УДК 519.681

Д.А. Дёмин, к.т.н.
Т.И. Каткова

МЕТОД ОБРАБОТКИ МАЛОЙ ВЫБОРКИ НЕЧЕТКИХ РЕЗУЛЬТАТОВ ОРТОГОНАЛИЗОВАННОГО ПАССИВНОГО ЭКСПЕРИМЕНТА

Национальный технический университет «ХПИ», г.Харьков

Рассмотрена задача оценки уравнения регрессии для малой выборки исходных данных в случае, когда результаты проведенных экспериментов – нечеткие числа с известными функциями принадлежности. Предложенная процедура обеспечивает возможность обработки результатов усеченного ортогонального подплана плана полного факторного эксперимента. Полученная методика позволяет осуществить обоснованное отсеивание малозначимых факторов и взаимодействий.

***Ключевые слова**: нечеткие числа, план полного факторного эксперимента, репликоподобный план, ортогональный подплан.*

Введение

Многочисленные задачи оценки и прогнозирования состояния сложных технических, экономических, социальных и других систем, функционирующих в условиях многофакторной среды, сводятся к отысканию соотношения, связывающего значения контролируемых параметров $f_1, f_2 \ldots, fm$ среды **(**факторов) со значением некоторого результирующего параметра y – отклика, характеризующего состояние системы. Удобная и традиционная используемая математическая модель связи «среда – система» имеет вид так называемого регрессионного полинома Колмогорова-Габора:

$$y = a_0 + \sum_{i=1}^{m} a_i F_i + \sum_{i_1=1}^{m} \sum_{i_2>i_1} a_{i_1 i_2} F_{i_1 i_2} + \ldots + \sum_{i_1=1}^{m} \sum_{i_2>i_1} \ldots + \sum_{i_1=1}^{m} \sum_{i_2>i_1} \sum_{i_m>i_{m-1}} a_{i_1 i_2 \ldots i_m} F_{i_1 i_2 \ldots i_m} . \quad (1)$$

Для оценивания параметров уравнения регрессии (1) используются результаты непосредственных измерений значений факторов и отклика. При этом формируется матрица H и векторы A, Y:

$$H = \begin{pmatrix} 1 & F_{11} & F_{12} \ldots F_{1m} & F_{11}F_{12} \ldots & F_{11}F_{12} \ldots F_{1m} \\ 1 & F_{21} & F_{22} \ldots F_{2m} & F_{21}F_{22} \ldots & F_{21}F_{22} \ldots F_{2m} \\ .. & .. & \ldots & & .. \\ 1 & F_{n1} & F_{n2} \ldots F_{nm} & F_{n1}F_{n2} \ldots & F_{n1}F_{n2} \ldots F_{nm} \end{pmatrix}, \quad (2)$$

$A^T = \left(a_0\, a_1\, a_2 \ldots a_m\ a_{12} \ldots a_{12\ldots m}\right)$, $Y^T = \left(y_1\, y_2 \ldots y_n\right)$.

Здесь F_{ji} – значение i-го фактора в j-м эксперименте, y_j– значение результирующего параметра в j-м эксперименте.

Теперь искомый вектор A отыскивается по методу наименьших квадратов по формуле:

$$A = \left(H^T H\right)^{-1} H^T Y . \quad (3)$$

Практические трудности реализации этого подхода связаны с малой выборкой исходных данных, что приводит к неудовлетворительному соотношению между числом оцениваемых параметров – $N=2^m$ и числом проведенных экспериментов n. Возникающие при этом проблемы решаются либо увеличением числа опытов, что не всегда возможно, либо обоснованным снижением числа оцениваемых параметров. Один из перспективных путей снижения размерности вектора A состоит в искусственной ортогонализации результатов пассивного эксперимента [1]. При этом проводится масштабирование измеренных значений факторов к интервалу [-1, 1] по формулам:

$$X_{ji} = \frac{2F_{ji} - (F_{i\max} + F_{i\min})}{F_{i\max} - F_{i\min}}, \quad F_{i\max} = \max_j F_{ji}, \quad F_{i\min} = \min_j F_{ji}, \quad i = 1,2,\ldots,m, \quad j = 1,2,\ldots,n$$

В результате все точки, соответствующие проведенным экспериментам, в масштабированном пространстве факторов окажутся внутри гиперкуба с центром в начале координат и длинами ребер, равными двум.

