ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 27 МАТЕМАТИКИ Выпуск 2
2020

А. Н. Тырсин (Екатеринбург, УрФУ). **Энтропия взаимосвязи как количественная оценка тесноты корреляционной связи между двумя случайными векторами.**

УДК 519.722+519.233.5

Резюме: Введена энтропия взаимосвязи двух случайных векторов, количественно характеризующая тесноту их корреляционной взаимосвязи. Показана связь между дифференциальной энтропией и корреляционным анализом.

Ключевые слова: корреляция, модель, случайный вектор, энтропия взаимосвязи.

В настоящее время достаточно распространено использование энтропии для описания поведения открытых стохастических систем в различных областях. Общим в этих работах является использование введенной К. Шенноном информационной энтропии [1]

$$H(S)=\sum_{i=1}^{L} p_i \ln p_i, \tag{1}$$

где $p_1,\dots,p_L$ — вероятности того, что система S находится в одном из конечного числа L соответствующих состояний $D_i\in\mathcal{D}$, $i=1,2,\dots,L$, т. е.

$$p_i=\mathbf{P}\left\{S\in D_i\right\},\qquad \bigcup_{i=1}^{L} D_i=\mathcal{D},\quad D_i\bigcap D_j=\varnothing,\quad i\neq j,\ \ i,j=1,2,\dots,L.$$

Согласно (1) модель системы представляется как функция от множества $\mathcal{D}$ ее состояний. Однако использование информационной энтропии в качестве модели такой системы имеет существенные недостатки [2].

1. Требуется оценить вероятности p_i. Это требует больших выборок, для некоторых состояний статистику получить практически невозможно.
2. Реальные системы обычно являются непрерывными.
3. Некоторые состояния систем заранее могут быть вообще не известны.
4. Затруднено моделирование взаимосвязей между элементами многомерных систем.
5. Не учитывается изменение дисперсии.
6. Адекватные модели разработаны только для частных задач.

Более адекватным подходом к описанию систем является использование модели «черного ящика», когда система определяется ее входами $\mathbf{X}=(X_1,\dots,X_n)$ и выходами $\mathbf{Y}=(Y_1,\dots,Y_m)$, т. е. $S=S(\mathbf{X},\mathbf{Y})$. Поэтому вместо информационной энтропии воспользуемся дифференциальной энтропией [1]

$$H(\mathbf{Y})=-\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} p_{\mathbf{Y}}(x_1,\dots,x_m)\,\ln p_{\mathbf{Y}}(x_1,\dots,x_m)\,dx_1\cdots dx_m,$$

где $p_{\mathbf{Y}}(x_1,\dots,x_m)$ — плотность распределения случайного вектора $\mathbf{Y}$.

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

$$H(\mathbf{Y})=\sum_{i=1}^{m}\ln\sigma_{Y_i}+\sum_{i=1}^{m}\varkappa_i+\frac{1}{2}\sum_{k=2}^{m}\ln\left(1-R^2_{Y_k\,|\,Y_1Y_2\dots Y_{k-1}}\right), \tag{2}$$

где

$$\varkappa_i = H\Big(\frac{Y_i}{\sigma_{Y_i}}\Big) = H(\widehat{Y}_i) = \int_{-\infty}^{+\infty} p_{\widehat{Y}_i}(x) \ln p_{\widehat{Y}_i}(x)\, dx$$

— 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, $k = 2, 3, \ldots, m$. Первые два слагаемых

$$H(\mathbf{Y})_V = \sum_{i=1}^{m} \ln \sigma_{Y_i} + \sum_{i=1}^{m} \varkappa_i$$

названы *энтропией хаотичности*, а третье

$$H(\mathbf{Y})_R = \frac{1}{2} \sum_{k=2}^{m} \ln \big(1 - R^2_{Y_k \,|\, Y_1 Y_2 \ldots Y_{k-1}}\big)$$

— *энтропией самоорганизации.*

Если $\mathbf{Y}^\circ$ — гауссовский случайный вектор, то

$$H(\mathbf{Y}^\circ)_V = \sum_{i=1}^{m} \ln \sigma_{Y_i} + m \ln \sqrt{2\pi e}, \quad H(\mathbf{Y}^\circ)_R = \frac{1}{2} \ln |\mathbf{R}|, \tag{3}$$

где $\mathbf{R} = \{\rho_{Y_i^\circ Y_j^\circ}\}_{m \times m}$ — корреляционная матрица.

Вид формул (2) и (3) говорит о наличии взаимосвязи между дифференциальной энтропией и корреляционным анализом.

О п р е д е л е н и е. Пусть заданы два непрерывных случайных вектора

$$\mathbf{X} = (X_1, \ldots, X_n), \quad \mathbf{Y} = (Y_1, \ldots, Y_m), \qquad n \geqslant 1, \quad m \geqslant 1.$$

Определим *энтропию взаимосвязи* между $\mathbf{X}$ и $\mathbf{Y}$ как

$$H(\mathbf{X} \cap \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{Z}) = H(\mathbf{X})_R + H(\mathbf{Y})_R - H(\mathbf{Z})_R \geqslant 0,$$

где $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$.

Нетрудно заметить, что:

1) $H(\mathbf{X} \cap \mathbf{Y}) = 0$ при взаимной независимости случайных векторов $\mathbf{X}$ и $\mathbf{Y}$;
2) чем выше теснота корреляционной взаимосвязи между $\mathbf{X}$ и $\mathbf{Y}$, тем больше значение $H(\mathbf{X} \cap \mathbf{Y})$;
3) $H(\mathbf{X} \cap \mathbf{Y}) = +\infty$ при наличии строгой функциональной взаимосвязи между хотя бы двумя компонентами у векторов $\mathbf{X}$ и $\mathbf{Y}$.

