Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 8, Пенза-2012, с. 49-52 (http://пниэи.рф/activity/science/BIT/T8-p49.pdf)

ОЦЕНКА ИЗБЫТОЧНОСТИ СИЛЬНО КОРРЕЛИРОВАННЫХ СЛУЧАЙНЫХ КОДОВ С ИСПОЛЬЗОВАНИЕМ ДВУХМЕРНОЙ НОМОГРАММЫ ИХ ЭНТРОПИИ

Фунтиков В.А., Иванов А.И. (г. Пенза)

Известно, что все естественные языки обладают некоторой кодовой избыточностью. Брюс Шнайер [1], ссылаясь на Томаса Кавера [2] считает, что энтропия одной буквы английского языка составляет 1,3 бита для 16-ти символьных блоков текста. Если считать, что английский алфавит имеет 26 букв, то энтропия должна составить 4,7 бита на букву ($\log_2 26 \approx 4{,}7$ бит) для случайных, бессмысленных блоков текстов произвольной длины. То есть избыточность английского языка составляет 3,4 бита на букву или 260%.

Из-за того, что коды слов и фраз со смыслом на том или ином языке обладают избыточностью, появляются корреляционные связи между парами случайно выбранных разрядов этих кодов. Если бы коды не имели избыточности (были случайными), то коэффициенты корреляции между любыми парами разрядов оказались бы нулевыми $\left|r(x_i, x_j)\right| = 0.0$ для i≠j. Для избыточных кодов ситуация меняется $\left|r(x_i, x_j)\right| \rangle 0.0$ в том числе и при i≠j. Для кодов с избыточностью следует вычислить математическое ожидание модуля корреляционных связей множества случайно выбранных пар разрядов выходного кода $r = E(\left|r(x_i, x_j)\right|)$. В этом случае высоко-размерная энтропия выходных кодов будет связана со средней входной энтропией каждого из разрядов следующим соотношением [3]:

$$H(x_1, x_2, ..., x_n) = (\beta(n, r) \cdot (n-1) + 1) \cdot E\{H(x_i)\} \qquad (1),$$

где $\beta(n, r)$ - это почти мультипликативный параметр, являющийся функцией двух переменных: числа разрядов – n и их корреляционной связанности – r. Номограмма двухмерной функции $\beta(n, r)$ представлена на рисунке 1.

Из рисунка 1 видно, что при небольших размерностях n ≤ 16 многомерная энтропия ведет себя совершенно иначе, чем высоко-размерная энтропия для n >> 16. Для пограничной ситуации функция $\beta(16, r)$ почти мультипликативной связи оказывается практически линейной.

Следствием качественных изменений функции $\beta(n, r)$ при переходе от низких размерностей к высоким размерностям является изменение отношения к оценке качества систем через контроль корреляционных связей между их параметрами. Для систем низкой размерности существенные корреляционные связи вполне допустимы, они практически не влияют на качество принимаемых нейронными сетями решений. Для систем высокой и сверхвысокой размерности все резко меняется, для них высокие корреляционные связи между контролируемыми параметрами уже недопустимы.

Похоже, что размерность – 16 является некоторой границей между низкими размерностями (менее 16) и высокими размерностями (более 16). При низкой

размерности матрицы нейросетевых функционалов очень плохо обращаются [5], однако чем выше оказывается размерность матрицы нейросетевых функционалов тем легче они обращаются (возникает положительный эффект благодати высоких и сверхвысоких размерностей).

