УДК 681.51.015

ИНФОРМАЦИОННЫЕ МЕРЫ СТАТИСТИЧЕСКОЙ СВЯЗИ ДЛЯ ИДЕНТИФИКАЦИИ МНОГОМЕРНЫХ ПО ВХОДУ ОБЪЕКТОВ

© 2007 Н.Н. Савченков, Д.К. Тюмиков

Самарская государственная академия путей сообщения

В статье приведено доказательство возможности представления информации статистической связи нескольких переменных в виде взвешенной суммы взаимных и взаимных условных информаций, вычисленных на различных срезах многомерной плотности распределения. На примерах показана возможность использования информационных мер для идентификации статистических связей, показано принципиальное отличие возможностей предлагаемого подхода от существующих.

Важным моментом при анализе многомерных статистических взаимосвязей является определение как степени связанности всех переменных, так и степени взаимосвязи отдельных групп переменных. Для этого уже разработан ряд подходов, например, на основе парных корреляций или дисперсионных отношений [1, 2]. Однако существует ряд видов статистических взаимосвязей, идентификация которых вышеуказанными методами невозможна.

В связи с этим актуальной является разработка нового подхода к определению степени связанности переменных. В ряде работ, например [1, 3], говорится, что взаимную информацию можно использовать в качестве меры связи, а в [1] также отмечается, что информационная мера должна дать более точные оценки при идентификации нелинейных объектов, однако дальнейшего развития данная идея не получила.

В статье предлагается подход, основанный на информационных мерах. Рассмотрим статический нелинейный объект с n входными переменными $x = \{x_1,...,x_n\} \in X^n \subset R^n$ и одной выходной переменной $y \in Y \subset R^1$, $y = f(x)$. Как известно /3,4/, совместная информация статистической связи нескольких переменных может быть представлена в виде суммы взаимных и взаимных условных информаций следующим образом:

$$I_{y(x_1,x_2,...x_n)} = I_{yx_1} + I_{yx_2|x_1} + \\ + I_{yx_3|x_1x_2} + ... + I_{yx_n|x_1x_2...x_{n-1}}, \qquad (1)$$

где n = dim x – количество входных переменных;

I_{yx_1} – взаимная информация связи выходной переменной y с входной x_1;

$I_{yx_2|x_1}$ – взаимная информация связи выходной переменной y с входной x_2, при условии, что известно значение входной переменной x_1;

$I_{yx_n|x_1x_2...x_{n-1}}$ – взаимная информация связи выходной переменной y с входной x_n, при условии, что известны значения входных переменных $x_1 \ldots x_{n-1}$.

Докажем, что информацию статистической связи n переменных можно представить в виде взвешенной суммы взаимных и взаимных условных информаций по всем возможным сочетаниям входных переменных.

Теорема. Пусть справедливо соотношение (1), тогда имеет место следующее разложение:

$$I_{y(x_1,x_2,...x_n)} = \frac{1}{C_n^1 \cdot C_1^0} \cdot \sum_{i_1} I_{yx_{i_1}} + \frac{1}{C_n^2 \cdot C_2^1} \times$$

$$\times \sum_{i_1} \sum_{i_2 (i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}} + \frac{1}{C_n^3 \cdot C_3^2} \times$$

$$\times \sum_{i_1} \sum_{i_2 (i_2 \neq i_1)} \sum_{i_3 [(i_3 \neq i_2),(i_3 \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}} + ... + \frac{1}{C_n^n \cdot C_n^{n-1}} \times$$

$$\times \sum_{i_1} \sum_{i_2 (i_2 \neq i_1)} ... \sum_{i_n [(i_n \neq i_{n-1}),(i_n \neq i_{n-2}),...,(i_n \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}...x_{i_n}}, \qquad (2)$$

где n – количество входных переменных;

C_n^k – количество сочетаний по k элементов из n;

$i_1, i_2, \ldots, i_n$ – индексы входных переменных, пробегающие значения в диапазоне 1…n;

$\sum_{i_k [(i_k \neq i_{k-1}),(i_k \neq i_{k-2}),...,(i_k \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}...x_{i_k}}$ — сумма по всем i_k, таким, что i_k не равно i_{k-1}, i_k не равно i_{k-2}, …, i_k не равно i_1.

Доказательство приведено в приложении.