Далее все множество результатов экспериментов разбивается на 2^m подмножеств (M_1, M_2,…, M_N) по правилу

$$M_l = \left\{ j : \min_s \left(X_s^0 - X_j \right)^T \left(X_s^0 - X_j \right) = \left(X_l^0 - X_j \right)^T \left(X_e^0 - X_j \right) \right\},$$

где X_s^0 - S-я вершина гиперкуба, $s = 1,2,...,N$, $X_j = \left(X_{j1}, X_{j2}, ..., X_{ji}, ..., X_{jm} \right)$ - точка в нормированном пространстве факторов, соответствующая *j-му* эксперименту, $j = 1,2,...,n$.

Теперь точки, принадлежащие каждому из полученных подмножеств, используются для получения локальных описаний поведения функций отклика в пределах соответствующей области факторного пространства. Эти локальные описания позволяют рассчитать значения функции отклика в вершинах гиперкуба, образующих в совокупности план полного факторного ортогонального эксперимента. Принципиальное достоинство такого плана состоит в возможности независимого оценивания каждого из параметров уравнения регрессии (1), позволяя осуществить обоснованное отсеивание малозначимых факторов и взаимодействий.

К сожалению, непосредственное использовании описанной технологии в реальных условиях затруднено вследствие неопределенности в отношении численных значений результатов экспериментов. Эти результаты, строго говоря, не могут трактоваться как случайные величины ввиду отсутствия законов их распределения, и поэтому их описание естественно осуществить в терминах нечеткой математики. В соответствии с этим **цель** статьи – получить метод оценивания коэффициентов уравнения регрессии в условиях нечетких исходных данных с использованием ортогонализации пассивного эксперимента.

Постановка задачи

Пусть значения y_1, y_2,…, y_n результатов эксперимента - нечеткие числа с функциями принадлежности $\mu(y_1)$, $\mu(y_2)$, …, $\mu(y_n)$. Поставим задачу оценивания параметров уравнения регрессии (1) с учетом нечеткости результатов эксперимента.

Основные результаты

Пусть M_l – множество из n_l точек масштабированного факторного пространства, примыкающих к вершине X_l^0, $l=1, 2,…, N$.

Введем локальное описание поведения функции отклика в соответствующей подобласти гиперкуба с помощью линейного по параметрам и по факторам полинома

$$y_l = b_{l0} + b_{l1} x_1 + ... + b_{lm} x_m \ . \qquad (4)$$

Параметры этого полинома определим методом наименьших квадратов:

$$B_l = \left(H_l^T H_l \right)^{-1} H_l^T Y_l,$$

где $H_l = \begin{pmatrix} 1 & x_{j_1 1} & x_{j_1 2} ... & x_{j_1 m} \\ 1 & x_{j_2 1} & x_{j_2 2} ... & x_{j_2 m} \\ .. & .. & ... & .. \\ 1 & x_{j_{n_l} 1} & x_{j_{n_l} 2} ... & x_{j_{n_l} m} \end{pmatrix}$, $B = \begin{pmatrix} b_{l_0} \\ b_{l_1} \\ ... \\ b_{l_m} \end{pmatrix}$, $y_l = \begin{pmatrix} y_{j_1} \\ y_{j_2} \\ ... \\ y_{j_{n_l}} \end{pmatrix}$,

Введем матрицу $R_l = \left(H_l^T H_l \right)^{-1} H_l^T = \left(r_{ij_k} \right)$, $i = 1,2,...,m$, $k = 1,2,...,n_l$.

Тогда $B_l = R_l Y_l = \left(\sum_{k=1}^{n_j} r_{1 j_k} y_{j_k} \quad \sum_{k=1}^{n_j} r_{2 j_k} y_{j_k} \quad \dots \quad \sum_{k=1}^{n_j} r_{m j_k} y_{j_k} \right)^T$.

