline{1,n} \;\; H(w,k) \le \widetilde{H}(n,k)$. Заметим, что определение (7) не предусматривает наличие какого-то конкретного слова из L_n, обладающего мажорирующей энтропией на всей области определения, функция $\widetilde{H}(n,k)$ получает свои значения для различных аргументов k, вполне вероятно, от различных слов из L_n, что и приводит к введению обобщающего (по языку L_n) первого аргумента этой функции в виде длины слова n.

Приведем метод вычисления значений функции $\widetilde{H}(n,k)$ при фиксированной длине слова n, не связанный напрямую с определением (7). Рассмотрим два сегмента значений аргумента, определяемых порогом потери разнообразия $k^* = \lfloor x^* \rfloor$ (см. формулу (6)), — сегмент $\left[1, k^*\right]$ и сегмент $\left[k^* + 1, n\right]$. Для каждого из них определим наибольшие возможные значения энтропии следующим образом:

1. Для сегмента $\left[1, k^*\right]$ — при фиксированном k из этого сегмента мы наблюдаем $m = n - k + 1$ подслов в окне сдвига 1, имея $N = |\Sigma|^k$ возможных подслов в данном алфавите. В данном сегменте по построению выполняется неравенство $m \ge N$. Тогда каждому подслову длины k первоначально назначается частота равная $m \operatorname{div} N$, кроме того остаток $m \bmod N$ распределяется по единице между любыми (в силу свойств энтропии) $m \bmod N$ подсловами, максимизируя тем самым значение функции энтропии.

2. Для сегмента $\left[k^* + 1, n\right]$ — по построению мы фиксируем потерю разнообразия в любом окне ширины k из этого сегмента. Между значениями m и N выполняется неравенство $m < N$, и мы предполагаем, что существует слово, доставляющее единичную встречаемость всех подслов в окне ширины k оператора сдвига 1. Тогда на основе формулы (4) имеем:

$$\widetilde{H}(n,k) = -\sum_{i=1}^{m} \left(\frac{1}{m}\right) \log_{|\Sigma|^k} \left(\frac{1}{m}\right) = -\log_{|\Sigma|^k} \left(\frac{1}{m}\right) = \log_{|\Sigma|^k} (n - k + 1), \qquad (8)$$

при этом очевидно, что $\widetilde{H}(n,n) = 0$.

Приведем пример вычисления значений функции $\widetilde{H}(n,k)$ при $n = 10$ в соответствии с указанным методом для бинарного алфавита. При $n = 10$ значение $k^* = 3$. Значения $\widetilde{H}(10,k)$ в первом сегменте $[1, 3]$ вычисляются следующим образом:

— при $k=1$ есть 10 позиций окна длины один и, следовательно, возможно получить равночастотную встречаемость символов алфавита (как, например, в слове $w_1=(ab)_5$), что влечет $\widetilde{H}(10,1)=1$.

— при $k=2$ есть 9 позиций окна длины 2 при четырех возможных подсловах. Поскольку $9\ div\ 4=2$, то всем подсловам первоначально назначается частота 2, а поскольку $9 \bmod 4=1$, то одно подслово будет иметь частоту 3. Тем самым одна из вероятностей в модели будет равна 3/9, а остальные три —2/9, что доставляет значение $\widetilde{H}(10,2)\approx 0{,}987$.

— при $k=3$ есть восемь позиций окна ширины три и восемь возможных подслов, таким образом, мы можем получить равночастотность и $\widetilde{H}(10,3)=1$. Пример — слово $w_2=($ *aaabbbabaa* $)$.

Потеря разнообразия происходит в окне ширины 4, и в сегменте $[4,10]$ значения вычисляются по формуле (8). Результаты приведены в таблице 3 и на рис.3.

Таблица 3.

Значения мажоранты функции энтропии слов $\widetilde{H}(10,k)$

k	1	2	3	4	5	6	7	8	9	10
$H(k)$	1,000	0,987	1,000	0,702	0,517	0,387	0,286	0,198	0,111	0,000

Заметим, что в первом сегменте значений аргумента k (до потери разнообразия) не все значения $\widetilde{H}(10,k)$ равны единице — для значения $k=2$ это объясняется тем, что в этом случае число позиций окна не кратно числу возможных подслов данного алфавита.