Каждый член разложения (2) представляет собой взаимную информацию, вычисленную на различных срезах многомерной плотности распределения. Поскольку в других методах идентификации плотность распределения используется только в качестве весовой функции, а сами моменты вычисляются относительно значений переменных, то предлагаемый метод должен иметь значительные отличия. Как будет видно ниже, основное отличие заключается в способности к определению степени связи между переменными в случае функционально неоднозначных взаимозависимостей.

Для иллюстрации использования информационных мер проведем их сравнение с известными мерами, а именно, с корреляционными и дисперсионными отношениями.

Пример 1. Рассмотрим линейную зависимость выходной переменной y от 3 независимых входных переменных $y = x_1 + x_2 + x_3$. Диапазон изменения входных переменных ограничим отрезком $[-5;4]$ с равномерным распределением значений по диапазону.

В этом случае $I_{yx_1} = I_{yx_2} = I_{yx_3} = 0.311$, $I_{yx_1|x_2} = I_{yx_1|x_3} = I_{yx_2|x_1} =$

$= I_{yx_2|x_3} = I_{yx_3|x_1} = I_{yx_3|x_2} = 0.632$,

$I_{yx_1|x_2x_3} = I_{yx_2|x_1x_3} = I_{yx_3|x_1x_2} = 2.49$

, $I_{y(x_1,x_2,\dots x_n)} = 3.433$, $\frac{1}{C_3^1 \cdot C_1^0} \cdot \sum_{i_1} I_{yx_{i_1}} = 0.311$,

$\frac{1}{C_3^2 \cdot C_2^1} \cdot \sum_{i_1} \sum_{i_2(i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}} = 0.632$,

$\frac{1}{C_3^3 \cdot C_3^2} \cdot \sum_{i_1} \sum_{i_2(i_2 \neq i_1)} \sum_{i_3[(i_3 \neq i_2),(i_3 \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}} = 2.49$. Следует обратить внимание на тот факт, что все отношения вида

$\frac{I_{yx_{i_1}}}{\sum_{i_1} I_{yx_{i_1}}} = 0.333$, $\frac{\sum_{i_2(i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}}}{\sum_{i_1} \sum_{i_2(i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}}} = 0.333$

и

$\frac{I_{yx_{i_1}|x_{i_2}x_{i_3}}}{\sum_{i_1} \sum_{i_2(i_2 \neq i_1)} \sum_{i_3[(i_3 \neq i_2),(i_3 \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}}} = 0.333$,

т.е. определяют вклад каждой входной переменной в выходную. Для сравнения рассчитаны коэффициенты корреляции, при этом отношение коэффициента корреляции каждой входной переменной к сумме этих коэффициентов также равно 0,333. Такое точное совпадение информационных мер с корреляционными будет наблюдаться только при одинаковом вкладе каждой переменной в выходную, однако оно показывает возможность применения приведенного разложения при анализе взаимозависимостей в статистических данных.

Пример 2. При анализе нелинейных связей сравнение предлагаемого подхода необходимо проведем с дисперсионным анализом, т.к. корреляционный подход не позволяет получить надежных оценок в нелинейном случае. Рассмотрим зависимость выходной переменной y от 3 независимых входных переменных $y = x_1 + x_2^{\,2} + x_3$. Диапазон изменения входных переменных ограничим отрезком $[-5;4]$.

Рассматриваемая зависимость является аддитивной относительно функций от входных переменных, поэтому множественное дисперсионное отношение можно представить в виде суммы парных дисперсионных отношений [2]:

$$1 = \eta_{y,x_1} + \eta_{y,x_2} + \eta_{y,x_3}.$$

Расчет показывает, что $\eta_{y,x_2} = 0.787$, $\eta_{y,x_1} = \eta_{y,x_3} = 0.106$, что показывает равный вклад первой и третьей переменной в дисперсию выходной, и значительное превышение вклада второй переменной по сравнению с другими.