Теорема. *Пусть у всех компонент векторов* $\mathbf{X}$ *и* $\mathbf{Y}$ *существуют дисперсии. Тогда*

$$H(\mathbf{X} \cap \mathbf{Y}) = -\frac{1}{2} \ln \big(1 - d_e(\mathbf{X}, \mathbf{Y})\big),$$

где

$$d_e(\mathbf{X}, \mathbf{Y}) = 1 - \frac{1 - d_e(\mathbf{Z})}{(1 - d_e(\mathbf{X}))(1 - d_e(\mathbf{Y}))}$$

— *коэффициент тесноты корреляционной взаимозависимости между* $\mathbf{X}$ *и* $\mathbf{Y}$,

$$d_e(\mathbf{X}) = 1 - \prod_{k=2}^{n} \big(1 - R^2_{X_k \,|\, X_1 X_2 \ldots X_{k-1}}\big)$$

$d_e(\mathbf{Y})$, $d_e(\mathbf{Z})$ — *коэффициенты тесноты совместной корреляционной связи в случайных векторах* $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ [3].

Отметим, что коэффициент $d_e(\mathbf{X},\mathbf{Y})$, в отличие от метода канонических корреляций [4], позволяет однозначно оценивать тесноту взаимозависимости между случайными векторами.

Следствие 1. *Пусть U и V — непрерывные случайные величины, у которых существуют дисперсии. Тогда энтропия взаимосвязи между U и V равна*

$$H(U\cap V) = -\frac{1}{2}\,\ln\left(1-R^2_{V\,|\,U}\right) = -\frac{1}{2}\,\ln\left(1-R^2_{U\,|\,V}\right),$$

а между $\mathbf{X}$ и U —

$$H(\mathbf{X}\cap U) = 1-\frac{1-d_e(\mathbf{X}\cup U)}{1-d_e(\mathbf{X})},$$

где $R_{U\,|\,V}$ $R_{V\,|\,U}$ — теоретические корреляционные отношения между U и V.

Следствие 2. *Если $\mathbf{X}^\circ$ и $\mathbf{Y}^\circ$ — гауссовские случайные векторы, то*

$$H(\mathbf{X}^\circ\cap\mathbf{Y}^\circ) \;=\; -\frac{1}{2}\,\ln\frac{|\mathbf{R}_{\mathbf{X}^\circ\cup\mathbf{Y}^\circ}|}{|\mathbf{R}_{\mathbf{X}^\circ}|\,|\mathbf{R}_{\mathbf{Y}^\circ}|},$$

$$H(X_i^\circ\cap Y_j^\circ) \;=\; -\frac{1}{2}\,\ln\left(1-\rho^2_{X_i^\circ Y_j^\circ}\right).$$

Следствие 3. *Пусть имеется уравнение регрессии*

$$\overline{Y}_k(\mathbf{X}) = \sum_{i=1}^{n} a_{ki}\,X_i$$

и пусть $\operatorname{cov}(X_i,X_j)=0$ при любых $i\neq j$. Тогда

$$\rho^2_{Y_k X_i} \;=\; \exp\big\{-2H(Y_k\cap X_i)\big\},$$

$$a_{ki} \;=\; \exp\big\{H(Y_k)-H(X_i)-H(Y_k\cap X_i)\big\},$$

где

$$H(Y_k)=\ln\left(\sigma_{Y_k}\sqrt{2\pi\,e}\right),\quad H(X_i)=\ln\left(\sigma_{X_i}\sqrt{2\pi\,e}\right).$$

Исследование выполнено при финансовой поддержке гранта РФФИ, проект № 20-51-00001.

СПИСОК ЛИТЕРАТУРЫ

1. *Shannon C. E.* Mathematical theory of communication. — Bell Syst. Tech. J. 1948, v. XXVII, № 3, p. 379–423, № 4, p. 623–656.
2. *Тырсин А. Н.* Энтропийное моделирование многомерных стохастических систем. Воронеж: Научная книга, 2016, 156 с. // *Tyrsin A. N.* Application of Data Mining Entropy Modeling of Multidimensional Stochastic Systems. Voronezh: Nauchnaya Kniga Publ., 2016, 156 p. (In Russian.)
3. *Тырсин А. Н.* Скалярная мера взаимозависимости между случайными векторами. — Заводская лаборатория. Диагностика материалов, 2018, т. 84, № 7, с. 76–82. // *Tyrsin A. N.* Scalar measure of the interdependence between random vectors. — Industrial Laboratory. Diagnostics of Materials. 2018, v. 84. № 7, p. 76–82. (In Russian.)
4. *Сошникова Л. А., Тамашевич В. Н., Уебе Г., Шефер М.* Многомерный статистический анализ в экономике. М.: ЮНИТИ-ДАНА, 1999, 598 с. // *Soshnikova L. A., Tamashevich V. N., Uebe G., Shefer M.* Multidimensional Statistical Analysis in Economics. Moscow: UNITY-DANA, 1999, 598 p. (In Russian.)

UDC 519.722+519.233.5

Tyrsin A. N. (Yekaterinburg, Russia, Ural Federal University)**. Relationship's entropy as a quantitative assessment of the correlation's tightness between two random vectors.**

Abstract: Relationship's entropy of the between two random vectors is introduced. It quantifies the closeness of their correlation relationship. The relationship between differential entropy and correlation analysis is shown.

Keywords: relationship's entropy, correlation, model, random vector.