ЭНТРОПИЙНАЯ МЕРА ДЛЯ ИДЕНТИФИЦИКАЦИИ НЕЛИНЕЙНЫХ СТАТИСТИЧЕСКИХ СВЯЗЕЙ

Савченков Н. Н., Тюмиков Д. К.

(Россия, Самара)

В статье показана возможность идентификации нелинейных взаимосвязей с использованием энтропийной меры, раскрыта сущностная сторона предложенных показателей.

Введение. В практических задачах идентификации статистических связей по выборке $\{x_i, y_i\}$, где i = 1..N, где N — число экспериментов и $x = \{x_1, ...x_n\} \in X^n$, $y \in Y^1$, в основном исследуются оценки линейных статистических связей на основе коэффициентов корреляции (ковариации) и оценки нелинейных статистических связей на основе дисперсионных отношений (ДО) (парных дисперсионных отношений (ПДО), ДО эффектов взаимодействия (ДОЭВД) и ДО эффектов взаимосвязей (ДОЭВС))[1].

Однако имеется достаточно большой класс объектов, для которых статистические зависимости имеют неоднозначность по аргументу и по значениям функции, например, в детерминированном случае: окружность, петля гистерезиса, модели катастроф. Такие зависимости перечисленными выше мерами статистических связей не идентифицируются [2].

Модель. В статье рассматривается энтропийная мера, оценивающая неопределенность выборки $\{x_i, y_i\}$. Энтропия (1) отражает степень статистической независимости: она имеет максимальное значение при статистически независимых переменных и минимальное значение при однозначной функциональной связи между ними [3]. В приводимых далее формулах используются логарифмы по основанию 2, хотя это не принципиально и меняет только единицу измерения.

$$H(y,x) = -\sum_{x}\sum_{y} p(y,x)\cdot\log(p(y,x))\,. \qquad (1)$$

При этом максимальное значение энтропии, равное

$$H_{\max}(y,x) = \log N_{общ}\,, \qquad (2)$$

где $N_{общ}$ — количество всех возможных сочетаний значений векторов переменных, имеет место при равных вероятностях отдельных сочетаний значений переменных.

Из (1) и (2) следует первый из возможных критериев независимости векторов x и y:

$$K1_{незав} = \frac{H(y,x)}{\log N_{общ}}\,.$$

Также, имеет место следующее неравенство [1,3]:

$$H(y,x) \le H(y) + \sum_{k=1}^{n} H(x_k)\,,$$

причем знак равенства имеет место тогда и только тогда, когда переменные статистически независимы, откуда еще одним возможным критерием статистической независимости является:

$$K2_{незав} = \frac{H(y,x)}{H(y) + \sum_{k=1}^{n} H(x_k)}\,.$$

Однако наиболее интересным является критерий, для одномерного по входу и выходу объекта предложенный в [4]:

$$K_{зав} = \frac{I(y;x)}{H(y,x)}\,, \qquad (3)$$

где $I(y;x)$ — взаимная информация между векторами y и x. Однако, для многомерного по входу объекта он имеет ряд особенностей, требующих дополнительного исследования.

Поскольку в [4] не была вскрыта сущностная сторона данной меры, ниже приведены необходимые выкладки и пояснения.

Из определения взаимной информации следует, что:

$$I(y;x) = H(x) - H(y \mid x), \tag{4}$$

где $H(y \mid x)$ — условная энтропия выходной переменной, вычисленная при известных значениях входных переменных.

Также известно [1,3], что:

$$H(y,x) = H(x) + H(x \mid y), \tag{5}$$

где $H(x \mid y)$ — условная энтропия вектора входных переменных, вычисленная при известных значениях выходной переменной.

Выражая из (5) $H(x)$, и подставляя полученное выражение в (4), получаем:

$$I(y;x) = H(y,x) - \left[H(y \mid x) + H(x \mid y)\right],$$

откуда $$\frac{I(y;x)}{H(y,x)} = 1 - \frac{\left[H(y \mid x) + H(x \mid y)\right]}{H(y,x)}. \tag{6}$$

Таким образом, степень связанности векторов данных выражается через две условные и одну безусловную энтропии. Кроме того, из критерия (6) можно выделить две оценки степени неоднозначности зависимости выходной переменной от вектора входных переменных:

$$H(y \mid x) \text{ и} \tag{7}$$

$$H(x \mid y), \tag{8}$$

оценивающие неоднозначность по функции и аргументу соответственно.