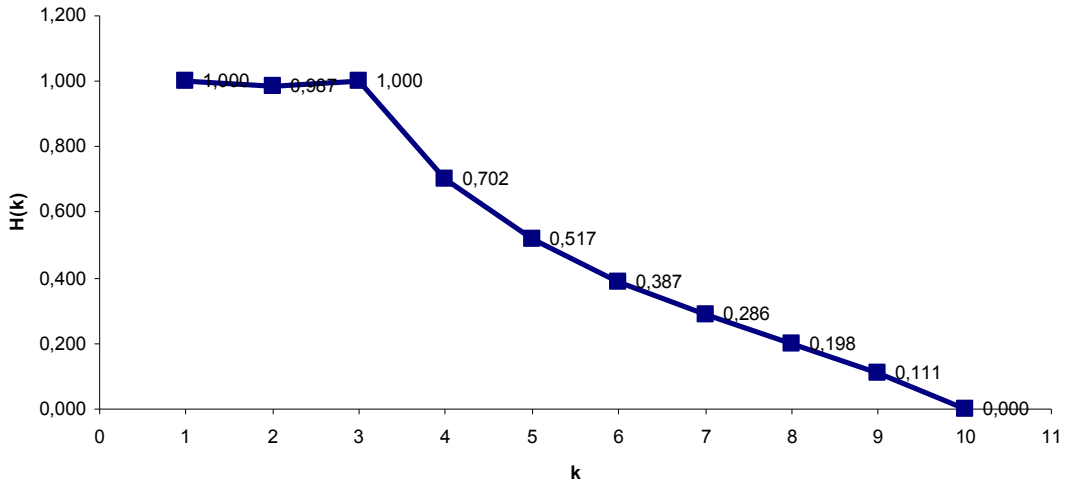

Рис. 3. График мажоранты функции энтропии слов $\widetilde{H}(10,k)$.

6. Характеристический функционал на множестве слов над конечным алфавитом

В силу определения (7) любое случайно выбранное подслово длины n из случайного (в смысле фон-Мизеса [10]) бесконечного слова не может иметь значения функции энтропии

большие, чем мажоранта $\widetilde{H}(n,k)$. Поскольку и функция энтропии исследуемого слова и мажоранта определены на сегменте $[1,n]$ мы можем использовать формализм метрического пространства с интегральным или чебышёвским расстоянием для функций, определенных на сегменте $[1,n]$.

Для решения поставленной задачи для слова $w, |w|=n$, введем в рассмотрение следующие две формы характеристического функционала, отражающие расстояние функции энтропии исследуемого слова $H(w,k)$ до функции $\widetilde{H}(n,k)$. При этом мы учитываем целочисленность аргумента обеих функций.

1. Нормированная форма, основанная на интегральном расстоянии

$$\mu_M(w)=\frac{1}{n}\sum_{k=1}^{n}\left(\widetilde{H}(n,k)-H(w,k)\right). \tag{9}$$

2. Форма, основанная на чебышёвском расстоянии

$$\mu_T(w)=\max_{k=1,n}\left(\widetilde{H}(n,k)-H(w,k)\right). \tag{10}$$

Поскольку все значения функции энтропии и мажоранты находятся в сегменте $[0,1]$, то и значения каждой из форм также нормированы в этот сегмент. По определениям (9) и (10) малые значения $\mu_M(w)$ и $\mu_T(w)$ говорят о близости исследуемого слова к «случайным» словам такой же длины. «Случайное» слово здесь понимается как любое случайно выбранное подслово длины n из случайного в смысле фон-Мизеса [10] бесконечного слова. Большие значения позволяют говорить о наличии периодичности или об отсутствии возможного разнообразия, например при отсутствии в исследуемом слове части символов используемого алфавита или их очень редкой встречаемости.

Вычисление функций $H(w,k)$ и $\widetilde{H}(n,k)$ на всем сегменте $[1,n]$, особенно функции $H(w,k)$ достаточно трудоемко, особенно для слов большой длины (геномы и т.д.). С другой стороны при больших значениях аргумента, близких к длине исследуемого слова, изучение поведения обеих функций не представляет особого интереса. Наибольшие различия проявляются в области от $k=1$ до аргумента, кратного точке потери разнообразия, т.е. до $\hat{k}\approx m\cdot k^*$. В связи с этим мы рекомендуем для длинных слов в формулах (9) и (10) изменить верхние пределы суммирования и поиска максимума (n) до порога, кратного k^*, например до $\hat{k}\approx 3\cdot k^*$. Для функционала в чебышёвской форме интерес представляет и значение аргумента, при котором достигается максимум — его сравнение с k^* также позволяет получить полезную информацию, например, если точка максимума разности лежит левее k^*, и значение $\mu_T(w)$

велико, то, скорее всего, слово w — периодическое или близкое к периодическому, например, слово Штурма [3].

