

А. А. ПАВЛОВ, д-р техн. наук, проф., декан факультета информатики и вычислительной техники НТУУ «КПИ», Киев;

А. В. ЧЕХОВСКИЙ, студент НТУУ «КПИ», Киев

МОДИФИЦИРОВАННЫЙ АЛГОРИТМ ПОСТРОЕНИЯ МНОГОМЕРНОЙ ПОЛИНОМИАЛЬНОЙ РЕГРЕССИИ ПО ИЗБЫТОЧНОМУ ОПИСАНИЮ

У статті наводиться ефективний алгоритм відновлення багатовимірної поліноміальної регресії за надлишковим описом, що не приводить до послідовного накопичення помилок.

В статье приводится эффективный алгоритм восстановления многомерной полиномиальной регрессии по избыточному описанию не приводящий к последовательному накоплению ошибок.

The paper presents an efficient algorithm for reconstructing multivariate polynomial regression on excessive description that does not lead to a successive accumulation of errors.

В [1] приведен метод построения многомерной полиномиальной регрессии по избыточному описанию в условиях активного эксперимента. Метод реализует минимально необходимое число экспериментов (в схеме с фиксацией значений всех переменных кроме одной), но может приводить к существенному увеличению ошибок оценок неизвестных коэффициентов регрессии в результате решения последовательно конструируемых систем линейных равенств в следствие многократного использования в этих равенствах ранее найденных оценок коэффициентов регрессии. Этот эффект наиболее часто наблюдается в задачах большой размерности. Рассмотрим модификацию изложенного в [1] метода, которая исключает последовательное накопление ошибок оценок коэффициентов многомерной полиномиальной регрессии.

Возможность для одномерного случая практически гарантированно находить степень полинома линии регрессии, вычислять с допустимой вероятностью с заданной погрешностью коэффициенты этого полинома [1] позволяет предложить достаточно эффективную процедуру восстановления многомерной полиномиальной линии регрессии (при условии реализации активного эксперимента).

Пусть многомерная модель задается в виде

$$y(\bar{x}) = \sum_{\forall (i_1, \dots, i_n) \in K} \sum_{\forall (j_1, \dots, j_n) \in K(i_1, \dots, i_n)} b_{i_1, \dots, i_n}^{j_1, \dots, j_n} (x_{i_1})^{j_1} \cdot (x_{i_2})^{j_2} \dots (x_{i_n})^{j_n} + E, \quad (1)$$

где $\bar{x} = (x_1 \dots x_n)^T$ – детерминированный вектор входных переменных;

x_i – i -тая компонента вектора \bar{x} ;

$b_{i_1, \dots, i_l}^{j_1, \dots, j_l}$ – неизвестные коэффициенты;

j_l – натуральные числа;

i_l – натуральные индексы из множества $\{1, \dots, n\}$;

E – случайная величина с нулевым математическим ожиданием и ограниченной неизвестной дисперсией σ_E^2 (как и в одномерном случае может быть известна верхняя оценка σ_E^2).

Модель (1) является избыточной – возможно, некоторые из коэффициентов $b_{i_1, \dots, i_l}^{j_1, \dots, j_l}$ равны нулю. Для удобства дальнейшего изложения линию регрессии модели (1) представим иначе:

$$\sum_{l=1}^n \sum_{\forall (i_1, \dots, i_l) \in K_l} \sum_{\forall (j_1, \dots, j_l) \in K_l(i_1, \dots, i_l)} b_{i_1, \dots, i_l}^{j_1, \dots, j_l} (x_{i_1})^{j_1} \cdot (x_{i_2})^{j_2} \dots (x_{i_l})^{j_l}. \quad (2)$$

Составляющие

$$\sum_{\forall (i_1, \dots, i_l) \in K_1} \sum_{\forall (j_1, \dots, j_l) \in K_1(i_1, \dots, i_l)} b_{i_1, \dots, i_l}^{j_1, \dots, j_l} (x_{i_1})^{j_1} \cdot (x_{i_2})^{j_2} \dots (x_{i_l})^{j_l} \quad (3)$$

содержат все слагаемые из (1), в каждую из которых входит компонента x_1 .