Пусть функции принадлежности нечетких результатов измерений имеют вид:

$$\mu\left(y_{j_k} \right) = \exp\left\{ -\frac{\left(y_{j_k} - \bar{y}_{j_k} \right)^2}{\sum \sigma_{j_k}^2} \right\}, k = 1,2,..n_l \ .$$

Тогда функция принадлежности нечеткого числа $b_{l_i} = \sum_{k=1}^{n_l} r_{ij_k} y_{j_k}$ будет описываться соотношением

$$\mu\left(b_{l_i}\right)=\exp\left\{-\frac{\left(b_{l_i}-\overline{b}_{l_i}\right)^2}{D_{l_i}}\right\}, i=1,2,..m, \tag{5}$$

где $\overline{b_{l_i}}=\sum_{k=1}^{n_l} r_{ij_k}\overline{y}_{j_k}$, $D_{l_i}=\sum_{v=1}^{n_l} r_{ij_k}^2\sigma_{j_k}^2$

Далее с использованием (5) рассчитаем нечеткое значение y_l функции отклика в вершине X_l^0 гиперкуба. При этом

$$y_l=b_{l0}+b_{l1}x_{l1}^0+....+b_{lm}x_{lm}^0, \tag{6}$$

где $X_l^0=\left(x_{l_1}^0\ x_{l_2}^0\ ...\ x_{l_m}^0\right)=\left[(-1)^{\varepsilon_1^{l-1}+1}\ (-1)^{\varepsilon_2^{l-1}+1}...\ (-1)^{\varepsilon_{n_l}^{l-1}+1}\right]$, $\left(\varepsilon_m^{l-1}\ \varepsilon_{m-1}^{l-1}...\ \varepsilon_l^{l-1}\right)$ - двоичное разложение числа *l-1*.

Функция принадлежности нечеткого значения y_l в соответствии с (6) имеет вид:

$$\mu\left(y_l\right)=\exp\left\{-\frac{\left(y_l-\overline{y}_l\right)^2}{2D_l}\right\}, l=1,2,..N, \tag{7}.$$

где $\overline{y}_l=\sum_{i=1}^{m}\overline{b_{l_i}}(-1)^{\varepsilon_i^{l-1}+1}$, $D_l=\sum_{i=1}^{m}D_{l_i}\left[(-1)^{\varepsilon_i^{l-1}+1}\right]^2=\sum_{i=1}^{m}D_{l_i}$

Описанная операция повторяется для всех вершин гиперкуба, *l=1, 2, …, N* . Очередная проблема, являющаяся следствием малой выборки располагаемых наблюдений, состоит в том, что число точек в каждой их подобластей может оказаться существенно различным. В [2] рассмотрен возможный путь преодоления этой трудности, основанный на формировании усечённого репликоподобного ортогонального подплана плана полного факторного эксперимента. При этом в [2] показано, что такой усечённый план определяется решением системы уравнений с булевыми переменными

$$\begin{aligned}
&\sum_{i_3=1}^{2^p} z_{i_1i_2i_3}=1,\ \ i_1=1,2,..2^p,\ \ i_2=1,2,..2^p,\\
&\sum_{i_2=1}^{2^p} z_{i_1i_2i_3}=1,\ \ i_1=1,2,..2^p,\ \ i_3=1,2,..2^p,\\
&\sum_{i_1=1}^{2^p} z_{i_1i_2i_3}=1,\ \ i_2=1,2,..2^p,\ \ i_3=1,2,..2^p
\end{aligned} \tag{8}$$

где *(i_1, i_2, i_3)* – номер строки полного факторного эксперимента в 2^p-ричной системе счисления, $p=\frac{m}{3}$, $z_{i_1i_2i_3}=\begin{cases}1, \textit{если строка}\left(i_1i_2i_3\right)\textit{включена в план},\\ 0\ \textit{в противном случае.}\end{cases}$

Там же предложен метод выбора наилучшего решения системы уравнений (8), основанный на минимизации критерия

$$D_{\max}=\max_l\left\{D_l\right\}, \tag{9}$$

задающего *l* уровень некомпактности соответствующего плана.

При этом соответствующая процедура опирается на достаточно жесткое предположение о равенстве ошибок измерения функции отклика во всех точках наблюдений. Снимем это предположение. Искомое решение системы уравнений (8) получим итерационно следующим образом.

Введем кубическую матрицу $\Delta=\left(\delta_{i_1i_2i_3}\right)$, положив $\delta_{i_1i_2i_3}=1$ для всех i_1*=1, 2,…, 2*^{p}, i_2*=1, 2,…, 2*^{p}, i_3*=1, 2,…, 2*^{p}. Дополним теперь систему уравнений (8) совокупностью неравенств

$$z_{i_1i_2i_3}\le\delta_{i_1i_2i_3} \tag{10}$$

Поскольку переменные $z_{i_1 i_2 i_3}$ - булевы, то неравенства (10) никак не влияют на решение системы уравнений (8). Получим какое-либо решение этой системы - $z_{i_1 i_2 i_3}$, i_l=*1, 2,…, 2^p*, i_2=*1, 2,…, 2^p*, i_3=*1, 2,…, 2^p*.

