УДК 519.2:519.711.3:519.722

ЭНТРОПИЙНОЕ МОДЕЛИРОВАНИЕ ДИСКРЕТНЫХ СЛУЧАЙНЫХ ВЕКТОРОВ НА ПРИМЕРЕ ГРУППИРОВОК И БАЛЛЬНЫХ ПОКАЗАТЕЛЕЙ

[1,2]Тырсин А.Н.

[1]ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)», Челябинск, e-mail: at2001@yandex.ru;
[2]ФГБУН Научно-инженерный центр «Надежность и ресурс больших систем и машин» УрО РАН, Екатеринбург, e-mail: at2001@yandex.ru

Энтропийное моделирование широко используется при исследовании открытых стохастических систем в различных областях. Однако при использовании дифференциальной энтропии для моделирования стохастических систем все компоненты случайного вектора должны быть непрерывными случайными величинами. На практике исследуемые явления обычно являются непрерывными, дискретность возникает при группировании данных или при переходе к балльным показателям. В статье описана методика энтропийного моделирования многомерных стохастических систем на примере группировок и балльных показателей. Показано, что дифференциальная энтропия не может использоваться при моделировании дискретных случайных величин. Однако для случаев, когда дискретные случайные величины получаются в результате группирования данных или перехода к балльным показателям, возможно использование дифференциальной энтропии. Это достигается за счет перехода от дискретных случайных величин к их аппроксимациям непрерывными случайными величинами, имеющими кусочно-линейные функции распределения. Рассмотрены два случая. Во-первых, когда дискретность возникает при группировках исходных непрерывных величин. Во-вторых, при переходе к балльным показателям. В статье приведен пример расчета дифференциальной энтропии дискретной компоненты, полученный в результате группировки нормально распределенной случайной величины.

Ключевые слова: дифференциальная энтропия, модель, система, дискретный случайный вектор, группировка, балльный показатель

ENTROPY MODELING OF DISCRETE RANDOM VECTORS ON THE EXAMPLE OF GROUPINGS AND SCORE INDICATORS

[1,2]Tyrsin A.N.

[1]South-Ural State University (National Research University), Chelyabinsk, e-mail: at2001@yandex.ru;
[2]Scientific-Engineering Center Reliability and Life of Large Systems and Machines, Ural Branch, Russian Academy of Science, Yekaterinburg, e-mail: at2001@yandex.ru

Entropy modeling is widely used in the study of open stochastic systems in various fields. However, when using differential entropy to model stochastic systems, all components of a random vector must be continuous random variables. In practice, the phenomena under study are usually continuous, and discreteness occurs when data is grouped or when moving to scoring indicators. The article describes the technique of entropy modeling of multidimensional stochastic systems on the example of groupings and score indicators. It is shown that differential entropy cannot be used in modeling discrete random variables. However, for cases where discrete random variables are obtained as a result of grouping data or when moving to scoring indicators, it is possible to use differential entropy. This is achieved by switching from discrete random variables to their approximations by continuous random variables having piecewise linear distribution functions. Two cases are considered. First, when discreteness occurs when the source data is grouped. Secondly, when moving to the point indicators. An example of calculating the differential entropy of a discrete component obtained as a result of grouping a normally distributed random variable is given.

Keywords: differential entropy, model, system, discrete random vector, grouping, score indicator

Энтропия – это одно из фундаментальных свойств стохастических систем. В настоящее время достаточно распространено использование энтропии для описания поведения открытых стохастических систем в различных областях [1–4]. Общим в этих работах является использование введенной К. Шенноном информационной энтропии [5].

Однако применение информационной энтропии в качестве модели многомерных стохастических систем сталкивается с затруднениями: необходимо оценивать вероятности всех возможных состояний системы (это требует больших объемов выборок, кроме того, некоторые состояния заранее могут быть неизвестны), а также затруднено моделирование взаимосвязей между элементами многомерных систем.

Этих недостатков лишена модель, использующая дифференциальную энтропию [6]. Она основана на представлении системы в виде случайного вектора и разложении его дифференциальной энтропии на компоненты – энтропии хаотичности и самоорганизации. Однако все компоненты вектора должны быть непрерывными

случайными величинами. Это существенно сужает область применения энтропийного моделирования, поскольку во многих приложениях, например в медицине, экономике, часто вместо фактических значений признаков используют их сгруппированные величины или вводят их балльные (рейтинговые) оценки [7–9].

