УДК 519.24

Л.Г. Раскин, О.В. Серая, В.В. Карпенко

Процедура повышения точности оценивания параметров многофакторного уравнения регрессии для малой выборки

Национальный технический университет "Харьковский политехнический институт"

Постановка задачи. При изучении сложных систем, процессов и объектов реальной действительности традиционно возникает задача описания зависимости какого-либо результирующего признака от набора влияющих факторов и их взаимодействий. В ходе решения такой задачи обычно выбирают какую-либо математическую модель процесса и на основе имеющихся экспериментальных данных проводят оценку параметров модели. Для большого класса систем можно получить решение этой задачи с помощью многофакторного уравнения регрессии, которое позволяет оценить вклад и самих влияющих факторов, и их взаимодействий. Оценки параметров модели находят с помощью метода наименьших квадратов.

Однако для некоторого круга задач прямое применение метода наименьших квадратов невозможно. Такая ситуация возникает при большом числе учитываемых признаков и относительно малом числе имеющихся экспериментальных данных. Часто вполне адекватное описание модели позволяет получить методика приближенного оценивания параметров многофакторного уравнения регрессии [1]. Методика основана на описании неизвестных параметров через меньшее число параметров, оцениваемых непосредственно по экспериментальным данным. Результаты, полученные с помощью данной методики, дают приближенную оценку влияния факторов и их взаимодействий на результирующий признак.

Тем не менее, во многих случаях такие оценки являются недостаточно точными, и в связи с этим необходимо более адекватное описание модели. Как известно, простая технология получения независимых оценок влияния факторов и их взаимодействий на результирующий признак состоит в использовании полного факторного ортогонального эксперимента. Но в реальных условиях для целого класса задач проведение активного эксперимента либо невозможно, либо сопряжено с серьезными трудностями. В связи с изложенным возникающая проблема может быть охарактеризована следующим образом. Прямое построение уравнения регрессии и оценивание его параметров невозможны ввиду малости выборки. С другой стороны, трудности организации и проведения активного эксперимента не дают возможности снижения размерности задачи вследствие предварительного отсеивания малозначимых факторов и их взаимодействий.

В этой ситуации возможный путь решения задачи состоит в искусственном формировании ортогонального плана полного факторного эксперимента (ОПФЭ). Цель состоит не в том, чтобы сразу получить адекватное уравнение регрессии, а в том, чтобы удалить из него малозначимые факторы и взаимодействия. При этом для формирования такого ОПФЭ может быть использована любая технология описания связи между влияющими и результирующими переменными, обладающая хорошими прогностическими свойствами. В частности, эта цель достигается с применением нейронной сети. При этом по реальным данным

осуществляется обучение и тестирование сети. Далее полученная сеть обеспечивает расчет оценок значений результирующего признака в ортогональных точках факторного пространства. Последующая обработка позволяет сформировать адекватную структуру уравнения регрессии. Удаление слабо влияющих факторов и взаимодействий во многих случаях очень существенно снижает размерность уравнения регрессии, параметры которого оцениваются в дальнейшем с использованием всего имеющегося статистического материала.

Таким образом, предлагается следующая трехшаговая процедура обработки данных в условиях малой выборки:

1. С использованием, например, нейронной сети устанавливают связь между входными переменными и результирующей.

2. Формируют план ортогонального полного факторного эксперимента и осуществляют отсев малозначимых факторов и взаимодействий.

3. Строят усеченное уравнение регрессии и по совокупности исходных данных оценивают его параметры.

Рассмотрим предложенную процедуру более подробно.

Установление связи между многомерным входом и выходом. Для установления связи между влияющими и результирующими переменными используем нейронную сеть. Важным достоинством этого метода является возможность его применения к широкому кругу задач, независимо от их природы. Кроме того, он легко позволяет моделировать связи при наличии сложным образом скомбинированных взаимозависимостей факторов.

Перед использованием нейронной сети вначале необходимо выбрать её конфигурацию. Как правило, нейронная сеть состоит из входного, выходного и нескольких скрытых слоев. Обычно число скрытых слоев невелико, так как вклад взаимодействий, порядок которых больше двух, невелик. Для многих задач достаточно всего одного скрытого слоя. Число узлов в слоях выбирается в зависимости от количества оцениваемых факторов. Во входном слое число узлов определяется по количеству влияющих факторов. В скрытых слоях число узлов подбирается, исходя из предполагаемого количества значимых факторов и взаимодействий. При этом необходимо провести несколько вычислительных экспериментов с разным количеством узлов для выявления оптимальной конфигурации сети. Для обучения нейронной сети используют алгоритм обратного распространения ошибки. Обучение происходит до момента, когда среднеквадратическая ошибка становится меньше некоторого, достаточно малого, числа. Критерием останова может служить также прекращение повышения точности расчета (среднеквадратическая ошибка перестает существенно уменьшаться). При обучении используют случайно отобранную часть исходных данных, на оставшихся проводят тестирование полученной сети.

Далее полученная сеть обеспечивает расчет оценок значений результирующего признака в ортогональных точках факторного пространства.

Формирование ОПФЭ и оценка параметров уравнения регрессии. Для формирования ортогонального плана полного факторного эксперимента строится соответствующая матрица, реализующая сочетание значений всех факторов и их взаимодействий на двух уровнях. Оценку коэффициентов уравнения регрессии с использованием полного факторного эксперимента осуществим следующим образом. $n = 2^m$

Пусть изучается m факторов, тогда $n = 2^m$ - общее число оцениваемых факторов и всех взаимодействий. Если обозначить вклад i-го фактора (или

взаимодействия) через x_i, $i=0,1,...,N$, где $N=n-1$, то результирующий признак рассчитывают по следующей формуле: $y=\sum_{i=0}^{N} b_i x_i$. Здесь b_i – коэффициент при i-м факторе уравнения регрессии.

