УДК: 612.087.1; 519.7; 519.66

ПРОГНОЗИРОВАНИЕ ЗНАЧЕНИЙ ЭНТРОПИИ ДЛИННЫХ КОДОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ПОРОЖДАЕМЫХ ЕСТЕСТВЕННЫМИ И ИСКУССТВЕННЫМИ ЯЗЫКАМИ

Малыгина Е.А., Иванов А.И., Язов Ю.К., Надеев Д.Н.

Расчет энтропии длинных кодировок букв осмысленного русского языка по Шеннону технически невозможен при существующих ограничениях на вычислительные ресурсы. На примере биометрических данных доказывается, что энтропия низкой размерности зависимых кодов связана с энтропией высоких размерностей экспоненциально. Как следствие, энтропия длинных кодовых последовательностей, порождаемых естественными и искусственными языками, описывается суперпозицией экспоненты и линейной составляющей. Последнее позволяет легко оценивать предельную избыточность естественных и искусственных языков.

Ключевые слова: прогнозирование, энтропия длинных зависимых кодов, независимые коды, «белый» шум.

Введение

В середине прошлого века Шеннон предложил измерять энтропию букв и их сочетаний в форме фрагментов текстов, опираясь на теорию информации. Формально информативность некоторого символа заданного алфавита можно определить следующим образом:

$$I("x_i") = \log_2(P("x_i")), \quad (1)$$

где $"x_i"$ – кодировка *i*-го символа, $P("x_i")$ – вероятность появление в тексте *i*-го символа, *i* – номер символа в алфавите. При таких обозначениях энтропия одиночных символов заданного алфавита определяется как математическое ожидание их информативности:

$$H("x_i") = E(I("x_i")). \quad (2)$$

Вычислить энтропию одиночных символов русского языка (как и любого иного языка) технически несложно. Несложно вычислить энтропию пары рядом стоящих символов, встреченных в анализируемом тексте:

$$H("x_i, x_j") = E(I("x_i, x_j")), \quad (3)$$

где

$$I("x_i, x_j") = \log_2(P("x_i, x_j")). \quad (4)$$

По индукции можно ввести понятие k-мерной энтропии $H("x_1, x_2, ..., x_k")$ для последовательности из k рядом стоящих символов, которая будет определяться через *k*-мерную информацию $I("x_1, x_2, ..., x_k")$ и через вероятности появления вариантов кодовых строк $P("x_1, x_2, ..., x_k")$.

Реальные вычислительные возможности современной техники позволяют оценить энтропию только относительно коротких кодировок до 9 знаков русскоязычного текста. Далее возникают технические проблемы, не позволяющие осуществлять прямую оценку энтропии по Шеннону. В [1] предложен метод обхода создавшихся проблем через переход в пространство расстояний Хэмминга. К сожалению, упрощения, на которых построен метод моделирования [1], вносят в вычисления существенную методическую ошибку.

Заметим, что проблема оценки энтропии длинных кодов возникает только в том случае, если коды зависимы. Если коды независимы («белый» шум), то их высокоразмерная энтропия легко вычисляется через биномиальный закон (закон испытаний Бернулли), являясь одномерной функцией вида

$$H("x_1, x_2, ..., x_k") \approx \log_2(1/32)\, k \approx 5k, \quad (5)$$

где $P("x") = P("x_1") = P("x_2") = ... = 1/32$ – вероятность появления каждого из кодов символов букв языка.

В случае «белого» шума (5) энтропия кодов будет почти совпадать с длиной кодовой последовательности. То есть для случайных паролей, набранных в кириллической 8-битной кодировке, энтропия составит величину, близкую 8*k* бит. Однако такая оценка неприемлема для понимаемых людьми текстов на русском языке. Во-первых, длинные коды из 8*k* бинарных разрядов оказываются коррелированными, во-вторых, появление каждого из *k* символов не является равновероятным. Прямые попытки учета этих факторов делают задачу вычисления энтропии как минимум *k*-мерной.

Экспоненциальная связь энтропии зависимых биокодов от их длины

Положение усложняется тем, что мы стоим на пороге появления специализированных искусственных языков с произвольной энтропией символов (слов) и произвольной длиной их кодов.