Составляющие

$$\sum_{\forall (i_1, \dots, i_l) \in K_1} \sum_{\forall (j_1, \dots, j_l) \in K_1(i_1, \dots, i_l)} b_{i_1, \dots, i_l}^{j_1, \dots, j_l} (x_{i_1})^{j_1} \cdot (x_{i_2})^{j_2} \dots (x_{i_l})^{j_l}, l = \overline{2, n}, \quad (4)$$

содержат все слагаемые из (1), в каждую из которых входит компонента x_l ,

Рассмотрим составляющую (3). Обозначим через $M_j^1, j = \overline{1, n_1}$, количество слагаемых, каждая из которых содержит x_1 в j -й степени; $M^1 = \max_j M_j^1, j = \overline{1, n_1}$, n_1 – максимальная степень полинома от переменной x_1 .

Фиксируем M^1 наборов значений компонент $x_2^s, \dots, x_n^s, s = \overline{1, M^1}$. На числа $x_i^s, i = \overline{2, n}, s = \overline{1, M^1}$, накладывается единственное условие – определенные ниже квадратные матрицы должны быть невырожденными.

Реализуем M^1 экспериментов, в каждом из которых (s -м, $s = \overline{1, M^1}$) переменные x_2, \dots, x_n принимают фиксированные значения $x_i^s (i = \overline{2, n})$, а переменная x_1 изменяется как при построении одномерной полиномиальной регрессии. При фиксированных значениях переменных x_2, \dots, x_n в s -м экспе-

рименте ($s = \overline{1, M^1}$) многомерная линия регрессии превращается в полином от переменной x_1 степени n_1 .

Для каждого s -го эксперимента ($s = \overline{1, M^1}$) находим значения дисперсий $D\widehat{\theta}_j^s$, $j = \overline{1, n_1}$, и эти числа ранжируем по возрастанию их значений при фиксированном j . Получим n_1 проранжированных последовательностей оценок коэффициентов $\theta_j^{s_1}, \dots, \theta_j^{s_{M^1}}$ ($j = \overline{1, n_1}$).

Эти результаты позволяют сформировать n_1 систем линейных уравнений, решениями которых являются значения всех коэффициентов $b_{i_1, \dots, i_j}^{j_1, \dots, j_i}$ в выражении (3).

Действительно, в каждом из s экспериментов неизвестные коэффициенты $\widehat{\theta}_j^s$ ($j = \overline{1, n_1}$) одномерной полиномиальной регрессии степени n_1 от переменной x_1 определяются следующим образом: необходимо из всех членов выражения (содержащих переменную x_1 в степени j) вынести x_1^j . Полученное выражение для θ_j^s содержит только M_j^1 неизвестных коэффициентов вида $b_{i_1, \dots, i_j}^{j_1, \dots, j_i}$, так как в каждом s -м эксперименте при изменении значений переменной x_1 переменные x_i , $i = \overline{2, n}$ принимают одно и то же фиксированное значение x_i^s , $i = \overline{2, n}$. Таким образом, для построения системы линейных уравнений для нахождения M_1^1 коэффициентов вида $b_{i_1, \dots, i_j}^{j_1, \dots, j_i}$ надо использовать M_1^1 чисел $\widehat{\theta}_1^{s_1}, \dots, \widehat{\theta}_1^{s_{M^1}}$ (они имеют наименьшую дисперсию).

Для определения верхних статистических оценок точности нахождения M_1^1 коэффициента вида $b_{i_1, \dots, i_j}^{j_1, \dots, j_i}$, полученную систему линейных уравнений условно запишем так:

$$A \begin{pmatrix} x_1 \\ \vdots \\ x_{M^1} \end{pmatrix} \begin{pmatrix} \widehat{\theta}_1^{s_1} \\ \vdots \\ \widehat{\theta}_1^{s_{M^1}} \end{pmatrix}, \quad (5)$$

где x_i , $i = \overline{1, M_1^1}$ – переменные (соответствующие M_1^1 переменным вида $b_{i_1, \dots, i_j}^{j_1, \dots, j_i}$).