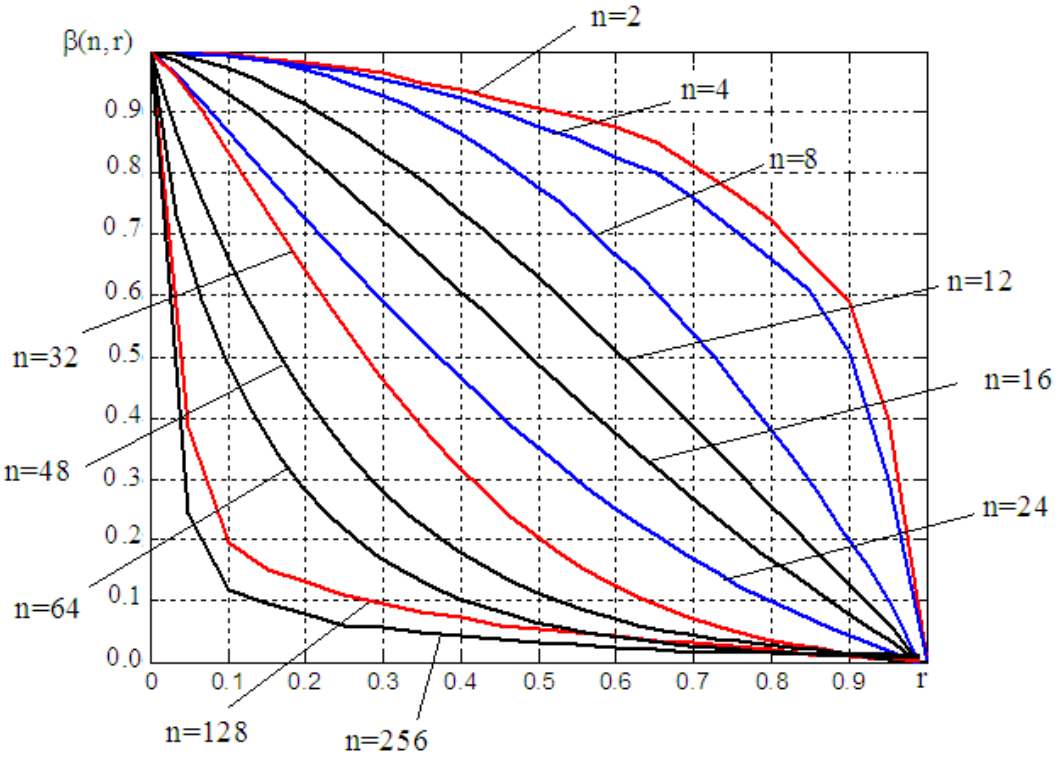

Рис. 21 Номограмма значений мультипликативной функции связи энтропии выходных кодов с коэффициентом равной коррелированности разрядов

Через вычисление многомерных функции энтропии (1) можно определить их многомерные корреляционные аналоги:

$$R(x_1, x_2,..., x_n) = (1 - \frac{H(x_1, x_2,..., x_n)}{n \cdot E\{H(x_i)\}}) \qquad (2).$$

При таком определении корреляционный функционал (2) всегда положителен и является дополнением классической функции многомерной энтропии. Максимум многомерной энтропии всегда соответствует минимуму многомерного корреляционного функционала. Связь корреляционного функционала (2) с классическими коэффициентами корреляции – r осуществляется через функцию $\beta(n,r)$.

Подчеркнем, что прямое вычисление высоко-размерной энтропии $H(x_1, x_2,..., x_n)$ и высоко-размерной корреляции $R(x_1, x_2,..., x_n)$ это очень сложные в вычислительном отношении задачи, требующие огромных массивов исходных данных. Прямое вычисление и $H(x_1, x_2,..., x_n)$, и $R(x_1, x_2,..., x_n)$ является технически не выполнимыми задачами. Вычисление же среднего значения модулей парных корреляций $r = E(|r(x_i, x_j)|)$ - это простая в вычислительном плане задача, для решения которой достаточно 1000 случайно выбранных длинных кодов с зависимыми разрядами. Рассчитывая параметр r, и пользуясь номограммой рисунка 1, мы получаем чрезвычайно быстрый алгоритм оценки высоко-размерных значений энтропии и высоко-размерных коэффициентов корреляции. Если идти этим путем, то возникает выигрыш по времени вычислений, а так же по размерам тестовых выборок на десятки и сотни порядков.

В частности, пользуясь этим методом удается рассчитать энтропию реальных биометрических длинных блоках кодов текстов трех естественных языков [4] русского, английского, татарского.

Сокращение интервалов времени, появляющееся при отказе от вычисления энтропии по Шеннону и переходе к процедурам вычисления энтропии через коэффициенты корреляции данных дан в таблице 1.