Введем матричные обозначения

$$Y=\begin{pmatrix} y_0 \\ y_1 \\ . \\ y_N \end{pmatrix},\ B=\begin{pmatrix} b_0 \\ b_1 \\ . \\ b_N \end{pmatrix},\ X=\begin{pmatrix} x_{00} & x_{01} & \dots & x_{0N} \\ x_{10} & x_{11} & \dots & x_{1N} \\ & & \dots & \\ x_{N,0} & x_{N,1} & \dots & x_{N,N} \end{pmatrix},$$

где X – матрица факторного эксперимента, столбцы которой ортогональны попарно.

Для вычисления коэффициентов вектора B используем метод наименьших квадратов. В матричных обозначениях функционал наименьших квадратов записывают в виде: $I=(Y-XB)^T(Y-XB)$. Как известно, искомый вектор B, минимизирующий сумму квадратов отклонений экспериментальных данных от рассчитываемых, вычисляют по формуле $B=(X^TX)^{-1}X^TY$, которая для ОПФЭ упрощается к виду $B=\frac{1}{N+1}X^TY$. Полученное соотношение используют для расчета значений функции отклика (результирующего признака) в ортогональных точках ОПФЭ.

В целях повышения точности оценивания этих значений и элиминирования случайности при разделении всех статистических данных на обучающую и проверочную выборки, целесообразно эту процедуру повторить несколько раз, используя разные способы формирования этих выборок.

Пусть проведено q соответствующих вычислительных экспериментов, в каждом из которых реализуется процедура обучения нейронной сети и расчета значений функции отклика в ортогональных точках ОПФЭ.

Тогда

$$y_j=\frac{1}{q}\sum_{k=1}^{q} y_{kj},\ s_{y_j}^2=\frac{1}{q-1}\sum_{k=1}^{q}(y_{kj}-\overline{y_j})^2,\ j=0,1,...,N,$$

где y_{kj} – оценка значения функции отклика в j-й точке ОПФЭ в k-м эксперименте.

Однородность дисперсий проверяют по критерию Кохрена: $G_P=s_{\max}^2\Big/\sum_{j=0}^{N} s_{y_j}^2$. Если расчетное значение G_P не превышает табличного G_T, то дисперсии усредняют по формуле $s_0^2=\frac{1}{N}\sum_{j=0}^{N} s_{y_j}^2$. Расчет дисперсии ошибок

$$s_{b_i} = \frac{s_0^2}{N}, \quad i = 0,1,...,N.$$

коэффициентов проводят по формуле $s_{b_i} = \frac{s_0^2}{N}, \quad i = 0,1,...,N.$ Далее осуществляют оценку значимости коэффициентов уравнения регрессии в соответствии со стандартной технологией, использующей критерий Стьюдента [2]. При этом вычисляют доверительные интервалы, накрывающие истинные значения коэффициентов с данной вероятностью. Если вычисленный доверительный интервал не захватывает нуль, то коэффициент значим, иначе его отбрасывают. Так проверяются все коэффициенты уравнения регрессии, в результате чего выявляется адекватная структура этого уравнения.

Оценка параметров усеченного уравнения регрессии. После отбрасывания незначимых факторов и взаимодействий получается усеченное уравнение $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_l x_l$, где l – число значимых факторов и взаимодействий, оставшихся после отбрасывания незначимых.

Оценку коэффициентов этого уравнения вновь проведем с помощью метода наименьших квадратов, обрабатывая совместно весь имеющийся статистический материал. Пусть $\underline{Y^r} = (y_1 \quad y_2 \quad ... \quad y_r)$ – вектор полученных в r-опытах значений результирующего фактора, $B_l^T = (b_0 \quad b_1 \quad b_2 \quad ... b_l)$ – искомые оценки коэффициентов уравнения, $X_l = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1l} \\ ... & ... & ... & ... & x_{2l} \\ 1 & x_{r1} & x_{r2} & ... & x_{rl} \end{pmatrix}$.

Тогда функционал наименьших квадратов записываем так: $I = (Y - X_l B_l)^T (Y - X_l B_l)$, а оценки коэффициентов вектора B получаем по формуле $\widetilde{B} = (X^T X)^{-1} X^T Y$.

Если после отсеивания малозначимых факторов и взаимодействий в соответствии с описанной методикой число оставшихся всё еще остается большим, то целесообразно использовать методику приближенного оценивания параметров многофакторного уравнения регрессии для малой выборки [1].

Выводы. Предложенная пошаговая процедура объединяет достоинства нейросетевых методов и ортогонального планирования экспериментов. Такой комбинированный подход позволяет в условиях недостаточности экспериментальных данных и большого числа оцениваемых факторов получить более точную, по сравнению с обычными методами, оценку параметров многофакторного уравнения регрессии.

Список литературы

1. *Раскин Л.Г., Серая О.В., Карпенко В.В.* Приближенное оценивание параметров многофакторного уравнения регрессии // Інформаційно-керуючі системи на залізничному транспорті. – 2003. – № 4. – С. 71–73.
2. *Колемаев В.А., Калинина В.Н.* Теория вероятностей и математическая статистика. – М.: ИНФРА, 2001. – 302 с.