В [10] описан частный случай энтропийного моделирования, когда несколько компонент были дискретными случайными величинами. Однако выбор вида закона распределения непрерывной случайной величины, аппроксимирующей дискретную компоненту, недостаточно обоснован. Также не приведено исследование точности энтропийного моделирования при наличии балльных компонент, а также не были учтены особенности смешанного (непрерывного и дискретного) состава компонент случайного вектора.

Целью статьи является описание методики энтропийного моделирования многомерных стохастических систем, все или часть компонент которых являются балльными показателями или получены с помощью группировки, и ее апробация на модельных данных.

Материалы и методы исследования

Представим многомерную стохастическую систему в виде случайного вектора $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$. Его дифференциальная энтропия равна

$$H(\mathbf{Y}) = -\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} p_{\mathbf{Y}}(x_1, x_2, \ldots, x_m) \ln p_{\mathbf{Y}}(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots dx_m , \quad (1)$$

где $p_{\mathbf{Y}}(x_1, x_2, \ldots, x_m)$ – плотность распределения случайного вектора $\mathbf{Y}$.

Формула (1) была предложена К. Шенноном в [5] как формальный аналог понятия информационной энтропии для m-мерного непрерывного случайного вектора $\mathbf{Y}$. Эта величина впоследствии А.Н. Колмогоровым совместно с И.М. Гельфандом и А.М. Ягломом была названа дифференциальной энтропией [11].

Предлагаемый подход основан на модели многомерной стохастической системы в виде случайного вектора $\mathbf{Y}$ с взаимно зависимыми компонентами, являющимися непрерывными случайными величинами и использует дифференциальную энтропию: $S \to S(\mathbf{Y}) \to H(\mathbf{Y})$.

Каждая компонента Y_i вектора $\mathbf{Y}$ является одномерной случайной величиной, характеризующей функционирование соответствующего элемента системы.

В [6] доказано, что если все компоненты Y_i имеют дисперсии $\sigma^2_{Y_i}$, то дифференциальная энтропия $H(\mathbf{Y})$ случайного вектора $\mathbf{Y}$ равна

$$H(\mathbf{Y}) = \sum_{i=1}^{m} \ln \sigma_{Y_i} + \sum_{i=1}^{m} \kappa_i + \frac{1}{2} \sum_{k=2}^{m} \ln(1 - R^2_{Y_k / Y_1 Y_2 \ldots Y_{k-1}}) , \quad (2)$$

где $\kappa_i = H(Y_i / \sigma_{Y_i}) = H(\widehat{Y}_i) = -\int_{-\infty}^{+\infty} p_{\widehat{Y}_i}(x) \ln p_{\widehat{Y}_i}(x) dx$ – энтропийный показатель типа закона распределения случайной величины Y_i; $R^2_{Y_k / Y_1 Y_2 \ldots Y_{k-1}}$ – индексы детерминации регрессионных зависимостей. Первые два слагаемых $H(\mathbf{Y})_V = \sum_{i=1}^{m} \ln \sigma_{Y_i} + \sum_{i=1}^{m} \kappa_i$ названы энтропией хаотичности, а третье $H(\mathbf{Y})_R = \frac{1}{2} \sum_{k=2}^{m} \ln(1 - R^2_{Y_k / Y_1 Y_2 \ldots Y_{k-1}})$ – энтропией самоорганизации.

Проблема состоит в том, что все компоненты Y_i в (1) должны быть непрерывными случайными величинами, что не позволит определить энтропийные показатели типа их законов распределения. Покажем это. Рассмотрим некоторую дискретную случайную величину Z, имеющую ряд распределения, представленный в табл. 1.

Таблица 1

Ряд распределения случайной величины Z

Z	z_1	z_2	z_3	...	z_{n-1}	z_n
$p_k = P(Z = z_k)$	p_1	p_2	p_3	...	p_{n-1}	p_n

Запишем функцию распределения $F_Z(x)$ случайной величины Z

$$F_Z(x)=\begin{cases}0, & x\le z_1,\\ \sum_{i=1}^{k} p_i, & z_k<x\le z_{k+1}, k=1,...,n-1,\\ 1, & x>z_n.\end{cases}$$

очевидно, что плотность вероятности $p_Z(x)$ случайной величины Z всюду, кроме точек z_k, равна нулю, а в точках z_k не существует, т.е.