Соотношение информационных мер также указывает на больший вклад второй входной переменной в выходную:

$I_{yx_1} = I_{yx_3} = 0.185$, $I_{yx_2} = 1.05$;

$I_{yx_1|x_2} = I_{yx_3|x_2} = 0.619$, $I_{yx_1|x_3} = I_{yx_3|x_1} = 1.464$

$I_{yx_2|x_1} = I_{yx_2|x_3} = 1.484$; $I_{yx_1|x_2x_3} = I_{yx_3|x_1x_2} = 2.35$,

$I_{yx_2|x_1x_3} = 2.37$ $I_{y(x_1,x_2,\dots x_n)} = 4.019$,

$\frac{1}{C_3^1 \cdot C_1^0} \cdot \sum_{i_1} I_{yx_{i_1}} = 0.473$,

$\frac{1}{C_3^2 \cdot C_2^1} \cdot \sum_{i_1} \sum_{i_2(i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}} = 1.189$,

$\frac{1}{C_3^3 \cdot C_3^2} \cdot \sum_{i_1} \sum_{i_2(i_2 \neq i_1)} \sum_{i_3[(i_3 \neq i_2),(i_3 \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}} = 2.357$.

Можно заметить, что отношения вида $\frac{I_{yx_{i_1}}}{\sum_{i_1} I_{yx_{i_1}}}$ близки к соответствующим парным дисперсионным отношениям (0.13/0.106, 0.739/0.787, 0.13/0.106), что также подтверждает возможность применения вышеприведенного разложения для анализа статистических взаимосвязей.

Пример 3. В завершении, рассмотрим зависимость выходной переменной y от 3 независимых входных переменных $y^2 = x_1^{\,2} + x_2^{\,2} + x_3^{\,2}$, являющуюся функционально неоднозначной ($y = \pm\sqrt{x_1^{\,2} + x_2^{\,2} + x_3^{\,2}}$). Диапазон изменения входных переменных ограничим отрезком $[-5;4]$.

Расчет как парных корреляций, так и парных дисперсионных отношений, дает нулевую связь выходной переменной с любой из входных, однако выходная переменная зависит от входных, точнее каждому набору значений входных переменных соответствуют два значения выходной переменной. Расчет информационных мер дает результат, говорящий о наличии статистической взаимосвязи:

$$I_{yx_1} = I_{yx_2} = I_{yx_3} = 0.382,$$

$$I_{yx_1|x_2} = I_{yx_1|x_3} = I_{yx_2|x_1} =$$

$$= I_{yx_2|x_3} = I_{yx_3|x_1} = I_{yx_3|x_2} = 0.831,$$

$$I_{yx_1|x_2x_3} = I_{yx_2|x_1x_3} = I_{yx_3|x_1x_2} = 2.163,$$

$$I_{y(x_1,x_2,\dots x_n)} = 3.376,$$

$$\frac{1}{C_3^1 \cdot C_1^0} \cdot \sum_{i_1} I_{yx_{i_1}} = 0.382,$$

$$\frac{1}{C_3^2 \cdot C_2^1} \cdot \sum_{i_1} \sum_{i_2(i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}} = 0.831,$$

$$\frac{1}{C_3^3 \cdot C_3^2} \cdot \sum_{i_1} \sum_{i_2(i_2 \neq i_1)} \sum_{i_3[(i_3 \neq i_2),(i_3 \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}} = 2.163.$$

В данном случае все переменные одинаково влияют на выходную, что следует из симметричности формулы относительно входных переменных и того условия, что они независимы и пробегают один диапазон значений с равномерным распределением, при этом все отношения вида

$$\frac{I_{yx_{i_1}}}{\sum_{i_1} I_{yx_{i_1}}} = 0.333, \quad \frac{\sum_{i_2(i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}}}{\sum_{i_1} \sum_{i_2(i_2 \neq i_1)} I_{yx_{i_1}|x_{i_2}}} = 0.333$$

и $\frac{I_{yx_{i_1}|x_{i_2}x_{i_3}}}{\sum_{i_1} \sum_{i_2(i_2 \neq i_1)} \sum_{i_3[(i_3 \neq i_2),(i_3 \neq i_1)]} I_{yx_{i_1}|x_{i_2}x_{i_3}}} = 0.333$,

т.е. показывают вклад каждой входной переменной в выходную.