Результаты (на примерах).

Поскольку, как уже было отмечено выше, использование критериев (6–8) при многомерном входе требует дополнительных исследований, в приводимых ниже примерах рассмотрим одно-

мерные по входу и выходу объекты, зависимость выхода от входа для которых представлена на рис. 1.

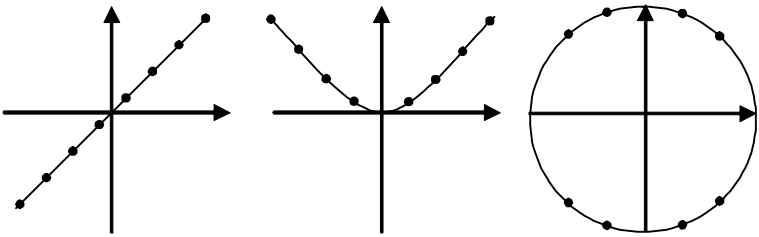

Рис 1. Виды зависимостей выхода от входа:
а) зависимость класса «прямая»;
б) зависимость класса «парабола»;
в) зависимость класса «окружность».

В приводимых ниже примерах будем использовать полученные в [1] выводы, относительно значений коэффициента корреляции и ПДО.

Пример 1. Пусть зависимость выхода от входа имеет вид y = x (рис. 1а), т.е. имеет место детерминированная линейная зависимость, при этом количество различных состояний входа ограничено и равно N_x, кроме того, при проведении активного эксперимента было выдержано равномерное распределение входной переменной с нулевым математическим ожиданием. Сравним результаты применения различных мер при данных условиях.

Т.к. вероятности отдельных значений x равны, то условные распределения для рассматриваемой зависимости имеют вид:

$$p(y \mid x) = \begin{cases} 1, & \text{при y=x,} \\ 0, & \text{при y} \neq \text{x.} \end{cases}, \quad p(x \mid y) = \begin{cases} 1, & \text{при x=y,} \\ 0, & \text{при x} \neq y. \end{cases},$$

а совместное распределение:

$$p(y,x)=\begin{cases}\dfrac{1}{N_x}, & \text{при } y=x,\\ 0, & \text{при } y\neq x.\end{cases}.$$

Рассчитаем значения мер (6–8):

$$H(y\mid x)=-\sum_x p(y\mid x)\cdot\log(p(y\mid x))=-\log 1=0\,.$$

$$H(x\mid y)=-\sum_x p(x\mid y)\cdot\log(p(x\mid y))=-\log 1=0\,.$$

$$H(y,x)=-\sum_x p(y,x)\cdot\log(p(y,x))=-\log\frac{1}{N_x}=\log N_x\,.$$

Таким образом, значение критерия (6) равно 1. Значения коэффициента корреляции и ПДО также равны 1 [1]. Меры (7, 8) имеют нулевые значения, что говорит о взаимнооднозначной зависимости между y и x.

Пример 2. Пусть зависимость выхода от входа имеет вид $y = x^2$ (рис. 1б), т.е. имеет место детерминированная квадратичная зависимость, при этом количество различных состояний входа ограничено и равно N_x, кроме того, при проведении активного эксперимента было выдержано равномерное распределение входной переменной с нулевым математическим ожиданием. Сразу отметим, что указанная зависимость является неоднозначной по аргументу. Сравним результаты применения различных мер при данных условиях.

Определим распределения вероятностей для рассматриваемой зависимости. Т.к. вероятности отдельных значений x равны между собой, то условные распределения имеют вид:

$$p(y\mid x)=\begin{cases}1, & \text{при } y=x,\\ 0, & \text{при } y\neq x.\end{cases},\quad p(x\mid y)=\begin{cases}0{,}5, & \text{при } x=\pm\sqrt{y},\\ 0, & \text{при } x\neq\pm\sqrt{y}.\end{cases},$$

а совместное распределение:

$$p(y,x)=\begin{cases}\dfrac{1}{N_x}, & \text{при } y=x,\\ 0, & \text{при } y\neq x.\end{cases}.$$

Рассчитаем значения мер (6–8):

$$H(y\mid x)=-\sum_x p(y\mid x)\cdot\log(p(y\mid x))=-\log 1=0\,.$$

$$H(x\mid y)=-\sum_x p(x\mid y)\cdot\log(p(x\mid y))=-\log 0.5=1.$$

$$H(y,x)=-\sum_x p(y,x)\cdot\log(p(y,x))=-\log\frac{1}{N_x}=\log N_x\,.$$

Таким образом, значение критерия (6) равно $1-1/\log(N_x)$. Очевидно, что при N_x стремящемся к бесконечности (т.е. в непрерывном варианте) значение критерия (6) будет стремиться к 1. Значение ПДО в данном случае равно 1, а значение коэффициента корреляции равно 0 [1].