Речь идет о системах мультибиометрии [2], где каждому биометрическому образу конкретного человека ставится в соответствие его новое конфиденциальное имя или код его нового личного ключа. При этом слабый биометрический образ будет иметь низкую энтропию, а сильный биометрический образ будет иметь высокую энтропию. Заранее нельзя указать энтропию того или иного биометрического образа, кроме того, появляется возможность менять длину кода, получаемого из этого биометрического образа.

Ситуация в естественных языках такова, что каждый из нас имеет только открытое всем имя. Так было далеко не всегда, древние славяне и иные народы практиковали использование наряду с открытым именем человека (семьи) их тайное имя, известное только очень узкому кругу людей. Фактически тайное родовое имя играло роль родового пароля, знание этого имени позволяло подтвердить свою принадлежность к роду (семье).

На новом технологическом витке второе (третье, четвертое и т.д.) тайное имя становится очень длинным и играет роль личного криптографического ключа, которому может соответствовать сертификат открытого ключа, а может его не быть (открытый ключ передают кому следует, не публикуя). Видимо, язык множества длинных тайных имен человека будет поддерживаться относительно слабой технологией «нечетких экстракторов» [3] или более сильной технологией нейросетевых преобразователей биометрия-код [4].

Особенностью выходных кодов биометрии является то, что их разряды не упорядочены по их значимости. Каждый разряд кода доступа является главным, достаточно одного неверного разряда, чтобы протокол аутентификации перестал работать. Еще одной особенностью биокодов является сильная коррелированность состояний их разрядов. Биокод длиной 256 бит, соответствующий ключу аутентификации будет иметь энтропию порядка 50 бит из-за того, что разряды этого биокода коррелированы (зависимы).

Очень важной особенностью биокодов является то, что изготовитель нейросетевого преобразователя биометрия-код может менять длину выходного кода так, как посчитает необходимым. В России принято иметь длину выходного кода – 256 бит, так как отечественные стандарты по шифрованию и формированию электронной цифровой подписи ориентированы именно на такую длину криптографического ключа.

Практика создания нейросетевых преобразователей биометрия-код [5] показала, что энтропия биокода зависит как от его длины, так и от информативности биометрического образа «Свой». На рис. 1 приведены кривые роста энтропии двух биокодов, соответствующих двум биометрическим образам разной информативности.

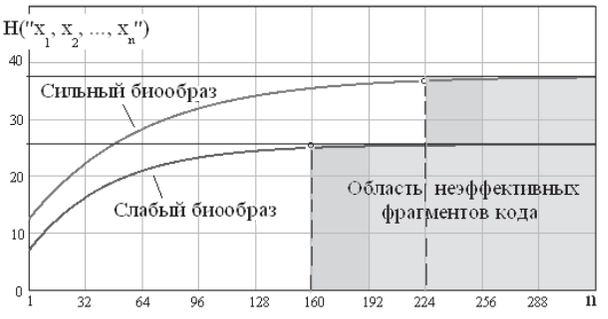

Рис. 1. Рост энтропии выходного биокода нейросети как функция от его длины

Из рис. 1 видно, что сильный биообраз позволяет получить энтропию выходного биокода порядка 38 бит при длине кода 244 бита. Дальнейшее увеличение числа выходов искусственной нейронной сети не приносит каких-либо результатов, энтропия кодов не увеличивается. Слабый биообраз дает предельное значение энтропии 26 бит уже при длине кода 160 бит, дальнейшее наращивание числа выходов нейронной сети смысла не имеет.

Установлено, что кривая роста энтропии точно описывается экспонентой:

$$H("x_1, x_2, ..., x_k") = E\left(H("x_i")\right) + A(1 - \exp\{-\tau k\}), \quad (6)$$

где $E(H("x_i"))$ – среднее значение энтропии выходных состояний каждого из выходов искусственной нейронной сети, τ – постоянная замедления роста энтропии, A-амплитуда потенциала роста энтропии.

Очевидно, что применение прямых оценок энтропии биокодов длиной 256 бит по Шеннону бесперспективно. События, которые приходится ждать при проведении численного эксперимента, не появляются в обозримые интервалы времени. Однако никто не мешает пойти иным путем. Нетрудно вычислить среднее значение энтропии одиночных решений, принимаемых на выходе каждого из нейронов. Нет технических ограничений для того, чтобы вычислить среднюю энтропию для нескольких вариантов 8-битных кодовых комбинаций – $E\left(H("x_1, x_2, ..., x_8")\right)$. Если далее найти среднюю энтропию 16-битных кодов

$E(H("x_1, x_2, ..., x_{16}"))$, то будет достаточно информации для того, чтобы найти два неизвестных параметра A и τ.