$$p_Z(x)=\begin{cases}\lim\limits_{\Delta x\to 0}\dfrac{p_k}{\Delta x}, & x=z_k, k=1,...,n,\\ 0, & x\ne z_k.\end{cases}$$

Рассмотрим дифференциальную энтропию дискретной случайной величины Z

$$H(Z)=-\int_{-\infty}^{+\infty} p_Z(x)\ln p_Z(x)dx=-\sum_{k=1}^{n}\lim_{\Delta x\to 0}(p_Z(z_k)\ln p_Z(z_k)\Delta x)=$$

$$=-\sum_{k=1}^{n}\lim_{\Delta x\to 0}\left(\frac{p_k}{\Delta x}\ln\frac{p_k}{\Delta x}\Delta x\right)=-\sum_{k=1}^{n}\lim_{\Delta x\to 0}\left(p_k\ln\frac{p_k}{\Delta x}\right).$$

Поскольку $\forall k\ \ 0<p_k<1$ и $\lim\limits_{\Delta x\to 0}\ln\dfrac{1}{\Delta x}=+\infty$, то предел в каждом слагаемом расходится и стремится к $+\infty$. Поэтому дифференциальная энтропия дискретной случайной величины Z не существует ($H(Z)\to-\infty$).

Таким образом, при использовании энтропийной модели (1)–(2) все компоненты случайного вектора **Y** должны быть непрерывными случайными величинами. Если некоторая компонента Y_i является дискретной случайной величиной, то ее необходимо заменить на непрерывную. В общем виде это делать нельзя, так как в зависимости от вида непрерывной функции распределения $F_{\tilde{Z}}(x)$, аппроксимирующей функцию $F_Z(x)$, можно получить практически любое значение энтропии $H(\tilde{Z})$, от некоторой константы до любой сколь угодно большой отрицательной величины (с ростом точности аппроксимации). Таким образом, энтропия (1) может использоваться для дискретных случайных величин только, если они получены из непрерывных путем преобразований (группировки, переход к балльным величинам и т.д.). В этом случае для определения необходимо восстановить исходную функцию распределения непрерывной случайной величины Z^0, которую заменили дискретной случайной величиной Z. Восстановить истинную функцию $F_{Z^0}(x)$ невозможно.

Поэтому ограничимся приближенным вариантом применительно к распространенным ситуациям, когда от Z^0 к Z переходят с помощью группировки данных и балльных показателей.

Результаты исследования и их обсуждение

Рассмотрим оба этих случая.

Случай 1. Группировки данных. Пусть ряд распределения дискретной случайной величины Z, представленный в табл. 2, получен путем группировки значений некоторой непрерывной случайной величины Z^0.

Таблица 2

Ряд распределения случайной величины Z

z_k	4	8	10	14	19	22
$p_k=P(Z=z_k)$	0,1	0,15	0,2	0,25	0,2	0,1

Обозначим середины всех внутренних интервалов групп как $z_{k,k+1} = (z_k + z_{k+1})/2$. Обычно при группировке данных левую границу $z_{0,1}$ первого интервала и правую границу последнего интервала определяют следующим образом [9]: $z_{0,1} = z_1 - \frac{z_2 - z_1}{2}$, $z_{6,7} = z_6 + \frac{z_6 - z_7}{2}$. В результате от ряда распределения из табл. 2 перейдем к группировке (табл. 3).

Таблица 3

Группировка для случайной величины Z

Группа	$(z_{0,1}, z_{1,2})$	$(z_{1,2}, z_{2,3})$	$(z_{2,3}, z_{3,4})$	$(z_{3,4}, z_{4,5})$	$(z_{4,5}, z_{5,6})$	$(z_{5,6}, z_{6,7})$
	(2; 6)	(6; 9)	(9; 12)	(12; 16,5)	(16,5; 20,5)	(20,5; 23,5)
p_k	0,1	0,15	0,2	0,25	0,2	0,1

Считая, что на каждом интервале некоторая непрерывная случайная величина Y распределена равномерно, с учетом заданных вероятностей p_k, достаточно просто восстановить плотность вероятности $p_Y(x)$: на каждом k-м интервале она будет постоянна и равна $p_k / (z_{k,k+1} - z_{k-1,k})$. На рисунке приведен график плотности вероятности $p_Y(x)$.