Таблица 1. Затраты времени вычисления энтропии по Шеннону и среднему модулю коэффициентов панной корреляции

Число букв (знаков) в группе	Вычисление энтропии по Шеннону		Предсказание энтропии по среднему модулю корреляции	
	Достаточный размер текстовой выборки	Время вычислений	Достаточный размер текстовой выборки	Время вычис-лений
1 знак	1 000 знаков	0.1 секунды	1 000 знаков	10 сек.
2 знака	33 000 знаков	3.3 секунды	1 000 знаков	10 сек.
3 знака	1 000 000 знаков	100 секунд	1 000 знаков	10 сек.
4 знака	33 млн. знаков	55 минут	1 000 знаков	10 сек.
5 знаков	$1*10^9$ знаков	33 часа	1 000 знаков	10 сек.
6 знаков	$33*10^9$ знаков	40 суток.	1 000 знаков	10 сек.
7 знаков	$1*10^{12}$ знаков	3.6 лет	1 000 знаков	10 сек.
8 знаков	$33*10^{12}$ знаков	120 лет	1 000 знаков	10 сек.
9 знаков	$1*10^{15}$ знаков	3800 лет	1 000 знаков	10 сек.
16 знаков	10^{22} знаков	больше возраста Земли	1 000 знаков	10 сек.
32 знака	10^{47} знаков	::::::::::::::::	1 000 знаков	10 сек.
64 знака	::::::::::::::::::::	::::::::::::::::	1 000 знаков	10 сек.

Из таблицы 1 видно, что попытки вычислять энтропию по Шеннону очень быстро приводят в тупик «проклятия размерности», когда с ростом длины кода экспоненциально растет объем необходимой тестовой выборки и время вычислений. Совершенно иначе обстоит дело при высокоразмерных оценках энтропии через усреднение модулей коэффициентов корреляции. В этом случае время вычислений вообще не зависит от длины исследуемых кодов.

Располагая высокоразмерной энтропией кодов $H(x_1, x_2,..., x_n)$ мы можем найти усредненную информативность кодов группы из n символов:

$$E\{I(x_1, x_2,..., x_n)\} = n \cdot E\{H(x_i)\} - H(x_1, x_2,..., x_n) \quad (3).$$

При необходимости от усредненной информативности группы букв (знаков), вычисленной по формуле (3), может быть осуществлен переход к информативности одной буквы алфавита, кодируемой несколькими разрядами кода. Далее может быть осуществлен переход к вычислению избыточности биометрических кодов (по аналогии с избыточности естественных языков):

$$Izb = \frac{H(x_1,x_2,....x_n)}{n \cdot E\{H(x_i)\} - H(x_1, x_2,..., x_n)} = \frac{H(x_1,x_2,....x_n)}{E\{I(x_1, x_2,..., x_n)\}} \quad (4).$$

Расчеты показывают, что избыточность биометрических кодов сильно зависит от длины анализируемого блока кода и от качества самого биометрического образа. Современные нейросетевые преобразователи биометрия-код, созданные в ОАО «Пензенский научно-исследовательский электротехнический институт», дают избыточность от 600% до 1500%, что намного хуже информативности русского, английского или татарского языка. На данный момент пока не удается добиться качества нейросетевых преобразований искусственного подсознания биометрических автоматов сравнимого с качеством нейросетевого подсознания естественного интеллекта человека.

Литература:

1. Брюс Шнайер Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си. М.: Триумф, 2002. 816 с.
2. Cover T.M. and R.C. Ring "A Convergent Gambling Estimate of the Entropy of English", IEEE Transactions on Information Theory v. IT-24, n. 4 Jul 1978, pp. 413-421.
3. Фунтиков В.А., Надеев Д.Н., Иванов А.И. Связь энтропии выходных состояний нейросетевых преобразователей биометрия-код с коэффициентами парной корреляции. «Нейрокомпьютеры: разработка, применение» № 3 2012 г.
4. Иванов А.И., Фунтиков В.А., Майоров А.В., Надеев Д.Н. Моделирование кодовых последовательностей с энтропией естественных и искусственных биометрических языков. Инфокоммуникационные технологии, том 8, № 4, 2010 г., с. 75-79
5. Язов Ю.К. «Нейросетевая защита персональных биометрических данных» Книга 1 серии «Нейросетевая биометрия» //Ю.К. Язов, В.И. Волчихин, И.Г. Назаров, В.А. Фунтиков, А.И. Иванов // М.: «Радиотехника» 2012 г., 210 с.

Материалы поступили 25.04.2012, опубликовано в Интернет 22.05.2012 по положительной рецензии д.т.н., профессора Малыгина А.Ю. (Пенза).