Значение критерия (8) указывает на наличие неоднозначности зависимости по аргументу (одному значению выходной переменной соответствуют два значения входной переменной).

Пример 3. Пусть зависимость выхода от входа выражается формулой $y^2=R^2-x^2$, $R=\text{const}$ (рис. 1в), т.е. имеет место детерминированная зависимость гистерезисного типа (окружность), при этом количество различных состояний входа ограничено и равно N_x, кроме того, при проведении активного эксперимента было выдержано равномерное распределение входной переменной с нулевым математическим ожиданием. Данная зависимость является неоднозначной по аргументу и по функции ($y=\pm\sqrt{R^2-x^2}$), т.к. каждому значению входной переменной соответствуют два значения выходной переменной, а каждому зна-

чению выходной переменной соответствуют два значения входной переменной. Сравним результаты применения различных мер при данных условиях.

Определим распределения вероятностей для рассматриваемой зависимости. Т.к. вероятности отдельных значений x равны между собой, то они имеют вид:

$$p(y \mid x) = \begin{cases} 0.5, \text{ при } y = \pm\sqrt{R^2 - x^2}, \\ 0, \text{ при } y \neq \pm\sqrt{R^2 - x^2}. \end{cases},$$

$$p(x \mid y) = \begin{cases} 0{,}5, \text{ при } x = \pm\sqrt{R^2 - y^2}, \\ 0, \text{ при } x \neq \pm\sqrt{R^2 - y^2}. \end{cases},$$

$$p(y,x) = \begin{cases} \dfrac{1}{2 \cdot N_x}, \text{ при } y = \pm\sqrt{R^2 - x^2}, \\ 0, \text{ при } y \neq \pm\sqrt{R^2 - x^2}. \end{cases}.$$

Рассчитаем значения мер (6–8):

$$H(y \mid x) = -\sum_x p(y \mid x) \cdot \log(p(y \mid x)) = -\log 0.5 = 1.$$

$$H(x \mid y) = -\sum_x p(x \mid y) \cdot \log(p(x \mid y)) = -\log 0.5 = 1.$$

$$H(y,x) = -\sum_x p(y,x) \cdot \log(p(y,x)) = \log(2 \cdot N_x).$$

Таким образом, значение критерия (6) равно $1 - 2/\log(2 \cdot N_x)$. Очевидно, что при N_x стремящемся к бесконечности (т.е. в непрерывном варианте) значение критерия (6) будет стремиться к 1. Значения коэффициента корреляции и ПДО в данном случае равны 0 [1].

Значения критериев (7, 8) указывают на наличие неоднозначности зависимости как по аргументу, так и по функции.

Заключение. На примерах показано, что дисперсионные критерии позволяют определить наличие взаимосвязи при ее неоднозначности по аргументу, однако неоднозначные по функции зависимости способна распознать только энтропийная мера степени связи между переменными. Также показана принципиальная возможность дискриминации более широко набора классов зависимостей, чем при использовании коэффициента корреляции или парных дисперсионных отношений.

СПИСОК ЛИТЕРАТУРЫ

1. Дисперсионная идентификация / Под ред. Н.С. Райбмана. – М.: Наука, 1981.
2. Методы структурной идентификации химико-технологических процессов: Учеб. пособ. / Д.К. Тюмиков; Куйбыш. политехн. ин-т. Куйбышев, 1990. 74 с.
3. Галлагер Р. Теория информации и надежная связь. Перев. с англ. под ред. М.С. Пинскера и Б.С. Цыбакова. М., «Сов. радио», 1974.
4. Златкис Ю.А. и др. Некоторые алгоритмы обработки статистической информации, вычисление степени статистической зависимости случайных величин и обработка группированных данных // Применение вычислительных методов и средств в физике. — 1978. с.61-88.

ENTROPY MEASURE FOR IDENTIFICATION OF NONLINEAR STATISTICAL RELATIONS

Savchenkov N. N., Tyumikov D. K.

(Russia, Samara)

In article the opportunity of identification of nonlinear statistical relations with use of entropy measures is shown, the intrinsic party of the offered parameters is opened.