Пусть оценки $\widehat{\theta}_l^{s_l}, l = \overline{1, M_1^1}$, с заданной статистически значимой вероятностью p оценивают $\theta_l^{s_l}$ с погрешностью, по модулю не превышающей чисел $\Delta_l^{s_l}, l = \overline{1, M_1^1}$. Тогда с вероятностью p максимальная погрешность нахождения точных значений M_1^1 соответствующих коэффициентов вида $b_{i_1, \dots, i_l}^{j_1, \dots, j_l}$ имеет вид

$$\max_{j=1, M_1^1} \left\{ \max \left(\sum_l^{(+)} a_{jl}^{-1} \Delta_l^{s_l}, \sum_l^{(-)} |a_{jl}^{-1}| \Delta_l^{s_l} \right) \right\}, \quad (6)$$

где $\sum_l^{(+)} a_{jl}^{-1} \Delta_l^{s_l}$ берется по всем $l = \overline{1, M_1^1}$, для которых $a_{jl}^{-1} \geq 0$; $\sum_l^{(-)} a_{jl}^{-1} \Delta_l^{s_l}$ берется по всем $l = \overline{1, M_1^1}$, для которых $a_{jl}^{-1} < 0$; a_{jl}^{-1} – jl -й элемент матрицы A^{-1} .

Как указывалось выше, предполагается, что $x_i^s, i = \overline{2, n}, s = \overline{1, M_1^1}$ выбраны так, что матрица A^{-1} существует.

Для каждой переменной $x_l, l = \overline{2, n}$ нахождение коэффициентов $b_{i_1, \dots, i_l}^{j_1, \dots, j_l}$ в выражениях (4) реализуется точно так же, как это описано для переменной x_1 .

Недостаток модифицированной схемы заключается в увеличении необходимого числа экспериментов, повторной оценке коэффициентов слагаемых, присутствующих в выражениях (4), которые были найдены на предыдущих этапах.

Преимущество заключается в отсутствии последовательного накопления ошибок и появлении косвенного критерия эффективного восстановления многомерной регрессии: практическое совпадение оценок повторяющихся коэффициентов многомерной регрессии в результате решения разных систем линейных равенств.

Изложенные выше методики требуют избыточное число экспериментов для практически точного нахождения коэффициентов в слагаемых полинома вида $a_i x_{i_1} x_{i_2} \dots x_{i_k}$. Так при фиксации значений всех переменных кроме одной получаем слагаемое первой степени относительно скалярной переменной. В этом случае удобен следующий прием:

кладем $x_{i_1} = x_{i_2} \dots x_{i_{k-1}} = x$, а значения x_{i_k} фиксируются; либо $x_{i_1} = x_{i_2} \dots = x_{i_k} = x$. Тогда в одномерных регрессиях соответствующие члены имеют вид $Q_{ke} a_i x^{k-1}$, (Q_{ke} фиксированные числа), либо $a_i x^k$. В этом случае одно и тоже количество экспериментов приводит на порядки более точному нахождению коэффициентов при x^{k-1} либо x^k в одномерных регрессиях и

следовательно решение соответствующих систем линейных равенств приводит к качественно более точным оценкам коэффициентов a_i . Представление независимых переменных $x_i, i = \overline{1, n}$ в одном из двух приведенных выше видов определяется структурой избыточного полиномиального описания многомерной регрессии и как следствие требованиями к системам линейных равенств для оценки неизвестных коэффициентов линии регрессии. Из анализа решения конкретных примеров следует что приведенный прием может приводить к построению существенно меньшего количества одномерных регрессий.

Список литературы: 1. Згуровский М. З. Принятие решений в сетевых системах с ограниченными ресурсами: Монография / М. З. Згуровский, А. А. Павлов // К. : Наукова думка, 2010. – 574 с.

Надійшла до редколегії 01.02.2012