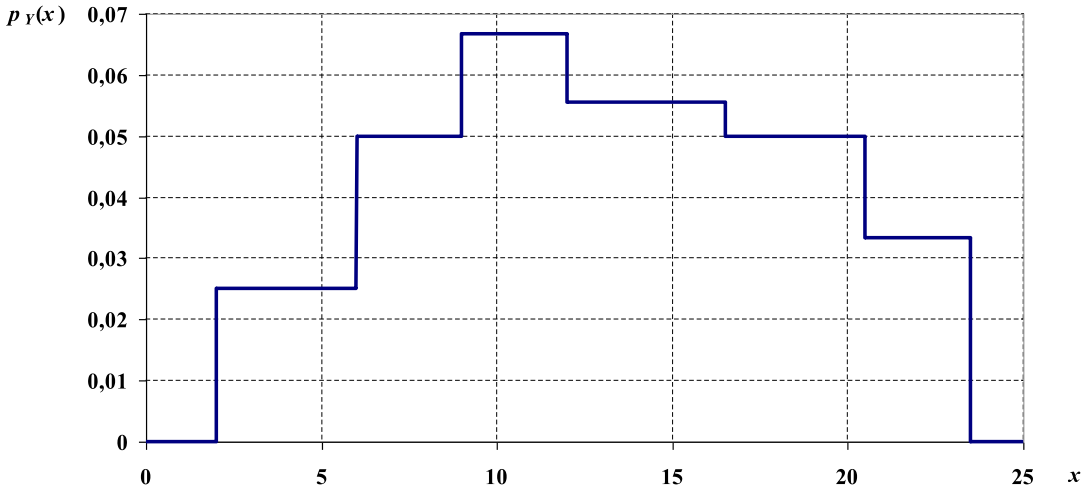

График плотности вероятности $p_Y(x)$

В общем случае для ряда распределения из табл. 1 аппроксимирующая плотность вероятности непрерывной случайной величины $p_{\tilde{Z}}(x)$ будет равна

$$p_{\tilde{Z}}(x) = \begin{cases} 0, & x \in (-\infty; z_{0,1}], \\ p_k^*, & x \in (z_{k-1,k}; z_{k,k+1}], \\ 0, & x \in (z_{n,n+1}; +\infty), \end{cases} \qquad (3)$$

где $p_k^* = \frac{p_k}{z_{k,k+1} - z_{k-1,k}}$, $z_{0,1} = z_1 - \frac{z_2 - z_1}{2}$, $z_{k,k+1} = \frac{z_k + z_{k+1}}{2}$, $z_{n,n+1} = z_n + \frac{z_n - z_{n-1}}{2}$, $k = 1, 2, \ldots, n-1$.

Теперь вычисляем оценку дифференциальной энтропии распределения (3) по формуле

$$H(\tilde{Z}) = -\sum_{k=1}^{n} p_k^* \ln p_k^* \cdot (z_{k,k+1} - z_{k-1,k}).$$

Случай 2. Балльные показатели. Пусть исследуемая непрерывная случайная величина Z^0 была в результате некоторых преобразований заменена на ряд балльных показателей (для определенности считаем баллы от 1 до M) (табл. 4).

Таблица 4

Ряд распределения балльной случайной величины Z

Z	1	2	3	...	M–1	M
$p_k = P(Z = k)$	p_1	p_2	p_3	...	p_{M-1}	p_M

Очевидно, что это частный случай рассмотренного выше случая группировок, если приравнять $z_k = k$, $k = 1, 2, ..., M$. Тогда вместо (3) получим формулу для аппроксимирующей плотности вероятности непрерывной случайной величины $p_{\tilde{Z}}(x)$:

$$p_{\tilde{Z}}(x) = \begin{cases} 0, x \in (-\infty; z_{0,1}], \\ p_k / 2, \\ 0, x \in (z_{M,M+1}; +\infty), \end{cases}$$

где $z_{k,k+1} = k + 0,5$, $k = 0, 1, ..., M$.

Пример. Сгенерируем выборку из стандартного нормального распределения Z^0 объема 100 чисел. Выборочное среднее квадратичное отклонение оказалось равным s = 0,9278. Дифференциальная энтропия равна [12]

$$H(Z^0) = \ln s\sqrt{2\pi e} = 1,344 .$$

Теперь сгруппируем данные на 7 интервалов (табл. 5). Ширина интервала каждой группы оказалась равной Δ = 0,59.