Введем множество индексов $N = \{(i, i\ i) : z = 1\}$. Теперь выберем $D_{1\max} = \max_{l \in N_l} D_l$ и в матрице $\Delta = \left(\delta_{i_1 i_2 i_3}\right)$ положим $\delta_{i_1 i_2 i_3} = 1$ для всех $(i_1, i_2\ i_3) \in \overline{N_1}$, где $\overline{N_1}$ - множество всех элементов Δ, для которых $D_l \geq D_{1\max}$.

Вновь будем решать систему уравнений (8) с учетом неравенств (10). Понятно, что теперь эти неравенства запретят использовать при получении решения те элементы матрицы $\left(z_{i_1 i_2 i_3}\right)$, для которых оценка уровня некомпактности тела неопределённости, задаваемого функцией принадлежности нечёткого значения функции отклика в соответствующих вершинах, превосходит $D_{1\max}$. Если такое решение получено, например, $\left\{z^{(2)}_{i_1 i_2 i_3}\right\}$, то для него, аналогично предыдущему, отыскиваем множество индексов $N_2 = \left\{(i_1, i_2\ i_3) : z^{(2)}_{i_1 i_2 i_3} = 1\right\}$, вычисляем $D_{2\max} = \max_{l \in N_2} D_l$ и, после этого, в матрице $\Delta = \left(\delta_{i_1 i_2 i_3}\right)$ дополнительно полагаем $\delta_{i_1 i_2 i_3} = 0$ для всех $(i_1, i_2\ i_3) \in \overline{N_2}$, где $\overline{N_2}$ - множество всех элементов Δ, для которых $D_l \geq D_{2\max}$. В этом случае решение задачи должно быть продолжено. Если же такое решение отсутствует, то план, полученный на первой итерации, является искомым. Соответствующий этому решению набор нечетких чисел y_l используем для независимой оценки параметров уравнения регрессии (1). При этом нечеткая оценка конкретного параметра a_p уравнения (1) рассчитывается как линейная комбинация нечётких значений y_l, *l=1, 2, …, N_0,* взвешенных со значениями элементов *p*-го столбца матрицы *X,* составленной из векторов X_l^0, *l=1, 2, …, N_0,* где N_0 – множество вершин, вошедших в усечённый план.

Функция принадлежности этого параметра a_p имеет вид:

$$\mu_p\left(a_p\right) = \exp\left\{-\frac{\left(a_p - \overline{a_p}\right)^2}{2D_p}\right\},$$

где $\overline{a}_p = \frac{1}{N}\sum_{l=1}^{N} x_{lp}\,\overline{y}_l$, $D_p = \frac{1}{N}\sum_{l=1}^{N} D_l$, $p = 1, 2, ..., N$.

Завершает процедуру оценивания параметров уравнения регрессии (1) наименее формализованный этап – принятие решения относительно значимости каждого из факторов и их взаимодействий. Возможный подход к решению этой задачи состоит в следующем. Будем считать, что *p*-й элемент уравнения регрессии (1) не значим, если функция принадлежности $\mu_p\left(a_p\right)$ накрывает нуль и значение $\mu_p(0)$ не ниже некоторого заданного δ (например, δ *=0,05).*

Выводы

Таким образом, получена процедура оценивания параметров уравнения регрессии, связывающего значения факторов, задающих характеристики режима эксплуатации системы, со значением некоторого результирующего параметра, определяющего качество (эффективность) системы, в условиях малой выборки нечётких исходных данных.

Список литературных источников

1. Серая О.В. Оценивание параметров уравнения регрессии в условиях малой выборки [текст]/Серая О.В., Дёмин Д.А. //Східно-Європейський журнал передових технологій. – 2009. - №6/4 (42). – с.14-20.

2. Раскин Л.Г. Искусственная ортогонализация пассивного эксперимента в условиях малой выборки [текст]/Раскин Л.Г., Дёмин Д.А. //Інформаційно-керуючі системи на залізничному транспорті. – 2010. - №1. – с.20-23.