Априорная информация об экспоненциальной связи энтропии и длины кода (6) настолько значительна, что достаточно достоверно знать три точки кривой роста энтропии для прогнозирования значений зависимых биокодов любой длины.

Обобщение результатов для прогноза энтропии длинных кодов любых языков

Простая экспоненциальная связь (6) энтропии зависимых биокодов с их длиной обусловлена тем, что информативность каждого биообраза конечна. Искусственная нейронная сеть по мере роста числа ее выходов все больше и больше извлекает информации из фиксированного числа входов (из фиксированного числа биометрических параметров). При этом растет общая коррелированность выходных состояний нейронной сети. В итоге рост энтропии прекращается при некотором числе выходов преобразователя биометрия-код.

В языках (естественных и искусственных) все обстоит иначе. Информативность сообщений должна монотонно расти по мере роста длины кода (по мере роста длины текста). Язык не выкачивает информацию из одного объекта, он описывает множество объектов и их взаимосвязи. Это означает, что энтропия длинных кодов любого языка должна описываться следующим уравнением:

$$H("x_1, x_2, ..., x_k") = H("x_1") + A(1 - \exp\{-\tau k\}) + (1 - a)k, \quad (7)$$

где $H("x_1")$ – энтропия одного символа алфавита языка, a – коэффициент избыточности языка, позволяющий устранять неоднозначности.

В уравнение (7) входит четыре неизвестных. Необходимо найти четыре точки, принадлежащие кривой роста энтропии. Численный эксперимент, проведенный для русского литературного языка (использовались четыре тома классического романа «Война и мир» Л.Н. Толстого), дал значения $H("x_1") = 2{,}64$ бита; $H("x_1, x_2") = 5{,}12$ бита; $H("x_1, x_2, x_3") = 7{,}46$ бита; $H("x_1, x_2, x_3, x_4") = 9{,}94$ бита; A = 80,1; $\tau = 0{,}047$; $a = 0{,}906$. То есть избыточность русского языка достигает 90%, что позволяет эффективно править ошибки и искажения в сказанном и услышанном.

Графическая интерпретация полученных результатов

Вышеприведенные данные численного эксперимента позволяют утверждать, что энтропия всех естественных и искусственных языков изменяется по одному и тому же закону, соответствующая кривая роста энтропии фраз русского языка приведена на рис. 2.

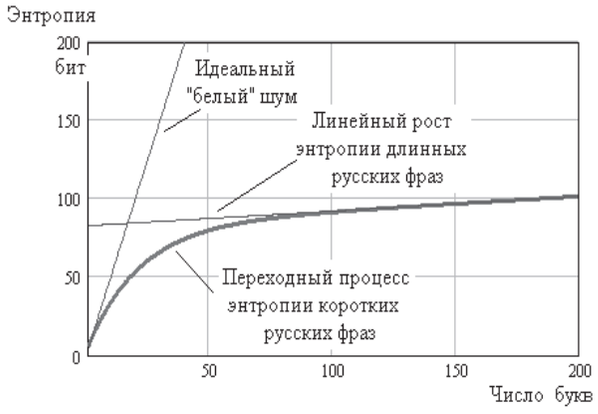

Рис. 2. Две предельные линейные функции изменения энтропии текстов на русском языке

Из данных, приведенных на рис. 2, видно, что каждый язык должен иметь линейный рост энтропии для длинных фраз. Для русского языка первоначальный быстрый экспоненциальный рост энтропии коротких фраз прекращается для текстов длиной порядка 100 букв. Далее начинается очень медленный линейный рост энтропии. Как следствие, даже очень длинные и легко запоминаемые людьми парольные фразы из 200 букв, набранные в одном регистре, будут иметь энтропию порядка 100 бит. Для того чтобы иметь энтропию парольной фразы более 100 бит, необходимо набирать на клавиатуре осмысленный текст парольных фраз в двух регистрах. При этом заглавные буквы в парольных осмысленных фразах русского языка встречаются намного реже, чем прописные буквы. Переход к текстам парольных фраз, набранных в двух регистрах, незначительно увеличивает их энтропию.