Таблица 5

Группировка для выборки из 100 наблюдений

$(z_{0,1}, z_{1,2})$	$(z_{1,2}, z_{2,3})$	$(z_{2,3}, z_{3,4})$	$(z_{3,4}, z_{4,5})$
(-2,19; -1,6)	(-16; -1,01)	(-1,01; -0,42)	(-0,42; 0,17)
$p_1 = 0,06$	$p_2 = 0,13$	$p_3 = 0,18$	$p_4 = 0,16$
$(z_{4,5}, z_{5,6})$	$(z_{5,6}, z_{6,7})$	$(z_{5,6}, z_{6,7})$	–
(0,17; 0,76)	(0,76; 1,35)	(1,35; 1,94)	–
$p_5 = 0,29$	$p_6 = 0,12$	$p_7 = 0,06$	–

Дифференциальная энтропия для распределения, задаваемого табл. 5, равна

$$H(Z) = -\int_{-\infty}^{+\infty} p_Z(x) \ln p_Z(x) dx = -\sum_{k=1}^{7} \frac{p_k}{\Delta} \ln\left(\frac{p_k}{\Delta}\right)\Delta = 1,291 .$$

Разница между величинами $H(Z)$ и $H(Z^0)$ составила менее 4 %, что говорит о достаточно точной оценке дифференциальной энтропии.

Выводы

Показано, что дифференциальная энтропия не может использоваться при моделировании дискретных случайных величин.

Для случаев, когда дискретные случайные величины получаются в результате группирования данных или перехода к балльным показателям, возможно использование дифференциальной энтропии. Это достигается за счет перехода от дискретных случайных величин к их аппроксимациям непрерывными случайными величинами, имеющими кусочно-линейные функции распределения.

Описана методика энтропийного моделирования многомерных стохастических систем, все или часть компонент которых являются балльными показателями или получены с помощью группировки.

Работа выполнена при финансовой поддержке гранта РФФИ, проект № 20-51-00001.

Список литературы

1. Малинецкий Г.Г., Потапов А.Б., Подлазов А.В. Нелинейная динамика: Подходы, результаты, надежды. 3-е изд. М.: ЛИБРОКОМ, 2011. 280 с.

2. Попков Ю.С. Математическая демоэкономика: Макросистемный подход. М.: ЛЕНАНД, 2013. 560 с.

3. Цветков О.В. Энтропийный анализ данных в физике, биологии и технике. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2015. 202 с.

4. Чумак О.В. Энтропия и фракталы в анализе данных. М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2011. 164 с.

5. Shannon C.E. A Mathematical Theory of Communication. The Bell System Technical Journal. 1948. Vol. 27. P. 379–423, 623–656.

6. Тырсин А.Н. Энтропийное моделирование многомерных стохастических систем. Воронеж: Научная книга, 2016. 156 с.

7. Сибурина Т.А. Базовая оценка и практика рейтинговых оценок в здравоохранении // Социальные аспекты здоровья населения. 2012. № 5 (27). [Электронный ресурс]. URL: http://vestnik.mednet.ru/content/view/427/30/ (дата обращения: 17.01.2021).

8. Орлов А.И. Организационно-экономическое моделирование. Ч. 2: Экспертные оценки. М.: Изд-во МГТУ им. Н.Э. Баумана, 2011. 486 с.

9. Ефимова М.Р., Ганченко О.И., Петрова Е.В. Практикум по общей теории статистики. 3-е изд., перераб. и доп. М.: Финансы и статистика, 2011. 368 с.

10. Тырсин А.Н., Шалькевич Л.В., Остроушко Д.В., Шалькевич О.В., Геворгян Г.Г. Исследование перинатального поражения центральной нервной системы у детей в неонатальном периоде методами многомерного статистического анализа // Системный анализ и управление в биомедицинских системах. 2017. Т. 16. № 3. С. 595–605.

11. Гельфанд И.М., Колмогоров А.Н., Яглом А.М. Количество информации и энтропия для непрерывных распределений // Труды III Всесоюзного математического съезда. Т. 3. М.: АН СССР, 1958. С. 300–320.

12. Тырсин А.Н., Соколова И.С. Энтропийно-вероятностное моделирование гауссовских стохастических систем // Математическое моделирование. 2012. Т. 24. № 1. С. 88–102.