Приведем примеры вычисления характеристического функционала для проанализированных выше модельных слов (данные из табл. 1, 2 и 3):

1. Для слова $w_1 = (ab)_5$: $\mu_M(w) = 0{,}238247$ и $\mu_T(w) = 0{,}49193$ при $k = 2$.

2. Для слова $w_2 = (\ aaabbbabaa\ )$: $\mu_M(w) = 0{,}002905$ и $\mu_T(w) = 0{,}02905$ при $k = 1$. В этом примере отметим, что для всех остальных значений аргумента разность $\widetilde{H}(n,k) - H(w,k) = 0$, $\mu_T(w)$ мало, и, тем самым, слово w_2 очень близко к случайному.

7. Возможные применения в задачах кластеризации, распознавания и прогнозирования

Предложенные выше формы функционала $\mu_M(w)$ и $\mu_T(w)$ могут быть применены в задачах анализа данных для выбора эталонных значений классов. В задачах кластеризации возникает проблема оценки числа кластеров и положения их центров. При недостаточно плотных выборках обучающих данных решение этих проблем приводит к значительным трудностям [11]. Очевидное требование — чтобы расстояния между элементами одного кластера были по возможности значительно меньше расстояний между разными кластерами. Для уменьшения первых желательно выбрать центр кластера, так, чтобы был обеспечен минимум максимального расстояния от этого центра до элементов этого кластера. Этот центр может не совпадать ни с одним элементом, но в большинстве методов за центр выбирается именно некоторый элемент. Выбор центра на основе предложенного характеристического функционала позволяет устранить это ограничение: как было отмечено, мажоранта функции энтропии не обязана соответствовать реальному элементу.

Аналогичная ситуация возникает при выборе эталонов классов в задачах классификации.

В задачах прогнозирования временных рядов функция мажоранты может быть полезна для оценки степени случайности анализируемых процессов. Для решения задачи прогнозирования необходимо выделять в анализируемых временных рядях некоторые закономерности. Наибольшие трудности при прогнозировании возникают при работе с последовательностями данных, в которых таких закономерностей не наблюдается. Энтропия таких последовательностей велика и близка к мажоранте.

Тот же критерий пригоден и для проверки качества генераторов псевдослучайных чисел. Здесь, наоборот, энтропия получающихся последовательностей должна быть близка к максимуму. Поскольку критерий дает количественную оценку близости к случайности, он полезен и для оценки «вырождения» генераторов при увеличении размерности пространства, заполняемого последовательно снимаемыми с данного генератора кортежами чисел. Известно, что генераторы псевдослучайных чисел с равномерным распределением, основанные на

детерминированных процедурах, которые очень хорошо работают в одномерном случае, перестают корректно работать в многомерном уже при небольших размерностях. Порог размерности, при котором еще целесообразно применять генератор, можно оценить по изменению значений предложенных форм функционала при увеличении размерности заполняемого пространства.

8. Заключение

Таким образом, в статье для заданного конечного алфавита и множества слов над этим алфавитом определены две формы характеристического функционала, дающие количественную оценку близости изучаемого слова конечной длины к подслову такой же длины, случайно выбранному из случайного бесконечного слова над этим же алфавитом. Указаны возможности использования предложенных двух форм характеристического функционала в при решении различных задач анализа данных, в том числе задач кластеризации, распознавания и прогнозирования.