Вывод

Энтропия естественных и искусственных языков описывается суперпозицией экспоненты и линейной функции. Эта априорная информация позволяет уйти от «проклятия» размерности и отказаться от сложных вычислений, и ожидания редких событий. Появляется возможность предсказывать значение энтропии длинных зависимых кодов по значениям энтропии четырех коротких кодов. Для синтетического языка безопасного применения множества тайных биометрических образов человека (синтетического языка мультибиометрии) линейный рост энтропии длинных парольных фраз отсутствует. Объединение *k*-би-

ометрических образов приводит к появлению k последовательных переходных процессов (сглаженные экспонентой ступеньки). И в том, и в другом случае методики оценки значений энтропии длинных кодов оказываются похожими.

Литература

1. Иванов А.И., Фунтиков В.А., Майоров А.В., Надеев Д.Н. Моделирование кодовых последовательностей с энтропией естественных и искусственных биометрических языков // ИКТ. Т.8, №4, 2010. – С. 75-79.
2. Окончательная редакция проекта ГОСТ Р 52633.7-20 «Защита информации. Техника защиты информации. Высоконадежная мультибиометрическая аутентификация».
3. Dodis Y., Reyzin L., Smith A. Fuzzy Extrac-tors: How to Generate Strong Keys from Biometrics and Other Noisy // In EUROCRYPT, Data April 13, Pages 523-540, 2004.
4. Язов Ю.К., Волчихин В.И., Иванов А.И., Фунтиков В.А., Назаров И.Г. Нейросетевая защита персональных биометрических данных. М.: Радиотехника, 2012. – 157 с.
5. Волчихин В.И., Иванов А.И., Фунтиков В.А., Малыгина Е.А. Перспективы использования искусственных нейронных сетей с многоуровневыми квантователями в технологии биометрико-нейросетевой аутентификации // Известия ВУЗов. Поволжский регион. Технические науки. №4(28), 2013. – С. 88-99.

PREDICTION OF THE ENTROPY VALUES OF LONG CODE SEQUENCES GENERATED BY NATURAL AND ARTIFICIAL LANGUAGES

Malygina E.A., Ivanov A.I., Yazov Y.K., Nadeev D.N.

Calculation of entropy encoding letters long meaningful Russian by Shannon technically not possible with existing restrictions on computing resources. For example, biometric data is that the entropy of low dimension dependent code is related to the entropy of high dimensions exponentially. As a result, the entropy of long code sequences generated by natural and artifi-cial languages, described by a superposition of the exponent and the linear component. The latter allows you to easily assess the ultimate redundancy of natural and artificial languages.

Keywords: *forecasting, long dependent entropy coding, independent codes white noise.*

Малыгина Елена Александровна, аспирант Кафедры информационной безопасности систем и технологий Пензенского государственного университета. Тел. (8-841) 236-80-92. E-mail: mal890@yandex.ru

Иванов Александр Иванович, д.т.н., доцент, начальник Лаборатории биометрических и нейросетевых технологий (ЛБНТ) Пензенского научно-исследовательского электротехнического института (ПНИЭИ). Тел. (8412) 59-33-10; E-mail: ivan@pniei.penza.ru

Язов Юрий Константинович, д.т.н., профессор, начальник отдела ФАУ «Государственный научно-исследовательский, испытательный институт проблем технической защиты информации Федеральной службы по техническому и экспортному контролю» (г. Воронеж). Тел. (7432)-61-97-12; E-mail: gniii@fstec.ru

Надеев Дамир Наилевич, к.т.н., младший научный сотрудник ЛБНТ ПНИЭИ. Тел. (8412) 59-33-10; E-mail: ivan@pniei.penza.ru

УДК 621.391

ПОТЕНЦИАЛЬНЫЕ ГРАНИЦЫ ПРОПУСКНОЙ СПОСОБНОСТИ ДИСКРЕТНОГО КАНАЛА СВЯЗИ, УЧИТЫВАЮЩИЕ СТОХАСТИЧЕСКИЕ СВОЙСТВА ВХОДЯЩЕГО В ЕГО СОСТАВ НЕПРЕРЫВНОГО КАНАЛА

Батенков К.А.

Получен аналитический вид потенциальных границ пропускной способности дискретного канала связи, составленного из произвольных непрерывного канала, модулятора и демодулятора. Свойства непрерывного канала характеризуются функцией правдоподобия, а операции модуляции и демодуляции учитываются на основе стохастических свойств сигнала на выходе демодулятора.

Ключевые слова: дискретное отображение непрерывного канала связи, дискретный канал связи, непрерывный канал связи, функция правдоподобия.