ПРИЛОЖЕНИЕ

Доказательство теоремы. Для двух входных переменных формула (1) имеет вид:

$$I_{y(x_1,x_2)} = I_{yx_1} + I_{yx_2|x_1}. \qquad (3)$$

Очевидно, что порядок символов x_1 и x_2 в данном разложении несуществен, поэтому можно также записать:

$$I_{y(x_1,x_2)} = I_{yx_2} + I_{yx_1|x_2}. \qquad (4)$$

Попарно суммируя левые и правые части (3) и (4), получаем:

$$2 \cdot I_{y(x_1,x_2)} = I_{yx_1} + I_{yx_2|x_1} + I_{yx_1} + I_{yx_2|x_1}$$

После деления обеих частей на 2 и группировки подобных членов, получаем:

$$I_{y(x_1,x_2)} = \frac{1}{2} \cdot (I_{yx_1} + I_{yx_1}) + \frac{1}{2} \cdot (I_{yx_2|x_1} + I_{yx_2|x_1})$$

Аналогичные выкладки можно провести для любого числа переменных, например, при n=3, получаем:

$$I_{y(x_1,x_2,x_3)} = \frac{1}{3} \cdot (I_{yx_1} + I_{yx_2} + I_{yx_3}) + \frac{1}{6} \cdot (I_{yx_1|x_2} +$$
$$+ I_{yx_1|x_3} + I_{yx_2|x_1} + I_{yx_2|x_3} + I_{yx_3|x_1} + I_{yx_3|x_2}) +$$
$$+ \frac{1}{3} \cdot (I_{yx_1|x_2x_3} + I_{yx_2|x_1x_3} + I_{yx_3|x_1x_2});$$

а при n=4, получаем:

$$I_{y(x_1,x_2,x_3,x_4)} = \frac{1}{4} \cdot (I_{yx_1} + I_{yx_2} + I_{yx_3} + I_{yx_4}) + \frac{1}{12} \times$$
$$\times (I_{yx_1|x_2} + I_{yx_1|x_3} + I_{yx_1|x_4} + I_{yx_2|x_1} + I_{yx_2|x_3} + I_{yx_2|x_4} +$$
$$+ I_{yx_3|x_1} + I_{yx_3|x_2} + I_{yx_3|x_4} + I_{yx_4|x_1} + I_{yx_4|x_2} + I_{yx_4|x_3}) +$$
$$+ \frac{1}{12} \cdot (I_{yx_1|x_2x_3} + I_{yx_1|x_2x_4} + I_{yx_1|x_3x_4} + I_{yx_2|x_1x_3} +$$
$$+ I_{yx_2|x_1x_4} + I_{yx_2|x_3x_4} + I_{yx_3|x_1x_2} + I_{yx_3|x_1x_4} + I_{yx_3|x_2x_4} +$$
$$+ I_{yx_4|x_1x_2} + I_{yx_4|x_1x_3} + I_{yx_4|x_2x_3}) + \frac{1}{4} \cdot (I_{yx_1|x_2x_3x_4} +$$
$$+ I_{yx_2|x_1x_3x_4} + I_{yx_3|x_1x_2x_4} + I_{yx_4|x_1x_2x_3}).$$

Замечая, что

$$\frac{1}{2}=\frac{1}{C_2^1\cdot C_1^0}=\frac{1}{C_2^2\cdot C_2^1},$$
$$\frac{1}{3}=\frac{1}{C_3^1\cdot C_1^0}=\frac{1}{C_3^3\cdot C_3^2}$$

и т.д., и используя метод математической индукции, получаем разложение (2). Теорема доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Райбман Н.С., Чадеев В.М.* Построение моделей процессов производства. М.: "Энергия", 1975.
2. Методы структурной идентификации химико-технологических процессов: Учеб. пособ. / *Д.К. Тюмиков*. Куйбыш. политехн. ин-т. Куйбышев, 1990.
3. *Стратонович Р.Л.* Теория информации. М.:, Сов. радио, 1975.
4. *Фано Р.* Передача информации. Статистическая теория связи: Пер. с англ. / *Пер. И.А. Овсеевича, М.С. Пинскера;* Под ред. *Р.Л. Добрушина.* М.: Мир. 1965.

INFORMATION MEASURES OF STATISTICAL RELATION FOR IDENTIFICATION OF MULTIVARIATE ON INPUT OBJECTS

© 2007 N.N. Savchenkov, D.K. Tyumikov

Samara State Railway Academy

In clause the proof of an opportunity of representation of the information of statistical relation of several variables in the form of the weighed sum of the mutual and mutual conditional information calculated on various cuts of multivariate density of distribution is resulted. On examples the opportunity of use of information measures for identification of statistical relations is shown, basic difference of opportunities of the offered approach from existing is shown.