9. Литература

1. *Lind D., Marcus B.* An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge, UK. 1995. — 495 pp.
2. *Сметанин Ю. Г., Ульянов М. В., Пестова А. С.* Энтропийный подход к построению меры символьного разнообразия слов и его применение к кластеризации геномов растений // Математическая биология и биоинформатика. 2016. Т. 11. № 1. С. 114–126. doi: 10.17537/2016.11.114.
3. *Lothaire M.* Algebraic Combinatorics on Words. Cambridge University Press. 2002. — 455 с. http://www-igm.univ-mlv.fr/~berstel/Lothaire/.
4. *Сметанин Ю. Г., Ульянов М. В.* Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов // Бизнес-информатика. 2014. № 3(29). С. 40–46.
5. *Петрушин В. Н., Ульянов М. В., И. А. Чертихина, Е. В. Никульчев* Бикритериальный метод построения гистограмм // Информационные технологии и вычислительные системы. 2012. № 4. С. 22–31.
6. *Ульянов М. В., Сметанин Ю. Г.* Подход к определению характеристик колмогоровской сложности временных рядов на основе символьных описаний // Бизнес Информатика. 2013. № 2. С. 49–54.
7. *Шеннон К.* Работы по теории информации и кибернетике. — М.: Изд-во иностранной литературы, 1963. — 830 с.
8. *Мартин Н., Ингленд Дж.* Математическая теория энтропии. — М.: Мир, 1988. — 350 с.
9. *Smetanin Y. G., Ulyanov M. V.* Reconstruction of a Word from a Finite Set of its Subwords under the unit Shift Hypothesis. I. Reconstruction without for Bidden Words // Cybernetics and Systems Analysis. January 2014, Volume 50, Issue 1, pp 148-156.
10. *Верещагин Н. К., Успенский В. А., Шень А.* Колмогоровская сложность и алгоритмическая случайность. М.: МЦНМО, 2013. 2010.— 576 с.
11. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.

Англ. Перевод:

ON A CHARACTERISTIC FUNCTIONAL FOR WORDS OVER A FINITE ALPHABET

The objects studied in the article are finite words over some finite alphabet. These words are symbolical codes of the investigated objects and processes that are analyzed further. A formalization of the entropy on words function is proposed that is based on the of the entropy of discrete distributions. Using the results of analysis of sets of words with fixed length over a given alphabet, a notion of entropy function majorant is proposed. The entropy on words function and the majorant form the basis for the forms of a characteristic functional proposed in the article, which give a quantitative estimation of the proximity of the tested finite word to a subword of the same length, randomly selected from a random infinite word over the same alphabet. The proposed forms of the functional may be used for solving different tasks of data analysis, including clustering and pattern recognition tasks.

Keywords: information entropy, words over a finite alphabet, entropy on words function, majorant of the entropy function, characteristic functional.

Транслитерированная литература:

1. *Lind D., Marcus B.* An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge, UK. 1995. — 495 pp.
2. *Smetanin Yu.G., Uljanov M.V., Pestova A.S. Entropiinyi podkhod k postroeniiu mery simvol'nogo raznoobraziia slov i ego primrnrniie k klasterizatsii genomov rastenii // Matematicheskaia biologiia I bioinformatika. 2016. Vol. 11. No. 1. C. 114–126.*
3. *Lothaire M.* Algebraic Combinatorics on Words. Cambridge University Press. 2002. — 455 c. http://www-igm.univ-mlv.fr/~berstel/Lothaire/.
4. Smetanin Yu.G., Uljanov M.V. Mera simvol'nogo raznoobraziia: podkhod kombinatoriki slov k opredeleniiu obobshchennykh kharakteristik vremennykh riadov // Biznes-informatika. 2014. No. 3(29). C. 40–46.
5. *Petrushin V.N., Uljanov M.V., Chertikhina I.A.,Nikul'chev E.V. Bikriterial'nyi metod postroeniia gistogramm // Informatsionnye tekhnologii i vychislitel'nye sistemy. 2012. No. 4. C. 22–*31.
6. Smetanin Yu.G., Uljanov M.V. Podkhod k opredeleniiu kharakteristik kolmogorovskoi slozhnosti vremennykh riadov na osnove simvol'nykh opisanii // Biznes-informatika. 2013. No. 2. S. 49–54.
7. *Shennon K.* Raboty po teorii informatsii i kibernetike. – M.: Izd-vo inostrannoi literatury, 1963. – 830 s.
8. *Martin N.*, Inglend Dzh. Matematicheskaia teoriia entropii. – M.: Mir, 1988. – 350 s.
9. *Smetanin Y. G., Ulyanov M. V.* Reconstruction of a Word from a Finite Set of its Subwords under the unit Shift Hypothesis. I. Reconstruction without for Bidden Words // Cybernetics and Systems Analysis. January 2014, Volume 50, Issue 1, pp. 148-156.
10. Vereshchagin N.K., Uspenskii V. A., Shen' A. Kolmogorovskaia slozhnost' I algoritmicheskaia sluchainost'. M.: MTsNMO, 2013. – 576 s.
11. Aivazian S. A., Bukhshtaber V.M., Eniukov I.S., Mashalkin L.D. Prikladnaia statisyika: Klassifikatsiia I snizhenie razmernosti. – V.: Finansy I statistika, 1989. – 607 s.